ESSAYS ON MACROECONOMIC DYNAMICS
AND THE ECONOMETRICS OF EXPECTILES

BY
COLLIN S. PHILIPPS

DISСERТATION
Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Economics
in the Graduate College of the
University of Illinois Urbana-Champaign, 2021

Urbana, Illinois

Doctoral Committee:
Assistant Professor Minchul Shin, Chair and Director of Research
Professor Dan Bernhardt
Assistant Professor Pooyan Amir-Ahmadi
Professor George Deltas
Abstract

This is a collection of studies considering models that behave differently in different scenarios. In four essays, we apply this approach to (1) macroeconomic dynamics and (2) expectile regression, which is a latent topic in the literature.

In the first essay, we investigate government spending multipliers using a two-regime model and impulse response functions with fully endogenous regimes. While short-run multipliers vary depending on business cycle fluctuations, we find little evidence that medium or long-run multipliers vary between expansions and recessions. The reason for state dependence found in the literature is the constant-regime assumption used to create impulse response functions. Importantly, a fiscal policy shock has little effect on the duration of a recession.

In the second essay, we show that the multiplier does not depend on the monetary policy rule. What we find is that the monetary policy rule itself changes after a government spending shock and converges quickly to a similar regime regardless of the initial condition. This rapid change in monetary policy leaves the multiplier unaffected by the initial monetary policy regime. An exception to this characterization of monetary policy occurs when nominal interest rates are stuck at zero. We analyze the multiplier at the zero-lower bound and find that the multiplier exceeds one.

The third essay re-introduces expectile regression. In some cases where OLS assumptions are violated, an expectile regression estimator is also the BLUE for the mean regression: we give three examples. Expectile regression is the BLUE for quantile regression coefficients in special cases where they are equal. But expectiles can be used in some models where quantiles are not helpful, such as binary response models. In those cases, expectile regression
is the new best option.

The fourth essay dispels misinformation from the literature. Two different likelihood models have been suggested for estimating expectiles as a maximum likelihood estimator. After comparison, it becomes clear that they are not the same and only one of these models is appropriate for that purpose.
To My Grandmother.
Acknowledgments

I give my deepest thanks to the people who helped me complete this project. Many thanks go to the members of my committee who suffered through this treacherous process with me. I thank Minchul Shin for checking countless lines of math and Dan Bernhardt for proofreading many revisions of these papers. I thank George Deltas and Pooyan Amir-Ahmadi for their invaluable support. Additionally, I would like to thank Roger Koenker, whose guidance was instrumental during the early days of this project. Lastly, I thank my friends and colleagues at Illinois State and others around the world who gave their valuable comments and feedback over the years.
# Table of Contents

Chapter 1 Does the Government Spending Multiplier Depend On the Business Cycle? ................................................................. 1

Chapter 2 Government Spending Between Active and Passive Monetary Policy ................................................................. 29

Chapter 3 When is an Expectile the Best Linear Unbiased Estimator? . 64

Chapter 4 The MLE of Aigner, Amemiya, and Poirier is *not* the Expectile MLE ................................................................. 123

References ......................................................................................... 147

Appendix A: Appendix to “Does the Government Spending Multiplier Depend On the Business Cycle?” ........................................ 158

Appendix B: Appendix to “Government Spending Between Active and Passive Monetary Policy?” ........................................ 176

Appendix C: Appendix to “When is an Expectile the Best Linear Unbiased Estimator?” ......................................................... 209
Chapter 1

Does the Government Spending Multiplier Depend On the Business Cycle?

1.1 Introduction

We ask whether the government spending multiplier varies with the business cycle. Fiscal policy in general has received significant attention from economists and policymakers since the Great Recession. At that time, the federal funds rate and other major nominal interest rates hit the Zero Lower Bound (ZLB) which made further conventional monetary policy interventions impossible. Partly because of this, large fiscal policy packages were implemented by governments around the world to fight the recession. The effects of counter-cyclical fiscal policy are now a major public debate.

Since the seminal paper by Auerbach and Gorodnichenko [2012], this question about state-dependent government spending multipliers has been the focus of a fast-growing body of empirical literature. In spite of a substantial amount of research attention, the empirical literature on this subject has not converged towards any consensus. In point of fact, the literature provides evidence for three opposing scenarios. On the one hand, Auerbach and Gorodnichenko [2012, 2013, 2017], Bachmann and Sims [2012], Mittnik and Semmler [2012], Fazzari et al. [2015] and Biolsi [2017] find that the spending multiplier is higher during recessions than during expansions. The underlying argument is that, during expansions, the economy is close to the resource constraint which limits the impact of additional government spending. On the other hand, Caggiano et al. [2015] and Ramey and Zubairy [2018] provide evidence that there is no significant difference between multipliers in recessions and expansions. Finally, there is the view by Bognanni [2013] and Alesina et al. [2018] arguing that the multiplier is higher during expansions because, for example, economic uncertainty
is relatively high during recessions and uncertainty reduces economic activity.\(^1\)

The goal of this study is to shed light on the origins of these differences. We first identify the assumptions that the literature relies on. We find that the literature (i) calibrates important model parameters, (ii) focuses on the long-run multiplier, (iii) achieves identification via the recursive or the narrative strategy, and (iv) assumes that the economy remains in the initial state after it is hit by a government spending shock. Because these assumptions can be strong, we question whether the dominating finding of higher multipliers in recessions found in the literature is the result of identifying assumptions or indeed a feature of the data.

To examine this question, we replace each of the assumptions step-by-step with weaker restrictions. First, we fully estimate the nonlinear smooth-transition VAR (ST-VAR) model used by Auerbach and Gorodnichenko [2012] via Bayesian estimation techniques. This allows us to evaluate the full posterior distribution of the model, multipliers, and impulse response functions in places where the rest of the literature has fixed model parameters.

Second, we look not only at the long-run multiplier but also at the short- and medium-run multipliers. We know of no justification to exclusively focus on the long-run multiplier. Interesting short-run dynamics may remain undiscovered. In addition, policymakers may be even more interested in the short- and medium-run effects of policy interventions if they are concerned with recessions, which are often short.

Third, we replace the recursive identification used by Auerbach and Gorodnichenko [2012] and many others with sign restrictions on impulse response functions. Recursive identification in the empirical government spending literature has its origin in work by Blanchard and Perotti [2002] who find that government spending is contemporaneously unaffected by the business cycle. This finding can be implemented into SVAR models by a Cholesky ordering strategy with government spending ordered first. Recursive identification of this type imposes zero restrictions on the contemporaneous effects of some variables on others; these restrictions lack theoretical foundations and results thus obtained may vanish.

\(^1\)Moreover, Barnichon and Matthes [2017] provide mixed results.
as soon as these restrictions are removed [Uhlig, 2005]. In addition, if GDP and government spending both affect the other contemporaneously, then there exists no feasible Cholesky ordering for a VAR. For example, the Troubled Asset Relief Program (TARP) of 2008 was passed by Congress and spent several hundreds of billions in the same quarter (see Webel [2011]). This observation suggests that government spending can respond quickly to the level of output in times of crisis, challenging Blanchard and Perotti [2002]’s main result.

We also do not rely on the narrative identification developed by Ramey [2011] and used in Ramey and Zubairy [2018]. Ramey [2011] builds a narrative time series based on information from newspapers to quantify agents’ expectations regarding future government spending plans and orders it first in a recursive VAR. It is widely known that agent’s expectations matter for fiscal policy analysis. If they are omitted from the model, the VAR model becomes misspecified and loses the ability to consistently estimate impulse response functions—a problem called fiscal foresight. The narrative literature assumes that Ramey’s news defense shocks carry information about agents’ expectations, and including them in the model helps to overcome the fiscal foresight problem. However, in addition to the zero restrictions related to the recursive implementation, the macroeconometrics literature now recognizes that narrative time series, such as Ramey’s news defense shocks, are measured with error and should be viewed as instruments rather than as perfect measures of structural shocks [Stock and Watson, 2012, Mertens and Ravn, 2012].

In contrast, sign restrictions on impulse response functions are considered to be less restrictive than the recursive or the narrative identification. The researcher only restricts the sign of a subset of the impulse response functions to characterize a structural shock without imposing any further constraints. For example, we define a government spending shock as a shock that raises output, prices and government spending, which is consistent with a large class of models. Furthermore, this approach allows for contemporaneous co-movements between variables, because all variables can contemporaneously respond to all

---

2 Similarly, the American Recovery and Reinvestment Act (ARRA) of 2009 was introduced, passed by congress, and signed by then-President Barack Obama in only 22 days.

3 Additionally, Caldara and Kamps [2017] argue that Ramey’s news defense shocks are not a suitable instrument for government spending shocks because they represent expectations about future changes in government spending.
structural shocks. Hence, the sign restriction approach also allows for scenarios where the
government acts quickly, such as with the TARP.

Fourth, we use generalized impulse response functions (GIRFs) proposed by Koop et al.
[1996] instead of orthogonalized impulse response functions (OIRFs) as suggested by Sims
[1980]. The OIRFs require the researcher to assume that the economy remains permanently
in the initial state after it is hit by an expansionary government spending shock. This is
true in ST-VAR applications where state variables have been fixed as well as in some local
projection applications where the implied transition between states is constant with respect
to identified shocks. The ST-VAR model benefits from the possibility that the state may
be made endogenous, i.e., it may be allowed to vary relative to policy shocks. This feature
can be implemented using the GIRFs.

Our analysis requires us to make several methodological contributions to the literature.
Building on prior work of Gefang and Strachan [2010], we fully estimate a ST-VAR model via
Bayesian estimation techniques. Our model is similar to that used by Gefang and Strachan
[2010], but features state-dependent heteroscedasticity. To improve further, we identify the
GIRFs via sign restrictions. While sign restrictions have been used for OIRFs in linear and
time-varying models, sign restrictions on GIRFs are new to the literature. Finally, we do
not only consider the effects of a government spending shock on the variables of interest but
also on the economic state itself. This consideration allows us to provide an explanation for
our multiplier results.

To investigate the source of the disagreement in the literature, we release the assumptions
mentioned above one at a time until we give maximum weight to the data. With this
approach, we can see if the finding of higher multipliers in recessions is truly a feature of
the data or whether it is a result of a particular identifying assumption. This procedure
helps to resolve the discrepancy in the literature. In the first exercise, we estimate the ST-
VAR model via Bayesian estimation methods but maintain the recursive identification and
the constant state assumption. We find that the spending multiplier is higher in recessions
than in expansions after one year and onward, confirming Auerbach and Gorodnichenko
In a second exercise, we replace the recursive identification by the sign restriction approach. Again, we find that the spending multiplier appears to be higher in recessions than in expansions. This is noteworthy, because results obtained by recursive identification may be fragile and may not be exist when weaker identification schemes such as sign restrictions are used. However, this does not seem to be the case in our context.

In another exercise, we release the constant state assumption by using GIRFs rather than OIRFs, allowing the economy to leave the initial state. We find that the expansion multiplier is higher than the recession multiplier in the short-run, which is in line with Bognanni [2013] and Alesina et al. [2018]. But this result is short-lived. In the medium- and long-run, we do not find any evidence for state-dependent multipliers. This evidence is in strong contrast to Auerbach and Gorodnichenko [2012] and others. Moreover, when we look at the response of the economic state to an expansionary government spending shock, we observe that the economy leaves its initial state rather quickly if it is hit during a recession and reaches a similar state after around one year as if it was hit during an expansion. If the economy appears to be in a similar state after around one year, then there should be little reason to believe that the multiplier is state-dependent in the medium- and the long-run. From the findings of this exercise, we conclude that the main finding of higher recession multipliers in the long-run is the result of the constant state assumption rather than a characteristic of the data. We also repeat this exercise for different government spending components and Ramey’s news defense shocks and find similar results in each case.

After establishing that the dominating finding of the literature is the result of an identifying assumption, we explore new territory. We use the multiplier estimates related to the different government spending components and compare them among three dimensions: government investment vs. government consumption spending, defense vs. non-defense spending and federal vs state and local spending. We find that the expansion and recession multipliers related to government investment, non-defense and state and local spending are higher than their government consumption, defense and federal spending counterparts more or less for the entire estimated horizon.

4We also find that very short-run multipliers are larger in expansions. This feature appears to be durable: it also appears in our GIRFs discussed later.
The paper proceeds as follows. Section 1.2 describes the model and the identification strategy; section 1.3 presents the results, and section 1.4 concludes.

1.2 Methodology

1.2.1 The Model

In order to treat the economic state as an endogenous variable, we employ the smooth-transition VAR model introduced by Auerbach and Gorodnichenko [2012]

\[
X_t = (1 - F(z_{t-1})) \Pi_E(\ell) X_{t-1} + F(z_{t-1}) \Pi_R(\ell) X_{t-1} + u_t. \tag{1.1}
\]

This model allows the economy to evolve as a linear combination of different regimes. We assume two regimes: an expansionary regime \( E \) and a recessionary regime \( R \). \( X_t \) is a \( n \times 1 \) vector of macroeconomic variables representing the underlying economy. In our baseline specification, \( X_t \) consists of real government consumption expenditures and gross investment, real government receipts net of transfers, real GDP, Ramey’s news defense shocks, the federal funds rate, and the GDP Deflator. Our sample range is from 1954Q3 to 2008Q4. \( \Pi_E(\ell) \) and \( \Pi_R(\ell) \) are two sets of VAR coefficients associated with regime \( E \) and regime \( R \), respectively. The residual vector \( u_t \) is the reduced form error and is assumed to be normally distributed with zero mean and covariance matrix \( \Omega_t \), e.g.

\[
u_t \sim N(0, \Omega_t) \tag{1.2}
\]

where the state-dependent heteroscedasticity follows

\[
\Omega_t = (1 - F(z_{t-1})) \Omega_E + F(z_{t-1}) \Omega_R. \tag{1.3}
\]

The state indicator function is called \( F \), where we employ a logistic sigmoid function:

\[
F(z_t) = \frac{\exp(-\gamma(z_t - c))}{1 + \exp(-\gamma(z_t - c))}. \tag{1.4}
\]
Thus, $F(z_t)$ is a continuous and monotone function, bounded between 0 and 1. The scalar $z_t$ is the state determining variable. We follow Auerbach and Gorodnichenko [2012] and choose the 7-month moving average real GDP growth rate as $z_t$. The threshold parameter $c$ determines the central "neutral" state where $F(z_t) = .5$. Then, as $z_t$ moves from below the threshold $c$ to above the threshold $c$, the dynamics of the model transition smoothly from $\Pi_R, \Omega_R$ to $\Pi_E, \Omega_E$. The rate of transition is governed by the parameter $\gamma$.

One useful feature of the logistic transition function is the fact that this smooth-transition model nests several other models commonly used. For example, when $\gamma = 0$, $F(z_t)$ is always equal to 0.5, and the ST-VAR model collapses to a traditional linear VAR model. On the other hand, when $\gamma \to \infty$, the economy jumps from one state directly to another and the ST-VAR model becomes a threshold VAR (T-VAR) model, which can also nest the Markov-type models.\(^5\)

In contrast to Auerbach and Gorodnichenko [2012], we will estimate the transition function via Bayesian estimation techniques. Thus, we can evaluate the full posterior distribution of the model and easily implement sign restrictions on impulse response functions. Bayesian methods require us to define prior distributions for all model parameters, which are then combined with the likelihood of the data to form the posterior distribution. This procedure is described in the next subsection.

The posterior distributions of the transition function allow the data to inform the dynamic structure of the model. In particular, we find that the parameter $\gamma$ has little posterior density near zero, yet has a mode well below our prior mean. This would seem to argue against either the linear or the threshold-type models. The estimated posterior density of $F(z)$ is bimodal near zero and one, which suggests that our two-state model is a reasonable approximation of the true data generating process.\(^6\)

---

\(^5\)See Krolzig [1998] for a discussion of how these fit into larger families as well.  
\(^6\)Posterior distributions for $c$, $\gamma$, and the median of $F(z)$ are shown in the appendix.
1.2.2 Priors

For the model expressed in (1.1), we need to specify prior distributions for $\Pi_E, \Omega_E, \Pi_R, \Omega_R, \gamma$ and $c$. First, we choose the conjugate Normal-Inverse-Wishart prior for $\Pi_s$ and $\Omega_s$ in both regimes. In particular, we assume

$$p(\Omega_s) = \mathcal{IW}(S_0, v_0)$$ (1.5)

$$p(\Pi_s|\Omega_s) = \mathcal{N}(\pi_0, \Omega_s \otimes N_0^{-1})$$ (1.6)

where $S_0, v_0, \pi_0$ and $N_0$ are hyperparameters, and $s \in \{E, R\}$. The choice of the Normal-Inverse-Wishart prior is convenient because it has the same functional form as the likelihood of the data: the corresponding posterior distributions are from the same family of distributions. Matlab has built-in functions to draw from a Normal and from an Inverse-Wishart distribution which are well-optimized, so it is very easy to sample from the corresponding marginal posterior distributions. This has some advantage over other Monte-Carlo Markov-Chain algorithms, which may converge slowly.

Next, we assume a Gamma prior for $\gamma$, i.e.,

$$p(\gamma) = \mathcal{G}(\mu_\gamma, \nu_\gamma)$$ (1.7)

where $\mu_\gamma$ is the mean and $\nu_\gamma$ is the degrees of freedom. We set them equal to 10 and 10, respectively. Finally, we assume that $c$ has a uniform prior, i.e.

$$p(c) = \mathcal{U}(u, \overline{u})$$ (1.8)

where $u$ and $\overline{u}$ represent the lower and upper limit, which we choose to be the first and ninth decile of $z_t$. These prior distributions are similar to those used for the homoscedastic model in Gefang and Strachan [2010].
1.2.3 Likelihood

Because \( u_t \) is normally distributed, the likelihood for the model in (1.1) is written as

\[
\mathcal{L} \propto \prod_{t=1}^{T} |\Omega_t|^{1/2} \exp \left\{ -\frac{1}{2} \sum_{t=1}^{T} u_t^t \Omega_t^{-1} u_t \right\}.
\]  

(1.9)

but can be expressed differently. The residual vector is \( u_t = X_t - (1 - F(z_{t-1}))\Pi_E(l)X_{t-1} - F(z_{t-1})\Pi_RX_{t-1} \). By writing

\[
W_t = [(1 - F(z_{t-1}))X_{t-1}, F(z_{t-1})X_{t-1}, ..., (1 - F(z_{t-1}))X_{t-p}, F(z_{t-1})X_{t-p}]
\]

we have \( u_t = X_t - \Pi'W_t' \) with \( \Pi = [\Pi_E(l), \Pi_R(l)]' \). Stacking over \( t \), we have

\[
Y = \begin{pmatrix} X_1' \\
\vdots \\
X_T' \end{pmatrix}, \quad W = \begin{pmatrix} W_1 \\
\vdots \\
W_T \end{pmatrix}, \quad U = \begin{pmatrix} u_1' \\
\vdots \\
u_T' \end{pmatrix}
\]

and the model may be written in the familiar form:

\[
Y = W\Pi + U.
\]  

(1.10)

The entire model is linear conditional on \( \{\gamma, c, \Omega_E, \Omega_R\} \). Thus, Auerbach and Gorodnichenko point out that the conditional MLE for \( \Pi \) is available by minimizing \( \sum_{t=1}^{T} u_t^t \Omega_t^{-1} u_t \) with respect to \( \Pi \). This yields \( \hat{\Pi} \) as

\[
\text{vec}(\hat{\Pi}) = \left( \sum_{t=1}^{T} \Omega_t^{-1} \otimes W_t'W_t \right)^{-1} \text{vec} \left( \sum_{t=1}^{T} W_t'X_t\Omega_t^{-1} \right).
\]  

(1.11)

This will be used to locate the mode of the likelihood and, later, the posterior mode. The covariance matrix is \( \Sigma = \text{diag}(\Omega_t) \). Using \( U \) in equation (1.10) and \( u = \text{vec}(U') \), the
likelihood (1.9) is proportional to

$$L \propto \prod_{t=1}^{T} |\Omega_t|^{1/2} \exp \left\{ -\frac{1}{2} \text{tr}(u'\Sigma^{-1}u) \right\}.$$  

Then \(\text{tr}(u'\Sigma^{-1}u)\) may be rewritten as

$$\text{tr}(u'\Sigma^{-1}u) = (\text{vec}(Y') - (W \otimes I)\text{vec}(\Pi'))'\Sigma^{-1}(\text{vec}(Y') - (W \otimes I)\text{vec}(\Pi'))$$

$$= s^2 + (\pi - \hat{\pi})'V^{-1}(\pi - \hat{\pi}). \quad (1.12)$$

In this formulation, \(s^2\) is \(\text{vec}(Y')'M_V\text{vec}(Y') = \hat{u}'\Sigma^{-1}\hat{u}\), where \(M_V\) is the weighted annihilator matrix corresponding to (1.11), i.e. the GLS annihilator matrix with weights corresponding to the variance \(V\). The covariance matrix \(V\) is given by either of the following representations

$$V^{-1} = (W' \otimes I)\Sigma^{-1}(W \otimes I)$$

$$= K \sum_{t=1}^{T} \Omega_t^{-1} \otimes W_t'W_t,$$

where \(K\) is the commutation matrix. A more explicit derivation of this result is given in the appendix. Ultimately, equation (1.12) allows us to write the likelihood as Gaussian with mean \(\hat{\pi}\);

$$L \propto \prod_{t=1}^{T} |\Omega_t|^{1/2} \exp \left\{ -\frac{1}{2} \left( s^2 + (\pi - \hat{\pi})'V^{-1}(\pi - \hat{\pi}) \right) \right\}. \quad (1.13)$$

And finally

$$L \propto |\Sigma|^{1/2} \exp \left\{ -\frac{1}{2} \text{tr}(\Sigma^{-1}\hat{u}\hat{u}') \right\} \exp \left\{ -\frac{1}{2} \text{tr} \left( (\pi - \hat{\pi})'V^{-1}(\pi - \hat{\pi}) \right) \right\}.$$  

There are two obvious results from this expression. First, we have the standard Normal-Inverse-Wishart representation of the model with an Inverse-Wishart distribution for \(\Sigma\) and a Gaussian distribution for \(\pi\). Second, the Inverse-Wishart form of the block-diagonal matrix \(\Sigma\) implies that the marginal likelihood of \(\Omega_E\) and \(\Omega_R\) also has an Inverse-Wishart
1.2.4 Posterior

Posterior conditional densities for $\Pi$ and $\Sigma$ are as follows. Conditional on $\Sigma, \gamma, c$, we have a Gaussian posterior for $\Pi$ which is proportional to

$$
\exp\left\{-\frac{1}{2}\text{tr}\left( (\pi - \hat{\pi})'V^{-1}(\pi - \hat{\pi}) \right) \right\} \exp\left\{-\frac{1}{2}\text{tr}\left( (\pi - \pi_0)'V^{-1}_o(\pi - \pi_0) \right) \right\}
$$

(1.14)

which is a standard textbook example with mean $\bar{\pi} = [\bar{\pi}_E', \bar{\pi}_R']' = \bar{V} (V^{-1}\hat{\pi} + V^{-1}_o\pi_0)$ and variance $\bar{V} = (V^{-1} + V^{-1}_o)^{-1}$. Integrating over $\Pi, \gamma, c$, the conditional posterior density for $\Sigma$ is proportional to expression (1.15).

$$
|\Sigma|^{-1/2} |\Omega_E\Omega_R|^{-(v_0+p+1)/2} \exp\left\{-\frac{1}{2}\text{tr}(\Sigma\hat{\mu}'\hat{\mu}') \right\} \exp\left\{-\frac{1}{2}\text{tr}\left( S_0\Omega^{-1}_E + S_0\Omega^{-1}_R \right) \right\}
$$

(1.15)

Because $\Sigma$ is parameterized entirely by $\Omega_E$ and $\Omega_R$, we may replace $\Sigma$ and expand expression (1.15) slightly.\footnote{See the appendix for a brief discussion.}

The posterior distributions for $\gamma$ and $c$ do not have a closed-form representation, but the Metropolis-Hastings algorithm of Chib and Greenberg [1995] can be used within the Gibbs sampler to draw $\gamma$ and the Griddy Gibbs sampler of Ritter and Tanner [1992] may be used to draw $c$. These two methods merely require comparisons between the posterior density at proposed values of their parameters. Each of these steps is described in detail in the online appendix.

We use a multi-move Gibbs sampler to sample from the posterior for the full model. The steps are shown in Algorithm (A.1).

1.2.5 Impulse Response Functions

Our core research question requires us to analyze how output (GDP) is expected to change in response to a government spending shock and whether the expected change differs between
Algorithm 1.1 Multi-Move Gibbs Sampler

1. Initialize: Choose $\pi^0_E, \pi^0_R, \Omega^0_E, \Omega^0_R, \gamma^0, c^0$;
2. Draw $\Omega_E|\pi_E, \pi_R, \gamma, c \sim IW(E_E'E_E, T)$ and $\Omega_R|\pi_E, \pi_R, \gamma, c \sim IW(E_R'E_R, T)$;
3. Draw $\pi_E, \pi_R|\Omega_E, \Omega_R, \gamma, c \sim N(\bar{\pi}, \bar{V})$;
4. Draw $\gamma|\pi_E, \pi_R, \Omega_E, \Omega_R, c$ using a Metropolis-Hastings step;
5. Draw $c|\pi_E, \pi_R, \Omega_E, \Omega_R, \gamma$ from a Griddy Gibbs sampler;
6. Repeat steps 2 through 5 and keep the desired number of draws after a burn-in phase.

expansions and recessions. This can be done via impulse response functions (IRFs). IRFs are broadly defined as

$$IRF(h, s) = E(X_{t+h}|s_t, \epsilon_t = \delta) - E(X_{t+h}|s_t, \epsilon_t = 0)$$ \hspace{1cm} (1.16)

where $X_{T+h}$ is the forecast of $X$ at horizon $h$ and state $s$. The impulse $\epsilon_t$ is a vector of structural shocks at time $t$, and $\delta$ represents the shock size.

Structural impulse response analysis in nonlinear models is more difficult than in linear models. Importantly, the impulse responses depend on the timing and size of the shocks. In our context, we must distinguish government spending shocks that occur during expansions from those that appear during recessions. Because of the nonlinearity, orthogonalized impulse response functions (OIRFs), as suggested by Sims [1980], do not carry the interpretation given in equation (1.16). However, they may be used to study the dynamics implied by each set of state coefficients, $\Pi_E, \Omega_E$ and $\Pi_R, \Omega_R$, separately. Orthogonalized impulse response functions are defined as

$$OIRF(h, s) = \Pi_s^h Chol(\Omega_s)Q$$ \hspace{1cm} (1.17)

where $OIRF(h, s)$ is a $n \times n$ matrix whose $i, j$-th element represents the response of variable $i$ in period $h$ and state $s$ to shock $j$. The matrix $Chol(\Omega_s)$ is the Cholesky decomposition of the state-dependent covariance matrix, and Q is an orthogonal matrix, i.e., $QQ' = I$, also called a rotation matrix. As further explained in the next section, the choice of $Q$ represents
the researcher’s identification strategy and requires some discussion.

The OIRFs are commonly used in the literature, e.g., by Auerbach and Gorodnichenko [2012], and are the standard type of IRFs for linear models. However, they are not representative of the behavior of the real economy in a nonlinear model, which varies between states. In our particular example, if an expansionary government spending shock affects the growth rate of real GDP, it will also influence the transition variable \( z_t \) and the economy may follow a different sequence of states than it would otherwise. To deal with this issue, we utilize the generalized impulse response functions (GIRFs) proposed by Koop et al. [1996]. These GIRFs use equation (1.18) to estimate the expression in equation (1.16) directly:

\[
GIRF(h, s) \approx \frac{1}{R} \sum_{i=1}^{R} X_{T+h}(u_{t,\lambda_i}, \lambda_i) - \frac{1}{R} \sum_{i=1}^{R} X_{T+h}(u_t, \lambda_i).
\]

The procedure to estimate the GIRFs is sketched in the next section and fully described in the appendix. In short, the GIRFs simulate the future path of the economy, with and without a structural shock, for comparison. The difference represents the dynamics of the variables that can be attributed to the structural shock. This method also allows for the possibility that the economy leaves its initial state: for each period \( h \), the GIRFs simulate the future trajectory of \( z_t \), and hence, of \( F(z_t) \). Therefore, the weight on the recessionary state is updated. In our model, this reweighting may be made endogenous, e.g., conditional on the response of GDP or unemployment, as in Caggiano et al. [2015] or Ramey and Zubairy [2018]. Finally, as explained in the next section, the GIRFs may also be identified via sign restrictions which is a novelty in this literature.

1.2.6 Identification

A fundamental challenge in VAR models is the identification of structural shocks. Macroeconomic theory assumes that the reduced form errors, \( u_t \), are a linear combination of structural shocks, \( \epsilon_t \),

\[
u_t = Chol(\Omega_t)Q\epsilon_t.
\]
The multivariate Gaussian likelihood of our model is not informative about the rotation matrix $Q$. Therefore, our choice of $Q$ represents our underlying identification strategy. We may either choose the Cholesky-type identification by setting $Q = I$, or draw from the set of possible rotation matrices $Q$ using some prior. The Cholesky matrix is lower diagonal, which restricts the contemporaneous effects of our variables on each other to be one-directional. Because government spending and output may affect each other contemporaneously, no such Cholesky ordering is feasible for this problem.

However, (1.19) implies that some linear combination of our identified shocks $Q\epsilon_t$ does have this world-ordering property. We can identify this using the sign restriction approach. Here, we reduce the set of possible $Q$ to only those that produce impulse responses with the correct signs following a shock. While sign restrictions are standard for the OIRFs, e.g., Uhlig [2005] or Rubio-Ramirez et al. [2010], they are new for the GIRFs.

The GIRF procedure can be summarized as follows. We begin by separating the data between expansions and recessions, where $F(z_t) > 0.8$ indicates a recessionary period. Then, we randomly draw a sequence of $H$ periods of $X_t$ and $u_t$ which is called a history, $\lambda_i$, and serves as the initial condition. Next, we consider two scenarios. First, we transform $u_{t,\lambda_i}$ into $\epsilon_{t,\lambda_i}$ using the inverse of (1.19), where $Q$ is drawn using some prior. Next, we add one unit to the first period of the government spending shock sequence, e.g., $\tilde{\epsilon}_{GS,1} = \epsilon_{GS,1} + \delta$, and transform $\tilde{\epsilon}_{t,\lambda_i}$ into $u_{t,\lambda_i}$ using equation (1.19). Lastly, we roll $X_t$ forward using (1.1) conditional on $u_{t,\lambda_i}$. For comparison, we also roll $X_t$ forward using (1.1) but using $u_{t,\lambda_i}$. The first scenario represents one future path of the economy with the shock and the second yields one future path of the economy without a shock. Taking the difference gives us one candidate realization of the generalized impulse response functions. If the realization of the GIRFs satisfies our sign restrictions, we keep it. If not, we discard it. To choose which draws are appropriate, we define an expansionary government spending as a shock that raises output, prices and government spending during the impact period and the four following periods. This is the same restricted horizon as in Mountford and Uhlig [2009].

Many researchers draw rotation matrices from a uniform prior on the Haar measure.
Baumeister and Hamilton [2015] point out that this prior is informative on impulse response functions, but a uniform distribution over rotations is correct if we have no other prior information about $Q$. In principle, this expands the exact identification of Cholesky—which does not produce the correct result—to a set identification strategy which does include the correct result in the identified set [Uhlig, 2005]. We give further comparisons between these approaches in section 1.3.1.

1.2.7 Government spending multiplier

The central object of interest in the literature is the government spending multiplier. It measures by how many dollars output changes if the government raises its spending by $1$. However, the literature does not provide a unique way to compute spending multipliers. First, Blanchard and Perotti [2002] propose using the peak multiplier defined as the maximum output response parameter over the initial government spending response parameter. This method only provides the multiplier at one particular point in time. In contrast, Auerbach and Gorodnichenko [2012] apply the sum multiplier which is defined as

$$\text{Multiplier}_h = \frac{\sum_{j=0}^{h} y_j}{\sum_{j=0}^{h} g_j} \times \bar{y}g.$$  \hspace{1cm} (1.20)

where $y_j$ and $g_j$ are the output and the government spending response parameter of period $j$. $\bar{y}$ is the sample mean of the real GDP to government spending ratio. This formula allows us to estimate the multiplier in the short-, medium- and long-run which is the main reason of why we prefer the sum over the peak multiplier.\(^9\)

1.3 Empirical Results

In Figure (1.1), we show the posterior distribution of the state indicator $F(z_t)$ over time. The median is shown in blue with the corresponding 80 percent credible bands in red. Grey

\(^9\)In addition, Mountford and Uhlig [2009] suggest using the present value multiplier defined as $\text{Multiplier}_h = \frac{\sum_{j=0}^{h} (1+i)^{-j} y_j}{\sum_{j=0}^{h} (1+i)^{-j} g_j} \times \frac{\bar{y}}{g}$, where $i$ is the sample mean of the federal funds rate. Because we want to compare our results to those of Auerbach and Gorodnichenko [2012], we choose the sum multiplier over the present value multiplier. The difference is negligible if $i$ is sufficiently close to zero.
Figure 1.1: Recession Probability $F(z_t)$ Over Time

Note: The estimation of $F(z_t)$ clearly identifies NBER recessions. Posterior 80 percent credible intervals are shown.

areas represent NBER recessions. Clearly, our model captures the evolution of the economy between growth and recession. Large values of $F(z_t)$ correspond with NBER recessions, and low values correspond with times of strong growth. The overall behavior of our estimated state indicator is similar to the calibrated regimes found in the literature; Caggiano et al. [2015], for instance.

The posterior contours of $F(z_t)$, as a function of $z$, are shown in Figure 1.2. The blue line shows the posterior median and posterior deciles are in light grey. For reference, we include the calibrated function used by Auerbach and Gorodnichenko [2012] as a dashed black line. Our estimated state indicator function agrees with the calibrated function in its extremities (especially high or low $z$) and in the center, which explains why Figure (1.1) is so visually similar to the corresponding figure made using the calibrated transition function.\footnote{See Auerbach and Gorodnichenko [2012] or Caggiano et al. [2015].}

The noticeable difference is the rate of transition between states, which is determined by the parameter $\gamma$. Our state indicator usually transitions more rapidly than the calibrated version, as shown by the steep slope of the posterior median in the center of the image. However, the calibrated state indicator of Auerbach and Gorodnichenko does not leave our
Note: The posterior deciles of $F(z_t)$ are shown as a function of $z$. The posterior median is quite close to the values chosen by Auerbach and Gorodnichenko in the extreme cases, but the speed of the transition differs.

80 percent credible band for any value of $z$.

1.3.1 Orthogonalized Impulse Response Functions

In the following, we estimate the model and identify government spending shocks, releasing assumptions one at a time. As a starting point, we employ Auerbach and Gorodnichenko’s approach but estimate the full posterior distribution of the model. We keep the recursive identification and maintain the constant state assumption. Then, we replace the recursive identification with the sign restriction approach. The orthogonalized impulse responses keep the state indicator $F(z)$ constant over time, as in Auerbach and Gorodnichenko [2012] and other studies. A constant $F(z)$ implies that the economy remains in the initial state after a shock occurs.

The recursive identification strategy used by Auerbach and Gorodnichenko [2012] sets $Q = I$. This assumption imposes a strict ordering on contemporaneous effects between variables in the VAR; shown in equation (1.21).
\begin{align*}
OIRF(0, s) &= Chol(\Omega_s) = \\
&\begin{pmatrix}
c_{11} & 0 & \cdots & 0 \\
c_{21} & c_{22} & \ddots & \vdots \\
\vdots & \vdots & \ddots & 0 \\
c_{n1} & c_{n2} & \cdots & c_{nn}
\end{pmatrix}
\end{align*} \tag{1.21}

More generally, we may choose \( Q \) to be set-identified such that the resulting set of impulse response functions satisfy certain restrictions. This sign-restricted approach is preferable if the restrictions in (1.21) are strong or implausible. \( (1.22) \) shows that all variables can contemporaneously respond to all structural shocks: the order of variables does not matter. Identification of shock \( j \) is achieved by restricting a subset of the \( d^j \) to be either positive or negative.

\begin{align*}
OIRF(0, s) &= Chol(\Omega_s)Q = \\
&\begin{pmatrix}
d_{11} & d_{12} & \cdots & d_{1n} \\
d_{21} & d_{22} & \cdots & d_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
d_{n1} & d_{n2} & \cdots & d_{nn}
\end{pmatrix}
\end{align*} \tag{1.22}

Multipliers estimated using recursive identification are shown in Figure 1.3 and sign-restricted multipliers are in Figure 1.4.\footnote{The corresponding sets of impulse response functions are shown in the appendix.} Figure 1.3 provides evidence that the recession multiplier is higher than the expansion multiplier, beginning one year after the shock. This finding is in line with Auerbach and Gorodnichenko \citeyear{Auerbach2012} and several others. The sign-restricted multipliers in the short-run in Figure 1.4 appear to be similar but the recession multipliers have a longer tail, see Table 1.1. The distributions start to fully diverge one year after the shock. For both identification strategies, we conclude that the recession multiplier is higher than the expansion multiplier at horizons over one year. Uhlig \citeyear{Uhlig2005} warns that results obtained by the recursive multiplier may be fragile or artefactual, but we find that the core result—larger recession multipliers at a five-year horizon—is not an artefact of the recursive restriction.\footnote{We find similar results using Ramey’s narrative identification strategy.}

\[11\]
1.3.2 Generalized Impulse Response Functions

In the last subsection, we demonstrated that the most common result in the literature—higher recession multipliers—does not hinge on calibrated model parameters or on recursive identification. However, when using the OIRFs to estimate the dynamic effects of an expansionary government spending shock on the economy, we enforce the restriction that the economy remains in the initial state after the shock occurs. Koop et al. [1996] show that this does not reflect the true impulse responses in nonlinear models, especially when the identified shock may affect the state variable. We remove the constant-state restriction by using GIRFs instead of OIRFs, which allows the economy to leave its initial state. As in the previous example, an expansionary government spending shock is identified via positive sign restrictions on output, prices, and government spending.

Figure A.12 reveals interesting results. In the short-run, the expansion multiplier seems to be higher than the recession multiplier. This finding is consistent with Bognanni [2013] and Alesina et al. [2018]. However, the difference starts to vanish after one year. In the medium- and long-run, recession and expansion multipliers appear to be the same,
Figure 1.4: Spending Multipliers using OIRFs and Sign Restrictions

Note: Figure shows the distributions of estimated multipliers for expansions and recessions across different time horizons. Using Orthogonalized impulse response functions and the recursive identification to estimate the multipliers, we find similar multiplier estimates in the short-run but they diverge in the long-run, supporting the findings of Caggiano et al. [2015] and Ramey and Zubairy [2018].

Table 1.1 places sign-restricted OIRF and GIRF multipliers side-by-side. This makes the difference between the two sets of multipliers clear. First, the OIRF procedure produces surprisingly large multipliers in the recession, especially at short horizons. In contrast, the GIRF multipliers produce smaller multipliers starting from $h = 1$. The ratio of expansion multipliers to recession multipliers is less than two and approaches unity as the horizon increases, which strongly differs from the OIRF results.

1.3.3 State Indicator Response Functions

Our results from generalized impulse response functions contrast sharply with those found using the constant-regime restriction in the OIRFs. The response of the economic state indicator variable can illuminate the source of the difference. Our GIRFs allow us to predict

\[13\] The right tail of the posterior distribution of the recession multiplier is longer than its expansion counterpart, resulting in a median multiplier that is larger for recessions. In contrast, the posterior mode is higher for expansions. Compare Table 1.1 with Figure 1.4.
Figure 1.5: Multiplier after various horizons - GIRFs

Note: Figure shows the distributions of estimated multipliers for expansions and recessions across different time horizons. Using GIRFs to estimate the multipliers, there is no evidence for state-dependence, neither in the short run nor in the long run.

\[ F(z_t) \] at any forecast horizon with or without the government spending shock. By doing this, we can examine how an expansionary government spending shock affects the evolution of \( F(z_t) \) and compare the effect across states. These results are displayed in Figure 1.6.

The two plots at the top of Figure 1.6 represent an economy that starts in an expansion, while the plots below represent an economy that starts in a recession. On the left, black lines represent the evolution of \( F(z_t) \) for the economies without a government spending shock. The dotted red and blue lines are the corresponding evolutions when a government spending shock occurs. The median and 68 percent probability intervals are shown. If the economy is hit by a government spending shock during an expansion, the shock decreases \( F(z_t) \) substantially. After one year, \( F(z_t) \) bounces back and fluctuates around a value of 0.5 onward. In contrast, if the shock hits during a recession, the shock lowers \( F(z_t) \) only moderately. The difference is small, but the dynamic response is similar: after one year, \( F(z_t) \) bounces back and fluctuates around a value of 0.5 moving forward. On the right, we plot the change of \( F(z_t) \) with and without the shock. There, we see that the government spending shock has a stronger impact on \( F(z_t) \) if it occurs during an expansion. This result
is similar to Alesina et al. [2018] who find that fiscal consolidation plans have a stronger impact on $F(\phi_t)$ during expansions than during recessions.

Figure 1.6 reveals more interesting details. Depending on the initial state of the economy, a shock influences the future state for one year after the shock. However, after that horizon, the economy is in a similar growth state regardless of whether it began in an expansion or recession. This points to the possibility that long-run multipliers should be similar regardless of the initial state. Our multiplier estimates from Figure A.12 support this hypothesis.\(^{14}\)

After comparing the different multipliers in this section, we find that the constant state assumption is responsible for the result from the literature—that the multiplier is higher during recessions. Once that assumption is released, multipliers appear to be very similar between expansions and recessions, except at very short horizons, and results are in line with Caggiano et al. [2015] and Ramey and Zubairy [2018]. Because Figure 1.6 shows that the economy leaves the recession in approximately one year (even if there is no shock), we view the constant state assumption as implausible.

\(^{14}\)We find similar state indicator response functions when we use different government spending variables, e.g., government investment, government consumption, defense, nondefense, federal or state and local spending as well as for Ramey’s news defense shocks. The shape and duration of the response appears to be robust.
Figure 1.6: Evolution of the state indicator function $F(z_t)$

Note: The economy invariably leaves the recession and returns to nominal expansion levels within 4-6 periods regardless of policy intervention. Interestingly, the effect of a policy shock on the state $F(z_t)$ differs by an order of magnitude between expansions and recessions. Both changes are extremely small and revert to zero after $h = 5$.

1.3.4 Components of Government Spending

Our GIRFs for total government expenditures (section 1.3.2) do not show significant differences between long-run spending multipliers in recessions and expansions. In this section, we first repeat our main exercise but use different government spending types and compare multipliers between expansions and recessions. As Tables 1.2 - 1.4 show, we find results similar to our baseline exercise: expansion multipliers are higher than recession multipliers in the short-run but multipliers appear to be similar in the long-run.\footnote{We also study the effects of large positive and negative shocks which confirm our baseline result. See the appendix for the corresponding multiplier figures.}

Because efficient expansionary fiscal policy should maximize the marginal effect of each dollar spent, it is also important to ask which types of spending yield the highest multipliers conditional on the initial state of the business cycle. We now compare the recession and expansion multipliers along three dimensions.
In Table (1.2), GIRF multipliers for government consumption are compared to those for government investment. Both are somewhat larger than the same multipliers for total government expenditures, found in Table (1.1). No matter when the economy is hit by a government spending shock, the multipliers for government investment are higher than those for government consumption at all estimated horizons. The difference shrinks substantially with the horizon $h$. This result is comparable to the result in Ilzetzki et al. [2013], who also find higher multipliers for investment in high-income countries, though the difference was not statistically significant. Our findings also supports the result of Ellahie and Ricco [2017], who find that multipliers related to government investment spending (defense, nondefense and state and local investment) are higher than those related to government consumption spending.

<table>
<thead>
<tr>
<th>Horizon</th>
<th>GIRF (Expansion)</th>
<th>GIRF (Recession)</th>
<th>GIRF (Expansion)</th>
<th>GIRF (Recession)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.90</td>
<td>2.20</td>
<td>11.04</td>
<td>4.68</td>
</tr>
<tr>
<td></td>
<td>[3.07,8.62]</td>
<td>[1.38,3.69]</td>
<td>[5.74,24.78]</td>
<td>[2.78,9.36]</td>
</tr>
<tr>
<td>4</td>
<td>3.62</td>
<td>2.50</td>
<td>5.59</td>
<td>3.88</td>
</tr>
<tr>
<td></td>
<td>[2.61,5.13]</td>
<td>[1.80,3.52]</td>
<td>[3.14,9.31]</td>
<td>[2.72,5.57]</td>
</tr>
<tr>
<td>8</td>
<td>3.91</td>
<td>2.45</td>
<td>4.58</td>
<td>3.39</td>
</tr>
<tr>
<td></td>
<td>[8.6,8.35]</td>
<td>[1.04,4.57]</td>
<td>[2.14,9.49]</td>
<td>[2.03,5.96]</td>
</tr>
<tr>
<td>12</td>
<td>2.68</td>
<td>2.07</td>
<td>3.71</td>
<td>3.04</td>
</tr>
<tr>
<td></td>
<td>[-2.29,7.62]</td>
<td>[-1.17,4.54]</td>
<td>[1.26,8.99]</td>
<td>[1.01,6.13]</td>
</tr>
<tr>
<td>16</td>
<td>2.15</td>
<td>1.83</td>
<td>3.40</td>
<td>2.83</td>
</tr>
<tr>
<td></td>
<td>[-2.80,7.33]</td>
<td>[-1.77,4.31]</td>
<td>[0.57,8.79]</td>
<td>[1.6,6.28]</td>
</tr>
<tr>
<td>20</td>
<td>1.80</td>
<td>1.66</td>
<td>3.20</td>
<td>2.65</td>
</tr>
<tr>
<td></td>
<td>[-2.74,6.29]</td>
<td>[-1.07,4.17]</td>
<td>[0.00,8.58]</td>
<td>[-63,6.35]</td>
</tr>
</tbody>
</table>

Table 1.2: Comparison between the SUM type multipliers for government consumption and government investment at different horizons; posterior 68% credible sets in brackets.

Similar results are seen in Table 1.3, where non-defense spending has a substantially larger multiplier than defense spending for both initial states. In that case, the difference is still noticeable at the longer horizon. This finding is consistent with the result of Barro and De Rugy [2013], who suggest that defense spending is an inefficient expansionary fiscal policy. It is worth noting that defense spending shocks are relatively well-identified (see e.g., Ramey and Shapiro [1998]) and are sometimes used as an instrument for overall federal spending shocks. The difference in the multiplier between defense and non-defense spending could be a difficulty for that strategy.
have a lower multiplier than non-defense spending. For example, states and local governments spend very little on defense, which we find to be a different mix of goods than the federal government. Differences in fiscal institutions between states induce different multipliers among the many state and local governments, decreases over time. Clemens and Miran [2012] argue that differences in fiscal institutions affect multipliers for federal spending. This is true for all estimated horizons whereas the difference in multipliers for state and local spending are higher than those for federal spending. This is documented in Table 1.3.

Table 1.3: Comparison between the SUM type multipliers for defense and non-defense spending at different horizons; posterior 68% credible sets in brackets.

<table>
<thead>
<tr>
<th>Horizon</th>
<th>GIRF (Expansion)</th>
<th>GIRF (Recession)</th>
<th>GIRF (Expansion)</th>
<th>GIRF (Recession)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.23</td>
<td>2.26</td>
<td>11.56</td>
<td>5.83</td>
</tr>
<tr>
<td></td>
<td>[2.48,8.31]</td>
<td>[1.45,3.84]</td>
<td>[6.93,21.92]</td>
<td>[3.82,9.29]</td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
<td>2.15</td>
<td>5.96</td>
<td>4.79</td>
</tr>
<tr>
<td></td>
<td>[2.91,6.34]</td>
<td>[1.19,3.20]</td>
<td>[3.61,9.68]</td>
<td>[3.43,6.85]</td>
</tr>
<tr>
<td>8</td>
<td>4.00</td>
<td>1.84</td>
<td>5.79</td>
<td>4.63</td>
</tr>
<tr>
<td></td>
<td>[2.79,9.51]</td>
<td>[1.72,3.81]</td>
<td>[2.93,11.02]</td>
<td>[2.66,7.93]</td>
</tr>
<tr>
<td>12</td>
<td>2.37</td>
<td>1.43</td>
<td>5.06</td>
<td>4.27</td>
</tr>
<tr>
<td></td>
<td>[-2.54,7.64]</td>
<td>[-42.34]</td>
<td>[1.51,11.15]</td>
<td>[1.01,8.48]</td>
</tr>
<tr>
<td>16</td>
<td>1.69</td>
<td>1.12</td>
<td>4.62</td>
<td>3.95</td>
</tr>
<tr>
<td></td>
<td>[-3.96,6.87]</td>
<td>[-66.36]</td>
<td>[0.27,11.16]</td>
<td>[.08,8.72]</td>
</tr>
<tr>
<td>20</td>
<td>1.31</td>
<td>1.04</td>
<td>4.18</td>
<td>3.82</td>
</tr>
</tbody>
</table>

Table 1.4: Comparison between the SUM type multipliers for federal versus state and local spending at different horizons; posterior 68% credible sets in brackets.

<table>
<thead>
<tr>
<th>Horizon</th>
<th>GIRF (Expansion)</th>
<th>GIRF (Recession)</th>
<th>GIRF (Expansion)</th>
<th>GIRF (Recession)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.21</td>
<td>1.98</td>
<td>7.20</td>
<td>4.65</td>
</tr>
<tr>
<td></td>
<td>[3.41,10.73]</td>
<td>[1.37,3.19]</td>
<td>[4.25,15.61]</td>
<td>[3.00,8.26]</td>
</tr>
<tr>
<td>4</td>
<td>3.85</td>
<td>2.09</td>
<td>5.61</td>
<td>5.71</td>
</tr>
<tr>
<td></td>
<td>[2.79,5.65]</td>
<td>[1.51,3.04]</td>
<td>[3.79,7.87]</td>
<td>[4.28,7.76]</td>
</tr>
<tr>
<td>8</td>
<td>3.82</td>
<td>1.99</td>
<td>5.14</td>
<td>5.73</td>
</tr>
<tr>
<td></td>
<td>[1.33,8.32]</td>
<td>[1.17,3.72]</td>
<td>[2.32,9.97]</td>
<td>[4.88,12.25]</td>
</tr>
<tr>
<td>12</td>
<td>2.70</td>
<td>1.66</td>
<td>4.32</td>
<td>4.30</td>
</tr>
<tr>
<td></td>
<td>[-1.65,7.69]</td>
<td>[43.37]</td>
<td>[-57.99]</td>
<td>[-2.34,12.29]</td>
</tr>
<tr>
<td>16</td>
<td>2.03</td>
<td>1.50</td>
<td>3.61</td>
<td>3.55</td>
</tr>
<tr>
<td></td>
<td>[-1.97,6.85]</td>
<td>[-0.7,3.78]</td>
<td>[-1.94,9.34]</td>
<td>[-4.43,11.66]</td>
</tr>
<tr>
<td>20</td>
<td>1.64</td>
<td>1.39</td>
<td>3.29</td>
<td>3.10</td>
</tr>
<tr>
<td></td>
<td>[-2.27,6.30]</td>
<td>[-25.37]</td>
<td>[-2.17,9.11]</td>
<td>[-4.01,11.01]</td>
</tr>
</tbody>
</table>

The multipliers in Table 1.4 compare federal spending with state and local spending. Table 1.4 documents that multipliers related to state and local spending are higher than those for federal spending. This is true for all estimated horizons whereas the difference decreases over time. Clemens and Miran [2012] argue that differences in fiscal institutions between states induce different multipliers among the many state and local governments, which is a reasonable explanation for our finding. Different states’ budget constraints and savings behavior could be the cause of the difference, though it is also certain that states and local governments spend money on a different mix of goods than the federal government. For example, states and local governments spend very little on defense, which we find to have a lower multiplier than non-defense spending.
Figure 1.7: Multiplier Comparisons: Recession

Note: The estimated multipliers for real non-defense spending and real state and local spending appear to be the strongest during recessions. Consumption and investment vary only slightly, while defense spending and federal spending have small multipliers.

For policy makers, an important question is what sort of spending has the highest multiplier in a recession. Figure 1.7 compares the recession multipliers for the six disaggregated variables from this section. The clear winners in both the short- and long-run are (1) non-defense spending and (2) state and local spending, which have the largest short-run multipliers and the largest multipliers at longer horizons. At the one-year horizon, state and local spending seems to be the best.

1.4 Discussion

We study whether government spending multipliers differ across the business cycle. Disparate results in the literature can be explained in part by the assumptions used to obtain them: the constant regime assumption has a noticeable effect on impulse response functions. Calibrated model parameters and timing restrictions (i.e. Cholesky) are less influential. To be as agnostic as possible, we replace those three restrictions with weaker assumptions:
Generalized Impulse Response Functions with endogenous regimes, fully-estimated model parameters, and sign restrictions as suggested by Uhlig [2005]. Thus, our methodology gives maximal weight to the data. Among these methodological improvements, the endogenous-regime GIRFs have the most substantial effect on estimated multipliers.

We find that multipliers do not differ in the medium- or long-run, which is sharply in contrast to much of the literature. Using orthogonalized impulse response functions with constant regimes, we find that the long-run multipliers are larger in recessions, similar to Auerbach and Gorodnichenko [2012]. Our generalized impulse response functions yield short-run multipliers that are larger in expansions which is consistent with Bognanni [2013] and Alesina et al. [2018]. However, the difference disappears completely after one year, which is in line with Caggiano et al. [2015] and Ramey and Zubairy [2018].

Our analysis also casts some light on the design of efficient counter-cyclical fiscal policy. None of our six major components of government spending have multipliers that vary with the business cycle, except in the short run. In all cases, short-run multipliers appear to be higher during expansions. However, long-run multipliers do vary across these spending components. Government investment has a higher multiplier than consumption. Non-defense spending has a higher multiplier than defense spending. State and local spending have a higher multiplier than federal spending. In particular, the highest median 5-year multipliers in recessions come from non-defense spending (3.82) and state and local spending (3.10). Thus, countercyclical government spending can be tailored to maximize growth. This is a promising topic for future research.

Using our methodology, it is also apparent that the state of the economy is the same one year after a recession, regardless of whether there is a countercyclical fiscal policy shock. This result supports the use of the generalized impulse response functions. Similarly, this result is consistent with our observation that recessions usually only last for about one year. If the goal of countercyclical policy is to treat the symptoms of the recession during the recession, then short-run multipliers deserve the most attention. Non-defense spending and government investment have the highest median one-quarter-ahead multipliers, at 5.83 and 4.68, respectively.
The key result for policymakers is that short-run multipliers vary more than long-run multipliers, and that the multiplier varies depending on how each dollar is spent. Others may wish to study the mechanisms that cause these result. Similarly, the differences between spending components should be studied in greater detail. There is much work to be done.
Chapter 2

Government Spending Between Active and Passive Monetary Policy

2.1 Introduction

Does the government spending multiplier depend on monetary policy? Conventional wisdom suggests that the multiplier is larger when monetary policy is passive. We show that this consensus misleads. Models supporting these predictions estimate multipliers while keeping the monetary policy rule constant after an increase in government spending. The shortcoming of that approach is that it fails to consider how the central bank adjusts its policymaking in response to the economic conditions that arise after the government raises its spending unexpectedly. We demonstrate that, in fact, the monetary policy rule itself changes quickly after a government spending intervention, and that it reaches a similar regime regardless of its initial conditions. This rapid change of monetary policy challenges the widely accepted link between the multiplier and monetary policy.

This paper proposes a new methodology to analyze this relationship empirically without constraining monetary policy after a government spending intervention. We estimate a Taylor rule with time-varying coefficients, and use its sequence of inflation parameters to inform the monetary policy regimes in a flexible, nonlinear, structural vector autoregression (SVAR) model. We find that the monetary policy regime varies substantially over time, and that the central bank changes its policy regime in response to economic conditions over

\footnote{When monetary policy is passive, the central bank raises nominal interest rates less than one for one to increasing inflation, and the real interest rate decreases. The lower real interest rate leads households to increase consumption. Consequently, output increases more than government spending and the multiplier is predicted to be larger than one. In contrast, when monetary policy is active, the central bank responds more than one for one to increasing inflation, and the multiplier is predicted to be smaller than one.}

\footnote{See e.g., Kim [2003], Canova and Pappa [2011], Woodford [2011], Christiano et al. [2011], Davig and Leeper [2011], Zubairy [2014], Nakamura and Steinsson [2014], Dupor and Li [2015], Leeper et al. [2017] and Cloyne et al. [2020].}
the sample period. In particular, the central bank becomes more active ("hawkish") when inflation is high, and more passive ("dovish") during recessions. This leads us to allow the central bank to update its monetary policy regime in response to future economic conditions when we estimate the dynamic effects of a government spending shock. We find that the central bank responds quickly after the shock, and that it transitions rapidly to an active regime (on average) even if the initial regime had not been active.

The response of the monetary policy regime has a fundamental impact on the multiplier. Once we account for the regime’s reaction to the government spending shock, we find little evidence that the initial monetary policy regime affects the multiplier. By contrast, when we force the monetary policy regime to remain constant after a shock, our results match those of theorists. This exercise leads us to conclude that the conventional wisdom is primarily driven by the constant-regime assumption, which itself is difficult to reconcile with the data. In fact, the relationship between the multiplier and monetary policy vanishes once we relax that assumption. Our analysis suggests that to boost the impact of government spending on the economy, the central bank would need to accommodate inflation for a long period of time. However, this tactic creates a dilemma, both because it would conflict with the central bank’s main mandate of maintaining price stability, and because it would violate the Taylor principle.

Our work analyzing the relationship between the multiplier and monetary policy makes several contributions. First, we analyze the interaction between fiscal and monetary policy empirically. This distinguishes our work from most of the literature, which either studies fiscal and monetary policy separately or the interaction of them theoretically. Questions about this interaction have become an important research topic, especially since the 2008 Financial crisis, when central banks and governments in many countries around the world worked closely together to fight the Great Recession.

Second, we combine the smooth-transition VAR (ST-VAR) model popularized by Auerbach and Gorodnichenko [2012] with the Taylor rule. This combination is essential because the ST-VAR model requires a state variable that governs the state of the economy. To identify the appropriate state variable for monetary policy, we refer to theory. When
theory makes its predictions about how the multiplier depends on monetary policy, it refers to the central bank’s responsiveness to inflation to distinguish between active and passive monetary policy regimes. To capture this responsiveness, we use the inflation parameter of a Taylor rule. Following empirical evidence that structural parameters of the real economy change over time, provided by Cogley and Sargent [2001, 2005] or Primiceri [2005], we estimate a Taylor rule with time-varying parameters. We then use the estimated inflation parameter to determine the states in the ST-VAR model. This allows us to estimate the path of monetary policy during our sample period. This is important because it helps us to design our multiplier analysis. Crucially, we find that the central bank changes its policy regime in response to economic conditions. For example, the U.S. Federal Reserve raised nominal interest rates aggressively in response to high inflation after Paul Volcker became its chairman, yet it lowered nominal interest rates in response to the recessions of 1957-58, 1990-91 and 2000-01.

These observations, which we capture precisely, raise the following question: When the central bank responds to economic conditions during our sample period, why should that be different after a government spending shock? Our third contribution is to analyze how the government spending multiplier depends on monetary policy while allowing the central bank to adjust its policy regime after a government spending shock. We find that the central bank responds quickly after the government spending shock, adopting promptly a similar policy regardless of the initial state of the economy: shortly after the shock and regardless of its initial condition, the central bank reacts actively to inflation. The response of monetary policy has a fundamental impact on the multiplier. When we account for the regime’s reaction, we find little evidence that the multiplier depends on monetary policy. By contrast, when we force the monetary policy regime to remain counterfactually constant after a shock, we find multiplier estimates that are consistent with the theoretical consensus.

Fourth, we identify a government spending shock using sign restrictions on impulse response functions. We define a government spending shock as a shock that drives up output, inflation, government spending, government tax revenue, and government debt. These restrictions represent joint predictions of theoretical models, and hence, are largely
uncontroversial. Most of the empirical literature relies on zero restrictions related to the standard Cholesky approach, with scant theoretical foundations [Uhlig, 2005]. We find that, regardless of the monetary policy regime at the time of a shock, our posterior median estimate for the multiplier is almost five in the short run, notably above most estimates in the literature. Our multiplier estimates then decline to about one after five years.

To further explore the implications of our main findings, we conduct counterfactual exercises. We first analyze what happens to the multiplier if the central bank were to keep its monetary policy regime temporarily constant after a shock. We repeat our main exercise but keep the monetary policy regime constant for one, two, and five years. In another exercise, we replicate the framework that underlies the conventional wisdom with our empirical model. To do this, we distinguish between the two most extreme monetary policy regimes, and we fix the regimes for the entire horizon after the shock. These exercises show that the multiplier may depend on monetary policy, but only if the central bank keeps its initial policy regime constant for a surprisingly long period of time after the shock (more than two years, according to our estimates). Because the central bank usually responds quickly after a government spending shock, we conclude that the key driver of this conventional wisdom is the constant-regime assumption — not the data.

The sole exception to our core finding that the monetary policy regime itself responds to a government spending shock occurs when nominal interest rates are stuck at zero in response to a severe economic downturn, such as the 2008 financial crisis. Between 2008Q4 and 2015Q4, the Fed held nominal interest rates at zero, and it did not change its (conventional) monetary policy regime, in spite of unprecedented fiscal policy interventions such as the American Recovery and Reinvestment Act (ARRA) of 2009. This leads us to analyze the multiplier at the zero lower bound separately. The scarcity of data poses a challenge: other authors ignore this period entirely (see e.g. [Leeper et al., 2017, Arias et al., 2019]). We estimate a VAR model for the period between 2008Q4 and 2015Q4. When we use our sign restriction approach for identification, we find multiplier estimates that exceed those reported in the empirical literature. Though imprecise, the posterior median estimate is 5.3 on impact. Five years later, it falls to 4.1. The timing restrictions related to the standard
Cholesky approach are clearly violated at the zero lower bound. The Cholesky approach assumes that governments change their spending in response to shocks to the business cycle only with a delay. Contrary to this premise, however, recent events such as the Troubled Asset Relief Program (TARP) of 2009 or the Coronavirus Aid, Relief and Economic Security (CARES) Act of 2020 reveal that governments can and do react quite quickly during crises, for example, when central banks cut nominal interest rates to zero in response to a deep economic recession. By contrast, the sign restriction approach reflects joint predictions of theoretical models, and allows for a contemporaneous government (spending) response. These final results provide some evidence that the multiplier at the modern zero lower bound in the United States is larger than has previously been found, and that recent fiscal stimulus packages, such as the ARRA and the CARES Act, might also have larger economic effects than the literature suggests.

This paper proceeds as follows: Section 2.2 introduces the methodology. Section 2.3 describes the evolution of monetary policy in the United States. Sections 2.4 and 2.5 present our main results and our counterfactual analysis. Section 2.6 includes our zero-lower bound exercise. Section 2.7 concludes.

2.2 Model

2.2.1 Model

This paper studies how the government spending multiplier depends on the responsiveness of monetary policy to inflation. To allow for policy-dependent multipliers, we use the following smooth transition VAR (ST-VAR) model:

\[
X_t = G(z_{t-1})\Pi_{AM}X_{t-1} + (1 - G(z_{t-1}))\Pi_{PM}X_{t-1} + u_t
\]

\[
u_t \sim N(0, \Omega_t)
\]

\[
\Omega_t = G(z_{t-1})\Omega_{AM} + (1 - G(z_{t-1}))\Omega_{PM}
\]
where $X_t$ is a vector of endogenous variables that represents the economy. $X_t$ consists of real government spending, real tax receipts, real GDP, Ramey’s news shocks, GDP inflation, the federal funds rate, and real government debt for the U.S. economy from 1954Q3 to 2007Q4. We include Ramey’s news shocks to account for the issue of fiscal foresight.

Equation (B.1) says that the economy $X_t$ evolves as a convex combination of two ideal monetary policy regimes: “purely active” monetary policy $(AM)$ and “purely passive” monetary policy $(PM)$ regimes. The state-determining variable $z_t$ characterizes the underlying state of the economy. To describe the evolution of monetary policy over time, we estimate a Taylor rule with time-varying coefficients. We then use the five-quarter moving average of the pointwise posterior median estimate of the inflation parameter as $z_t$. Hence, our choice of $z_t$ captures the central bank’s responsiveness towards inflation over time. Further details about our choice of $z_t$ can be found in Section 2.2.2. The transition function $G(\cdot)$ governs the transition between the $AM$ and $PM$ regimes. $G(\cdot)$ transforms $z_{t-1}$ into a value between zero and one, and represents our measure of monetary policy activism. This is because larger values of $G(\cdot)$ are equivalent to a more aggressive approach by the central bank towards inflation. In Equation (B.1), $u_t$ is a vector of reduced-form residuals. We assume that the residuals are normally distributed with zero mean and a time-varying covariance matrix $\Omega_t$. The covariance $\Omega_t$ is assumed to evolve as a convex combination of the pure regime covariance matrices, $\Omega_{AM}$ and $\Omega_{PM}$.

Our ST-VAR model approximates the variation in monetary policy over time as a smoothly evolving convex combination of the two pure $AM$ and $PM$ regimes. In each
period, $G(z_{t-1})$ represents the relative weight on the $AM$ regime and is given by Equation (2.4). We follow Auerbach and Gorodnichenko [2012] and Caggiano et al. [2015] and assume that $G(z_{t-1})$ is a logistic sigmoid function. Equation (2.4) says that $G(z_{t-1})$ is continuous, monotonically increasing, and bounded between zero and one. The transition function $G(z_{t-1})$ depends on the transition parameter $\gamma$, the state variable $z_t$, and on the threshold parameter $c$. As the variable $z_t$ moves from below $c$ to above $c$, the value of $G(z_{t-1})$ increases, and the model puts relatively more weight on the $AM$ regime. The rate of this transition between regimes is determined by $\gamma$. If $\gamma \approx 0$, then the ST-VAR model collapses to a linear VAR model, and no transition occurs. Conversely, if $\gamma \approx \infty$, then the model jumps directly from the $PM$ to $AM$ regime as soon as $z_t$ surpasses $c$. Thus, the ST-VAR model nests a threshold-VAR model, which also nests Markov-type models [Krolzig, 1998]. For any value of $\gamma$ between those two extremes, the transition between the two extreme regimes is smooth.

The literature suggests that monetary policy activism has varied substantially over time (i.e. Taylor [1999], Romer and Romer [2004]). To allow the data to speak freely about our measure of monetary policy activism, $G(z_{t-1})$, we use fully Bayesian methods to estimate the joint posterior distribution of the model. This approach lets the data inform the model structure – particularly for the state-transitioning parameters $\gamma$ and $c$. This would not be possible if those parameters were calibrated. First, we define prior distributions for all model parameters, i.e. $\Pi_{AM}, \Pi_{PM}, \Omega_{AM}, \Omega_{PM}, \gamma$ and $c$. Second, because the joint posterior distribution is analytically untractable, we employ a multi-move Gibbs sampler as proposed by Galvão and Owyang [2018] to let the data update our prior beliefs. The Gibbs sampler partitions the vector of model parameters into different groups. Then, the Gibbs sampler generates draws for each group separately from the corresponding marginal posterior distributions, conditional on the remaining parameters and the data $X_t$. This procedure simplifies the drawing process because the marginal posterior distributions are

---

7Auerbach and Gorodnichenko [2012], for instance, use an ST-VAR model similar to ours to explore how the government spending multiplier varies across the business cycle. Though they consider a fully-estimated model (see footnote 10 in that article), they instead present results based on maximum-likelihood estimation with calibrated parameters. Bayesian regularisation is helpful in models with a high degree of collinearity, such as these [Dormann et al., 2013], and has become a preferred approach. Logistic smooth-transition models, including the ST-VAR, have taken many forms; see Hubrich and Teräsvirta [2013]. However, Galvão and Owyang [2018] appear to have the first heteroscedastic ST-VAR model with fully-estimated posterior inference for (generalized) impulse responses.
often known and easy to sample from. Finally, the draws from the marginal posteriors approximate the joint posterior of the model.\footnote{See the appendix for a detailed description of our Bayesian Sampler.}

In our context, it is crucial that Bayesian methods allow the data to be informative about the transition parameters $\gamma$ and the threshold parameter $c$, which, in turn, inform our beliefs about the evolution of monetary policy. Conditional on the marginal posterior distributions of $\gamma$ and $c$, we compute the posterior distribution of $G(z_{t-1})$ as a function of $z_t$. The posterior of $G(z_{t-1})$ reveals that the monetary policy regime behaves differently from how theory models it. Figure 2.1 illustrates how the monetary policy regime evolves continuously over time and changes in response to inflation and recessions.\footnote{We describe the evolution of monetary policy in more detail in Section 2.3.}

\section{State-Determining Variable $z_t$}

The key ingredient of the smooth-transition VAR model, equations (B.1) - (2.4), is the state-determining variable $z_t$. The choice of $z_t$ characterizes the underlying nonlinearity of the economy. The standard approach in the literature is to choose a single observable variable as $z_t$. For example, to replicate the business cycle of the U.S. economy, Auerbach and Gorodnichenko [2012] use the growth rate of real GDP, and Ramey and Zubairy [2018] use the unemployment rate to distinguish between times of high and low slack. In our case, $z_t$ must embody economically meaningful and empirically informative attributes about the monetary policy regime, i.e., the central bank’s responsiveness towards inflation.

However, the central bank’s responsiveness towards inflation is not determined by any single variable, and there is no unique measure of it. This leads us to refer to theory. When theorists make predictions about how the effect of government spending depends on monetary policy, they typically refer to the central bank’s responsiveness to inflation as the determinant that distinguishes between the monetary policy regimes [Leeper et al., 2017]. When monetary policy is active, the central bank raises nominal interest rates more than one for one to increasing inflation. As a result, the real interest rate rises, inducing households to reduce consumption. Consequently, output increases, but it increases by less
than government spending, so theory predicts the government spending multiplier to be less than one. When, instead, monetary policy is passive, the central bank reacts only weakly to inflation, and real interest rates fall. As a result, households increase consumption. In this case, theory predicts the government spending multiplier to exceed one. We believe that this measure of policy responsiveness influences the dynamics of the real economy.

Since Taylor [1993], theorists have typically used the Taylor rule to model monetary policy. According to the Taylor rule, the central bank sets its policy instrument in response to inflation and output. Because ample evidence exists that monetary policy has changed over time, we estimate the following ex-post Taylor rule with time-varying parameters:

\[
i_t = c_t + \phi_{\pi,t}\pi_t + \phi_{y,t}y_t + v_t, \ v_t \sim N(0, \sigma_t^2) \tag{2.5}
\]

where \(i_t\) is the federal funds rate, \(\pi_t\) inflation and \(y_t\) is the growth rate of real GDP. \(c_t\) is a time-varying constant, and \(\phi_{\pi,t}\) and \(\phi_{y,t}\) capture the central bank’s time-varying responsiveness to inflation and output, respectively. Here, \(\sigma_t^2\) corresponds to the variance of the residuals that can also vary over time. We then use \(\phi_{\pi,t}\) to distinguish between different monetary policy regimes in each quarter of our sample period: when \(\phi_{\pi,t}\) exceeds one, the central bank raises nominal interest rates by more than one for one to increasing inflation, i.e., monetary policy regime is active in that period. By contrast, when \(\phi_{\pi,t}\) is smaller than one, the central bank responds by less than one for one to increasing inflation, i.e., the monetary policy regime is passive in that period.

Even though the literature has documented extensive evidence suggesting that the Taylor rule has varied substantially over time, the exact nature of this variation is unclear. To avoid imposing any specific structure, we assume that \([c_t, \phi_{\pi,t}, \phi_{y,t}]\) and \(\log(\sigma_t)\) follow Random Walks. The Random Walk assumption implies that the parameters in period \(t\) are equal to the parameters in period \(t - 1\), on average. However, our estimated parameters can change over time if the data argue strongly that they do. Using this assumption, the model in (2.5) can be written with a state-space representation:

\[^{10}\text{See e.g., Taylor [1999], Clarida et al. [2000], Cogley and Sargent [2005], Primiceri [2005], Boivin [2006], Wolters [2012] or Carvalho et al. [2019].}\]
\[ i_t = z_t' \phi_t + e_t, \quad e_t \sim N(0, \sigma_t^2) \]  
(2.6)

\[ \phi_t = \phi_{t-1} + \nu_t, \quad \nu_t \sim N(0, \eta) \]  
(2.7)

\[ \log(\sigma_t) = \log(\sigma_{t-1}) + \xi_t, \quad \xi_t \sim N(0, \omega) \].  
(2.8)

We employ Bayesian methods to estimate (2.6) - (2.8). We then use the five-quarter moving average of the pointwise posterior median estimate of \( \phi_{\pi,t} \) as \( z_t \) to determine the monetary policy regime for each quarter of our sample period.\(^{11}\) Therefore, our model is not the standard smooth-transition VAR model as in Auerbach and Gorodnichenko [2012], but involves a prior step in which we estimate the state variable as a parameter with time-varying coefficients. We refer to our model as smooth-transition VAR model with time-varying parameters (TVP-ST-VAR model). Our model extends the standard ST-VAR model to settings in which the underlying state of the economy is not well-captured by a single observable variable but can be characterized via a parameter that varies over time, as is the case with monetary policy.

The estimation of Taylor rules has received substantial prior attention in the literature, which guides the design of our model. In the appendix, we review the literature, and conduct a battery of robustness checks. We now describe the details of our baseline specification of (2.5).

First, we use current inflation and the current growth rate of real GDP as regressors. These have been criticized for endogeneity concerns [Clarida et al., 2000], but de Vries and Li [2013] argue that usual instruments (i.e., lagged regressors, real-time data or the instrument set of Clarida et al. [2000]) are not fully exogenous if the residuals of the Taylor rule are serially correlated; at the same time, however, they insist that the resulting bias will be small. In addition, Carvalho et al. [2019] estimate the Taylor rule before and after 1980 using ordinary least squares (OLS) and different instruments and uncover comparable estimates. We do find that the residuals of our baseline specification are serially

\(^{11}\)In the appendix, we change the lengths of the moving average to zero, three and sixth. In addition, we also consider the five-quarter moving average of the pointwise posterior 5th and 95th percentiles of \( \phi_{\pi,t} \) as \( z_t \). Our main results are robust to these alternative specifications.
correlated (see the appendix). Moreover, we find that the estimate of $\phi_{\pi,t}$ of our baseline specification is comparable to those of specifications with lagged regressors included to eliminate autocorrelation in the residuals, the strategy preferred by Clarida et al. [2000]. Ultimately, each method captures a similar pattern of variation over time. And, because our transition function contains estimated $\gamma$ and $c$, the choice of the state variable is only unique up to an affine transformation—a small bias in our Taylor rule coefficients may not affect the model’s regimes at all.

Second, Orphanides [2001, 2002, 2004] and Boivin [2006] advocate using real-time data rather than revised data because revised data may contain information that was not available when policy makers made their monetary policy decisions. We use revised data in our baseline specification of (2.5), because we are more interested in the interaction between policy and the economy, and less interested in how policy decisions are made. For example, a central bank’s decision can turn out differently than had been intended, and this can only be seen using revised data. In contrast, real-time data are likely to be more useful in capturing the central bank’s intentions. This consideration drives our baseline design.

Third, we use the growth rate of real GDP to measure output rather than the output gap as suggested by Taylor [1993]. As we detail in Section 2.2.4, we allow the central bank to adjust its monetary policy regime after the government spending shock. This idea requires us to update the inflation parameter in our forecasts of the real economy. Hence, we must include all variables of the Taylor rule in the VAR part of our model. Because the impulse response of output is a key ingredient in the formula for the multiplier, we measure output using the growth rate of real GDP instead of using the output gap.

In the appendix, we show that the estimates of $\phi_{\pi,t}$ using both revised and real-time data are very similar. A battery of alternative specifications leads us to conclude that the variables selected for our Taylor rule do not substantially affect the outcome. Even to the extent that estimates of $\phi_{\pi,t}$ vary slightly, we find that the estimated $G(z)$ is substantially the same in every case. We conclude that the monetary policy episodes described by Romer and Romer [2004] are not only well-documented by historical record, but that they are clearly visible in macroeconomic data. See 2.3 for discussion.

The choice of output growth rather than the output gap is reasonably common in the literature. For example, Orphanides and Williams [2002] argue that interest rate rules based on output growth are preferable to those based on the output gap of models in which the natural rate of unemployment is uncertain. Walsh [2003] shows that loss functions based on output growth rather than the output gap result in lower losses. Finally, Cubbin and Gorodnichenko [2011] illustrate that the Taylor principle breaks down under positive trend inflation, but that it can be restored if the central bank responds to output growth rather than to the output gap.
Finally, the Taylor rule in (2.5) can be contrasted with our model of the economy in (B.1), which also contains an equation for the nominal interest rate. Though the simple Taylor rule is nested in the larger model, the two equations do not embody the same information about the economy. While even simple versions of the Taylor rule have been shown to successfully approximate the central bank’s policy instrument [Taylor, 1993], the nominal interest rate equation of (B.1) includes variables that are typically not included in the Taylor rule (e.g., fiscal variables, and several lags of all endogenous variables). In addition, the current response parameters in Equation (B.1) are included in the impact matrix, which is not statistically identified. Because the Taylor rule is the most widely accepted way to model monetary policy activism, we use $\phi_{\pi,t}$ from equation (2.5) as the state variable that distinguishes between different monetary policy regimes in our analysis. This choice ensures that we capture a recognized component of monetary policy behavior without imposing any additional restrictions on the overall economy in Equation (B.1).

### 2.2.3 Generalized Impulse Response Functions

Impulse responses in nonlinear models depend on the impulse’s sign, size, and timing. In addition, the state of the economy can respond to economic conditions – both during the sample period and after shocks. To uncover how the impact of a government spending shock depends on monetary policy, and how monetary policy itself changes, we must consider the initial state of the economy and the dynamics that arise from both the direct effect of a shock on the variables and its indirect effects via the future evolution of the monetary policy rule after the shock.

To incorporate these features, we follow Koop et al. [1996], and employ generalized impulse response functions. Generalized impulse response functions are defined as the expected difference between two simulated paths of the economy. Formally, they can be written as
\[ GIRF(h) = E[(1 - G(z_{t+h-1})\Pi A \hat{X}_{t+h-1} + G(z_{t+h-1})\Pi P \hat{X}_{t+h-1} + \epsilon_{t+h}] 
- E[(1 - G(z_{t+h-1})\Pi A \hat{X}_{t+h-1} + G(z_{t+h-1})\Pi P \hat{X}_{t+h-1} + u_{t+h}] \]  

(2.9)

The first part of (2.9) represents the simulated path of the economy hit by a government spending shock, \( \hat{X}_{t+h} \). The second part corresponds to the simulated path when the economy is not hit by the government spending shock, \( \hat{X}_{t+h} \).

The generalized impulse response functions require initial conditions from a starting period. Given a particular starting period, we use (B.1) to roll the model \( H \) periods forward in both simulations. We estimate the effects of a government spending shock for different monetary policy regimes by choosing starting conditions that correspond to a particular policy regime. In Section 2.4, we divide monetary policy regimes into quintiles to distinguish between different monetary policy regimes (e.g., between “very active,” “weakly active,” “neutral,” “weakly passive,” and “very passive” regimes). Then, we randomly draw the initial conditions from these quintiles, and present the average difference between two simulations as in (2.9) for each quintile.

### 2.2.4 Updating Rules for Monetary Policy

In nonlinear models such as ours, the state of the economy can change in response to its economic environment, such as in response to shocks. In Section 2.3, we establish that the central bank changes its monetary policy regime frequently in response to inflation and recessions. The generalized impulse response functions also allow for scenarios in which the underlying state of the economy changes after the shock. Based on the forecasts for the economies with and without a government spending shock, \( X_{t+h}^\epsilon \) and \( X_{t+h}^u \), we can forecast the corresponding values of \( z_{t+h} \) and then of \( G(z_{t+h}) \). The new values of \( G(z_{t+h}) \) and \( 1 - G(z_{t+h}) \) represent the weights assigned toward the purely active and purely passive monetary policy regimes \( h \) periods after the shock. The updated weights then enter the forecasts of the economies with and without the government spending shock in the following periods, \( X_{t+h+1}^\epsilon \) and \( X_{t+h+1}^u \).
and $X^u_{t+h+1}$. In particular, the monetary policy regime can change after the government spending shock, thereby affecting how the government spending shock propagates through the economy.

To implement this feature, we run the Kalman filter based on the forecasted values of the federal funds rate, inflation rate, and output growth of $X^r_{t+h}$ and $X^u_{t+h}$ from equation (2.9). This is equivalent to the model originally used to estimate $\phi_{\pi,t}$, as described in Section 2.2.2. Each time we predict the economy one step ahead, the Kalman filter obtains new values of $\phi_{\pi,t+h}$ and $\phi_{y,t+h}$. At each step, we use the average of $\phi_{\pi,t+h}$ and its four previous values as $z_{t+h}$ to update the weights of the two monetary policy regimes.\textsuperscript{14} This method requires the Taylor rule to use the same information set as the rest of our model, in order to be consistent with the initial estimation of $\phi_{\pi,t}$. These updating rules reflect average behavior according to the historical data; the Federal Reserve relies on its own discretion when updating its policy.

2.2.5 Identification

Identification of structural shocks in our nonlinear model is similar to identification in linear models. The reduced-form errors, $u_t$, do not have an economic interpretation. Macroeconomic theory assumes that the reduced-form errors are linear combinations of structural shocks $\epsilon_t$. Formally,

$$u_t = A\epsilon_t,$$  \hspace{1cm} (2.10)

where

$$A = Chol(\Omega_t)Q.$$  \hspace{1cm} (2.11)

$Chol(\Omega_t)$ is the Cholesky decomposition of the covariance matrix $\Omega_t$, and $Q$ is an orthogonal

\textsuperscript{14}In the appendix, we also consider alternative updating rules. First, we update $\phi_{\pi,t}$ via an updating regression based on the forecasted values of the federal funds rate, inflation rate, and output growth of $X^r_{t+h}$ and $X^u_{t+h}$ from equation (2.9). Second, we include $\phi_{\pi,t}$ in $X_t$ and treat it as an observation variable. Third, we do not update $z_t$ at all and keep the monetary policy regime constant for the entire time after the shock (see also Section 2.5). Fourth, we update $z_{t+h}$ with an AR(1) process with a persistent parameter equal to 0.95.
matrix, $QQ' = I$. The matrix $Q$ is not identified statistically; consequently the matrix $A$ is also not identified. To identify the structural shocks, we must impose economically meaningful restrictions on $A$.

We identify government spending shocks using sign restrictions on impulse response functions. This strategy has been used in the literature by Mountford and Uhlig [2009], Canova and Pappa [2011] and Laumer [2020]. We follow Laumer [2020] and define a government spending shock as a shock that drives up output, inflation, government spending, government tax revenue, and government debt. Each of these restrictions is backed by a large set of neoclassical and new Keynesian models (e.g., Woodford [2011]), and none of them is controversial. We assume that this set of restrictions holds regardless of the monetary policy regime at the time of the shock.

The sign-restriction strategy has major advantages over other methodologies because sign restrictions impose only weak restrictions on the behavior of the economy. We only restrict the signs of certain impulse responses. Thus, sign restrictions allow all variables to react contemporaneously to all shocks. In contrast, other strategies (e.g., the Cholesky approach) impose specific timing restrictions that govern which variables can react to the shocks on impact, and as Uhlig [2005] underscores, the set timing restrictions may not be valid. Sign restrictions in linear VAR models have been widely applied and are well documented (see Baumeister and Hamilton [2018] for an overview). Sign restrictions on nonlinear impulse response functions have been introduced to the literature only recently, see e.g., Shin and Zhong [2020] or Laumer and Philipps [2020]. Here, we follow the approach in Laumer and Philipps [2020] which develops sign restrictions on generalized impulse response functions for the ST-VAR model.\footnote{We provide further details in the appendix.}

### 2.3 History of Monetary Policy

In this section, we describe how monetary policy has evolved over our sample period. We use quarterly data for the U.S. economy from 1954Q3 to 2007Q4. In Section 2.6, we extend
the sample period to 2015Q4 to include the zero lower bound epoch. Figure 2.1 plots the median estimate of $G(z_{t-1})$, the weight assigned to the purely active monetary policy regime in Equation (B.1), along with the 68 percent credible bands. We interpret $G(z_{t-1})$ as a measure of the central bank’s activism because larger values of $G(z_{t-1})$ occur when the central bank responds more aggressively to inflation. Low values of $G(z_{t-1})$ indicate a more “passive” monetary policy regime.

Figure 2.1 reveals that the monetary policy regime varies substantially over time. Our estimate of $G(z_{t-1})$ suggests that monetary policy was very passive during the second half of the 1950s, while the model puts some weight on the purely active monetary policy regime during the 1960s. In the 1970s, the monetary policy regime was mostly very passive, only interrupted by the summer of 1970 and the winter of 1974/75. After Paul Volcker became the chairman of the Federal Reserve in August 1979, our model indicates that monetary policy changed dramatically. Between 1979 and 1982, monetary policy transitioned from a very passive regime to a very active regime, and it remained very active throughout the 1980s. Then, during the first half of the 1990s, our model assigns a low weight to the active regime, while it assigns relatively high weight to the active regime during the second half. Since 2002, monetary policy has been almost “purely” passive, save for the very end of our sample period.

Narrative evidence from Romer and Romer [2004] lends support to our results regarding $G(z_{t-1})$. Romer and Romer [2004] describe how monetary policy was very passive during the late 1950s before the Federal Reserve increased nominal interest rates in response to high inflation in 1959 (Martin Disinflation). Moreover, Romer and Romer [2004] also illustrate that the Fed ran a very passive monetary policy throughout most of the 1970s, only interrupted in 1970 and the winter of 1974/75. Subsequently, the Fed raised nominal interest rates aggressively in response to high inflation. This occurred around the time of Paul Volcker’s appointment in August 1979 (Volcker Disinflation). In addition, Romer and Romer [2004] characterize the chairmanships of Paul Volcker (1979-1987) and Alan Greenspan (1987-2006) as periods in which the Fed responded actively to inflation. These periods were only interrupted during the first half of the 1990s and 2000s, when the Fed
Figure 2.1: Evolution of Monetary Policy between 1954 and 2008

Note: The figure shows the pointwise-posterior median estimate of $G(z)$, along with the 68 percent credible bands. $G(z)$ can be interpreted as a measure for monetary policy activism. Grey bars represent recessions as defined by the National Bureau of Economic Research (NBER). The figure provides evidence that monetary policy behaves differently from how theory models it. Monetary policy evolves continuously over time, changes smoothly, and responds several times to both inflation and recessions.

lowered nominal interest rates in response to the 1990-91 and 2000-01 recessions. Finally, Taylor [2007] argues that the Fed deviated substantially from the Taylor principle between 2002 and 2005. Our measure of policy activism captures each of these events, as well as others. The appendix contains a more detailed timeline of monetary policy in the United States.

The posterior of $G(z_t)$ shown over time in Figure 2.1 also sheds light on the theoretical frameworks used in the literature. These frameworks typically assume that monetary policy rules do not change in response to economic conditions such as inflation or recessions – or to government spending shocks, and only distinguish between active and passive monetary policy (e.g., Kim [2003], Zubairy [2014], Nakamura and Steinsson [2014] or Leeper et al. [2017]). Others use the Markov-switching approach to model policy changes (e.g., Davig and Leeper [2011]). That method implies that the central bank jumps randomly from one discrete regime directly to another.
Figure 2.1 provides overwhelming evidence that the monetary policy regime evolves differently from these models. First, our estimates of monetary policy and the narrative evidence given by Romer and Romer [2004] indicate that the central bank changed its policy regime several times in response to economic conditions during our sample period. For example, the Fed increased nominal interest rates aggressively (and became extremely active) in response to high inflation in 1959 and 1979-1982. Similarly, the Fed lowered nominal interest rates (and became more passive) in response to the recessions of 1957-58, 1990-91 and 2000-01. These observations suggest that impulse response functions that fix the Taylor rule after the government spending shock are misplaced, rendering their conclusions suspect.

Second, we find that monetary policy is not well described by just two regimes, but rather by a process that evolves continuously over time. For example, during the second half of the 1950s and in the late 1960s, the monetary policy regime reflects two different degrees of passiveness. Similarly, in 1970 and throughout the 1980s, the policy regime exhibits two different degrees of activeness. These observations imply that even within the active and the passive regimes themselves, differences occur.\footnote{In the appendix, we display the consequences of the binary interpretation of monetary policy if, in fact, the monetary policy regime evolves continuously over time. We find that the “average active” and the “average passive” monetary policy regimes are relatively similar to each other, i.e., \( G(z_1) = 0.16 \) and \( G(z_2) = 0.84 \), respectively. Consequently, it remains unclear whether the result of regime-independent multiplier estimates found in that exercise is coming from the fact that they truly do not depend on the initial monetary policy regime or whether the initial monetary policy regimes are too similar to detect any meaningful differences.}

Third, Figure 2.1 suggests that when monetary policy changes, it changes smoothly rather than abruptly. For example, the change in the monetary policy regime around 1980 took three years. The data do not support abrupt policy changes: the posterior argues strongly against \( \gamma \) being very large, indicating that Markov-switching approaches are incompatible with the data. This result supports Bianchi and Ilut [2017] and Chang and Kwak [2017], who similarly conclude that monetary policy evolves smoothly.

Taken together, these findings guide our approach to estimating the government spending multiplier conditional on the monetary policy regime in what follows. Most importantly, we observe that the central bank \textit{does} change its monetary policy regime in response

\footnote{In the appendix, we display the consequences of the binary interpretation of monetary policy if, in fact, the monetary policy regime evolves continuously over time. We find that the “average active” and the “average passive” monetary policy regimes are relatively similar to each other, i.e., \( G(z_1) = 0.16 \) and \( G(z_2) = 0.84 \), respectively. Consequently, it remains unclear whether the result of regime-independent multiplier estimates found in that exercise is coming from the fact that they truly do not depend on the initial monetary policy regime or whether the initial monetary policy regimes are too similar to detect any meaningful differences.}
to economic conditions during our sample period. Hence, when determining how the government spending multiplier depends on monetary policy, we must account for the endogeneity of monetary policy. Our main model allows the central bank to adjust its policy regime in response to the economic conditions after a government spending shock. For example, the central bank can switch to a more active policy when inflation rises rapidly due to a government spending intervention.

In the next section, we explore how the assumption of fixed monetary policy affects the relationship between the government spending multiplier and the monetary policy rule.

2.4 Results

This section presents our main results. We estimate our model using quarterly data for the U.S. economy from 1954Q3 to 2007Q4. The sample excludes the period in which monetary policy in the United States and other countries was constrained by the zero lower bound. This cutoff is often applied in the literature to avoid contamination through the effects of the Great Recession and unconventional monetary policy (e.g., Leeper et al. [2017] or Arias et al. [2019]). More importantly, though, the Fed kept nominal interest rates at zero between 2008Q4 and 2015Q4. This says that despite massive fiscal policy interventions such as the American Recovery and Reinvestment Act (ARRA) or the Troubled Asset Relief Program (TARP), the Fed held nominal interest rates constant.\footnote{We analyze the government spending multiplier at the zero lower bound in Section 2.6.}

We estimate our model with fully Bayesian methods that also include the marginal posterior distributions of the transition parameter $\gamma$ and the threshold parameter $c$. We then employ generalized impulse response functions to estimate the dynamic effects of a government spending shock. The generalized impulse response functions require an initial condition that allows us to estimate the multiplier for each initial monetary policy regime of our sample period. For comparison, we divide the initial regimes into quintiles and compare the multipliers when monetary policy is initially “very active,” “weakly active,” “neutral,” “weakly passive,” or “very passive” according to the value of $G(z_{t-1})$ during the impact.
period. If conventional wisdom is correct, we expect to see increasing multiplier estimates as we move smoothly from the most active quintile toward the most passive quintile. Finally, the generalized impulse response functions allow the central bank to change its policy regime in response to the government spending shock. For instance, the central bank can switch to a more active policy if inflation grows large after the shock.

To estimate the multiplier, we follow Auerbach and Gorodnichenko [2012] and use the formula for the sum multiplier

\[ \Phi_h = \frac{\sum_{j=0}^{h} y_j}{\sum_{j=0}^{h} g_j} \times Y/G, \]  

(2.12)

where \( y_j \) and \( g_j \) are the responses of output and government spending, respectively, in period \( j \) after the government spending shock. \( Y/G \) is the sample mean of the output-to-government-spending ratio. Using the sum formula, we estimate the multiplier for different time horizons after the shock.

Figure 2.2: Response of the Monetary Policy Regime to a Government Spending Shock

![Image of Figure 2.2](image)

Note: The figure shows the pointwise-posterior median evolution of \( G(z) \), along with the 68 percent credible bands, after a government spending shock when monetary policy is initially “very active,” “neutral,” or “very passive.” The figure provides evidence that the central bank responds quickly after the shock. Shortly after the shock and regardless of its initial condition, the central bank responds actively to inflation.
Figure 2.2 displays the response of the monetary policy regime after the government spending shock if the monetary policy regime is initially “very active,” “neutral,” and “very passive” using the evolution of $G(z_{t-1})$. The figure reveals interesting results. First, when the monetary policy regime is initially “very active” (left subplot), the central bank does not substantially adjust its responsiveness to inflation. The bank remains in the active sphere of the monetary policy spectrum for the entire past-shock horizon. Second, in contrast, if the monetary policy regime is initially “neutral” or “very passive,” then the central bank responds quickly, transitioning quickly to a more active regime. Thus, shortly after the shock and regardless of its initial regime, the central bank conducts policy in more or less the same way, responding actively to inflation. This result implies that the common modelling practice of keeping the monetary policy regime constant after the shock is misplaced, especially for passive regimes.

A natural question to ask is whether the response of the monetary policy regime affects the estimated government spending multiplier. It does. Figure 2.3 shows the estimated multipliers as a function of the initial regime.

Figure 2.3 compares the distributions of the government spending multiplier across the initial monetary policy regimes using boxplots. If the consensus in the literature holds, then, after accounting for the response of the monetary policy regime, the boxplots should shift upwards for more passive regimes. However, Figure 2.3 shows that there is little variation in the boxplots at each horizon after the shock. For example, one quarter after the shock, the posterior median estimate for the “very active” regime is 5.7 and 4.5 for the “very passive” regime. One year after the shock, the posterior median estimates are 4.2 and 3.6, respectively. Five years after the shock, they are 0.6 and 1.2, respectively. Even though there exist differences in the posterior median estimates, Figure 2.3 displays that the corresponding distributions highly overlap. Moreover, in the appendix, we show that the distribution of the difference between the “very active” and the “very passive” multipliers after five years is centered around zero. Thus, we find that the multiplier decreases in magnitude over time, but this decrease does not depend on the initial monetary policy regime contrary to the standard premise in the literature.
Figure 2.3: Estimated Multipliers when Monetary Policy is Fully Responsive

Note: The figure uses boxplots to show the estimated multiplier posterior distributions one quarter, one year, and five years after the shock across initial monetary policy regimes. The middle line and the box present the posterior median and the corresponding 50 percent credible bands. The upper and lower lines correspond to the highest and lowest value of the distribution that is not considered an outlier. The Figure suggests that the multiplier decreases over longer horizons but is almost completely unaffected by the initial monetary policy regime.

These results contradict the conventional wisdom that the government spending multiplier is larger when monetary policy is passive. Our findings suggest that the multiplier does not depend on monetary policy either in the short run or in the long run. This largely reflects the fact that the central bank responds quickly after a shock, and that it transitions rapidly to an active regime even when the initial regime had not previously been “very active.” The consensus in the literature does not account for this response of the monetary policy regime to the government spending shock. The consensus result hinges on a model assumption that lacks empirical support, and vanishes once that assumption is relaxed. Thus, we must conclude that the consensus – that government spending multipliers are larger when monetary policy is passive – is artificial, and is not a genuine feature of the data.

Furthermore, the results in Figure 2.3 suggest that short-run multiplier estimates are
considerably larger than in the long-run estimates. The empirical government-spending
literature largely agrees that the multiplier lies between 0.3 and 2.1 [Ramey, 2016, Ramey
and Zubairy, 2018]. However, the multiplier can also be negative [Perotti, 2014] and as high
as 3.5 [Edelberg et al., 1999]. While not directly comparable, our estimated short-run and
medium-run multiplier estimates are notably larger.

In most case, the literature has either focused only on longer-run multipliers, and/or
has employed identification strategies related to the Cholesky (e.g., Blanchard and Perotti
[2002]) or narrative (e.g., Ramey [2011]) approaches. The Cholesky approach imposes zero
restrictions, which require one to assume that governments adjust their spending only with a
delay when responding to shocks other than to government spending itself. This is a strong
assumption that is violated in practice. Recent events, such as the Troubled Asset Relief
Program (TARP) of 2008 or the CARES Act of 2020, have shown that governments can and
do react fast to changes in the business cycle, at least during times of crisis. The narrative
method introduced by Ramey [2011] uses a narrative time series based on newspaper articles
as an instrument for government spending shocks. However, this time series is mostly related
to military spending and military spending is an inefficient fiscal policy [Perotti, 2014]. In
addition, Ramey [2011] finds that her defense news shocks has low explanatory power for
the post-Korean-War period which is used in our study.

Accordingly, we expect the Cholesky approach to produce biased multiplier estimates.
To provide support for our conjecture, Figure 2.4 compares the multiplier estimates from
our main exercise (violet histograms) with those using the Cholesky approach (orange
histograms).18 We find smaller estimates in the short and medium runs for the Cholesky
approach. The estimates are similar after five years. This comparison of multiplier estimates
using different identification approaches is an important exercise. The findings indicate that
once we consider multipliers in the short and medium runs, and replace the zero restrictions
related to the recursive approach with uncontroversial sign restrictions, the multiplier may
be larger than previously suggested in the literature. We next provide additional evidence
that the constant-regime assumption is the key driver of the conventional wisdom.

---

18Because of the low explanatory power of Ramey’s defense news shocks for the post-Korean War period, we forgo the comparison with Ramey’s narrative approach.
2.5 Counterfactuals

In this section, we provide more evidence that the conventional wisdom is primarily driven by the constant-regime assumption, and not the data. First, we analyze what would happen to the government spending multiplier if the central bank were to keep its monetary policy regime temporarily constant after a shock. Second, we fully replicate the framework that underlies the consensus in the literature with our empirical model. To do this, we distinguish only between the purely active and the purely passive monetary policy regimes, and we keep the monetary policy regimes constant for the entire horizon after the shock. The findings of these exercises suggest that the constant-regime assumption drives results in the theoretical literature.

Because the policymaker can commit to a policy rule for an extended period of time, these contingencies represent hypothetical policy scenarios.
2.5.1 Government Spending Multiplier under Temporary Constant Monetary Policy

Section 2.4 revealed that the central bank adjusts its monetary policy regime after government spending shocks when the regime is not already “very active.” However, the central bank could have also chosen to fix its policy regime for some specific time after the shock. Indeed, monetary policy has remained nearly constant on a few occasions. For instance, Figure 2.1 shows that the monetary policy regime was purely active throughout the 1980s and the second half of the 1990s. By contrast, the monetary policy regime was purely passive in the second half of the 1970s and the first half of the 2000s. In addition, the central bank sometimes signals its intentions to hold policy constant in the future. For example, on April 29th, 2020, the Federal Reserve announced:

“The ongoing public health crisis will weigh heavily on economic activity, employment, and inflation in the near term, and poses considerable risks to the economic outlook over the medium term. In light of these developments, the Committee decided to maintain the target range for the federal funds rate at 0 to 1/4 percent. The Committee expects to maintain this target range until it is confident that the economy has weathered recent events and is on track to achieve its maximum employment and price stability goals.”

Hence, we address this possibility directly. First, we analyze what would happen to the government spending multiplier if the central bank were to keep its policy temporarily constant after the government spending shock. To do this, we employ the generalized impulse response functions but keep the inflation parameter $\phi_{\pi,t+h}$ constant for one year, two years and five years. After these periods expire, we begin updating $\phi_{\pi,t+h}$ via the “rolling” Kalman filter. Figures 2.5 and 2.6 display the corresponding multiplier estimates when the monetary policy regime is restricted to remain in its initial regime for one year and five years after a shock.

If the central bank keeps its policy regime constant for one year after a shock, the

\(^{20}\text{Source: https://www.federalreserve.gov/newsevents/pressreleases/monetary20200429a.htm}\)
Figure 2.5: Estimated Multipliers when Monetary Policy is Constant for One Year

Note: The figure uses boxplots to show the estimated multiplier distributions across initial monetary policy regimes when the regime is kept constant for one year after the shock. The middle line and the box present the posterior median and 50 percent credible bands of the corresponding distribution. The upper and lower lines correspond to the highest and lowest values of the distribution that are not considered to be outliers. The figure suggests that the multiplier decreases over time but is almost completely unaffected the initial monetary policy regime even if the regime is held constant for one year.

Results are similar to those in Section 2.4. The government spending multiplier decreases in magnitude over time, but it does not vary significantly with the initial monetary policy regime. We find similar results if we keep the monetary policy regime constant for two years.\textsuperscript{21} Only when the central bank keeps its policy regime constant for five years after the shock, do we observe a higher multiplier estimate in more passive regimes after five years. This is in line with conventional wisdom. While this finding indicates the possibility that the government spending multiplier may depend on monetary policy exists, the difference in the multiplier hinges crucially on the central bank’s willingness to maintain the monetary policy regime for extended periods of time. Our analysis indicates that the central bank must maintain its initial policy rule for more than two years before one can discern any meaningful differences in the government spending multiplier with respect to monetary policy.

\textsuperscript{21}Results are available upon request.
Figure 2.6: Estimated Multipliers when Monetary Policy is Constant for Five Years

Note: The figure uses boxplots to show the estimated multiplier distributions across initial monetary policy regimes when the regime is kept constant for five years after the shock. The middle line and the box present the posterior median and 50 percent credible bands of the corresponding distribution. The upper and lower lines correspond to the highest and lowest values that are not considered to be outliers of the distributions. The figure suggests that there multiplier is larger for the more passive regimes in the long run but only if the monetary policy regime is held constant for five years.

2.5.2 Government Spending Multiplier under Fully Constant Monetary Policy Regimes

We now “replicate” the theoretical framework with our empirical model. Recall that theory typically fixes the monetary policy regime for the entire time after the government spending shock when it computes impulse response functions. In addition, theory interprets monetary policy as binary, and only distinguishes between “active” and “passive” monetary policy regimes. Finally, some studies use the Markov-switching approach to model policy changes, which implies that the central bank jumps from one monetary policy regime directly to another.

To recreate these imposed assumptions, we estimate our model with an exogenous $\gamma$ that we calibrate to be very large. This choice ensures that the central bank jumps from one
regime directly to another. We then use traditional impulse response functions to estimate the dynamic effects of the shock. The traditional impulse response functions estimate the effects for the purely active (AM) and the purely passive (PM) regime, rather than an interior combination. Most importantly, though, the traditional impulse response functions keep the regimes constant for the entire forecasted horizon after the shock. Figure 2.7 presents the results.

Figure 2.7: Estimated Multipliers when Monetary Policy is Fully Constant

Note: The figure shows the estimated multiplier distributions for the most active (red) and most passive (blue) monetary policy regimes when the regimes are kept constant for the entire time after the shock. The figure suggests that the multiplier is larger when monetary policy is and remains “purely” passive.

Figure 2.7 illustrates that if we replicate the theoretical framework with our empirical model, then we also replicate the theoretical consensus. Up to one year after the shock, the distributions of the multiplier highly overlap so that there is no meaningful difference in the estimated multiplier between the two most extreme monetary policy regimes. One year after the shock, the estimated multiplier distributions start to separate. In the long run, the multiplier is estimated to be higher when monetary policy is and remains purely passive. This evidence mirrors the results in Leeper et al. [2017], who find comparable multipliers in
the short run but a higher multiplier under passive monetary policy in the long run.\textsuperscript{22}

Because we find regime-dependent multipliers \textit{only when we restrict the monetary policy regime to remain unchanged for more than two years} as in these two exercises, we conclude the conventional wisdom in the literature is largely driven by that constant-regime assumption. Regardless of whether we interpret the monetary policy regime as a continuous or a binary process, or whether we model smooth or abrupt regime changes, the government spending multiplier diverges in the long run \textit{only if} we keep the monetary policy regime constant for at least two years. When we relax this assumption and allow the central bank to update its policy rule after the shock as we see over our sample period, the multipliers ceased to diverge.

These results indicate that—to boost the impact of government spending on the economy—the central bank would need to accommodate inflation for a surprisingly long period of time after a government spending shock. However, this would compromise the central bank's main mandate to maintain price stability, and it would violate the Taylor principle.

\subsection*{2.6 Government Spending Multiplier at the Zero Lower Bound}

In this section, we analyze the government spending multiplier when nominal interest rates are stuck at zero in response to a \textit{severe} economic downturn. In Section 2.4, we show that the monetary policy regime itself changes quickly after a government spending shock. The sole exception to that result is the zero lower bound period that began in December 2008. In December 2008, the Fed set nominal interest rates to zero, and kept them there until December 2015 despite unprecedented fiscal policy interventions such as the American Recovery and Reinvestment Act (ARRA).\textsuperscript{23} When we extend our sample period to 2015Q4

\textsuperscript{22}We also obtain similar results when we conduct the same analysis, but estimate the full posterior distribution of our model including an estimated $\gamma$. Results are available upon request.

\textsuperscript{23}During the zero lower bound period, the Fed kept nominal interest rates constantly at zero for almost seven years after the ARRA was signed into law. Furthermore, nominal interest rates fell back toward zero in March 2020, and, as of this writing, many countries are \textit{still} at the zero lower bound. In addition, the Fed
and reestimate our model (see Figure 8), our measure of monetary policy activism \( G(z_t) \) is zero with very little uncertainty during the zero lower bound period. If there is an example where monetary policy remains purely passive over a long period of time, it would be the zero lower bound.

Figure 2.8: Evolution of Monetary Policy Between 1954 and 2016

Note: The figure shows the pointwise-posterior median estimate of \( G(z) \), along with the 68 percent credible bands. \( G(z) \) can be interpreted as a measure for monetary policy activism. Grey bars represent recessions as defined by the National Bureau of Economic Research. This Figure represents an extension of Figure 2.1. During the zero lower bound era (2008Q4-2015Q4), \( G(z) \) is consistently zero with very little uncertainty indicating that the the zero lower bound era is an “extremely passive” monetary policy regime.

However, Monetary policy is not the only structural change that occurred in 2008.\(^{24}\) Because many changes occurred simultaneously, the size of the multiplier post-2008 is a controversial topic. There is ample evidence of a macroeconomic structural break in 2008 (Silva and Hassani [2015], for instance). From a policymaker’s standpoint, the dynamics of

\[^{24}\text{Other major changes that occurred in 2008 and shortly thereafter include the Fed’s use of unconventional monetary policy (quantitative easing) begin in November 2008. Dramatic changes in banks’ lending behavior also began at roughly the same time [Ivashina and Scharfstein [2010]]) and sweeping regulatory reforms began shortly thereafter. For example, the Americans with Disabilities act (2009) and the Patient Protection and Affordable Care Act (2010) substantially changed labor market regulations, while the Dodd-Frank act (2010) imposed new financial regulations. Some of these changes—like the changes in the Fed’s behavior—have proven to be long-lived.}\]
the economy after 2008 are of primary importance. But estimating the dynamics of the real economy post-2008 is a challenge due to the scarcity of data. Indeed, most of the relevant literature [Leeper et al., 2017, Arias et al., 2019] relies entirely on data prior to the 2008 crisis. Policymakers are in uncharted territory.

Empirical studies using other data sets reach a variety of differing conclusions. Almunia et al. [2010] and Gordon and Krenn [2010] find multipliers above two for the zero lower bound period during the Great Depression. Miyamoto et al. [2018] estimate the multiplier at the modern zero lower bound in Japan to lie between 1.5 and 2. By contrast, Ramey [2011] and Crafts and Mills [2013] find multiplier estimates below unity for the zero lower bound epochs from the first half of the 20th century in the U.S. and the U.K., respectively, while Ramey and Zubairy [2018] uncover ambiguous results.

To make matters worse, the theoretical literature gives opposing predictions as well. Woodford [2011], Christiano et al. [2011], and Eggertsson [2011] predict that the government spending multiplier is higher when monetary policy accommodates inflation, and that it is especially high when monetary policy is constrained by the zero lower bound. However, this belief has been challenged by Aruoba and Schorfheide [2013], Mertens and Ravn [2014], Kiley [2016], and Wieland [2018], who argue that the multiplier at the zero lower bound can be below one. With no clear guidance from the theoretical or empirical literature, we turn our attention back to the data.

To analyze the government spending multiplier at the zero lower bound, we estimate a linear VAR model for the period between 2008Q4 and 2015Q4. The structural break in 2008 motivates the use of a separate model (substantial change in the correlation structure of the data is also apparent: see the appendix). In this exercise, we use the same economic variables as in Sections 2.4 and 2.5, save for the federal funds rate which we replace by the Wu and Xia [2016] shadow rate. Note that the corresponding (linear) impulse response functions keep the monetary policy regime constant after the government spending shock.

Table 2.1 presents the results. The first row shows the pointwise posterior median estimate of the government spending multiplier when we use sign restrictions on impulse responses...
response functions for identification, along with the 68-percent credible bands. The posterior median decreases from 5.2 one year after the shock to 4.1 five years after the shock. The 68-percent credible bands are above one for two years and turn partly negative after four and five years.

Table 2.1: Estimated Multipliers at Zero Lower Bound 2008Q4-2015Q4

<table>
<thead>
<tr>
<th></th>
<th>1 Quarter</th>
<th>1 Year</th>
<th>2 Years</th>
<th>3 Years</th>
<th>4 Years</th>
<th>5 Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR (ZLB)</td>
<td>5.3</td>
<td>5.2</td>
<td>5.5</td>
<td>5.0</td>
<td>4.5</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>[2.1, 13.7]</td>
<td>[2.2, 13.1]</td>
<td>[1.9, 17.0]</td>
<td>[0.9, 16.9]</td>
<td>[-0.3, 16.5]</td>
<td>[-6.4, 15.6]</td>
</tr>
<tr>
<td>SR (Pre-ZLB) PM</td>
<td>5.6</td>
<td>4.0</td>
<td>3.5</td>
<td>3.4</td>
<td>3.3</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td>[2.7, 12.1]</td>
<td>[2.1, 7.7]</td>
<td>[1.8, 7.2]</td>
<td>[1.6, 7.1]</td>
<td>[1.5, 7.3]</td>
<td>[1.3, 7.4]</td>
</tr>
</tbody>
</table>

Note: Upper part of the Table presents our posterior median estimates along with the 68 percent credible bands in brackets. Posterior median decreases over time and credible bands turn negative after 4 years. Bottom part of the Table provides an overview of estimates from the related literature. Estimated multipliers range from 0.3 to 3.7.

The bottom part of Table 2.1 provides an overview of estimated multipliers from the related literature. The overall literature provides an estimated range of multipliers between 0.3 and 3.7. In the empirical literature, the multiplier estimates lie between 0.3 and 2.7. Our results are consistent with the theoretical predictions given by Woodford [2011], Christiano et al. [2011], and Eggertsson [2011], and empirical estimates provided by Almunia et al. [2010], Gordon and Krenn [2010] and Miyamoto et al. [2018], at least for three years after the shock. However, when we compare our estimates with those from the empirical literature, we find that our sign-restricted estimates are larger.

The related empirical literature relies either on the Cholesky approach or Ramey’s narrative method. But as discussed in Section 2.4, the Cholesky approach imposes timing restrictions which are clearly violated during periods in which the central bank is constrained by the zero lower bound. Ordering government spending first in a recursive VAR implies
that governments can change their spending plans only with a delay when responding to shocks other than to government spending itself. In contrast, our set of sign restrictions is consistent with a large class of neoclassical and new Keynesian models, and hence, has theoretical foundations. More importantly, though, these sign restrictions also allow government spending to respond quickly, as in the CARES, PPP, and TARP examples. Thus, the difference between our multipliers and those found elsewhere may be attributable to a Cholesky bias. Our estimated multiplier for the zero lower bound is imprecise but does not suffer from the Cholesky bias prevalent in the literature: this suggests that the multiplier post-2008 may be larger than previously reported.

To complete the story, we compare our estimates during the zero lower bound to the counterfactual exercise in Section 2.5. Christiano et al. [2011] argue that, during the zero lower bound, the multiplier should be relatively large because monetary policy is perfectly passive. Our two-regime model incorporates this feature, and predicts a five-year multiplier of 3.3 when the passive regime remains in place, as was the case during the zero lower bound. Our estimates using data from the zero lower bound are less precise, but the posterior median estimate is similar (at 4.1). These two best-available estimates of the government spending multiplier at the zero lower bound are also consistent with the Christiano et al. [2011] baseline estimate of 3.7. Because all of these estimates agree, we cannot rule out the possibility that the government spending multiplier at the zero lower bound may be as large as 4.

2.7 Conclusion

Our work presents a flexible, nonlinear, structural vector autoregression model to investigate the relationship between the government spending multiplier and monetary policy. Conventional wisdom suggests that the multiplier is larger when nominal interest

---

26Cholesky identification of government spending shocks was once a popular approach. However, the CARES Act of 2020 represents the largest economic stimulus package of any kind in U.S. history, and it was debated and passed into law in only a few weeks. The Paycheck Protection Program (PPP) attached to that package disbursed $349 billion in less than two weeks. The TARP of 2008 similarly disbursed hundreds of billions in the same quarter in which Congress passed the bill. This demonstrates that governments can and do react quickly during times of crises, and that practices in the current era violate the timing restrictions of the recursive approach. As a result, estimates based on Cholesky restrictions are biased.
rates respond less than one for one to increasing inflation. However, models supporting this consensus keep the monetary policy regime constant after the government spending shock when they estimate multipliers. Our analysis shows that the central bank responds quickly after a shock, and that it transitions rapidly to an active regime even if the initial regime had not previously been active. The response of the monetary policy regime dramatically alters estimates of the multiplier. Once we account for the regime’s reaction, the relationship between the multiplier and monetary policy vanishes. By contrast, when we force the monetary policy regime to remain counterfactually constant after the shock, we find multiplier estimates that are consistent with the conventional wisdom. After the comparison, we conclude that the key driver of convention wisdom is the constant-regime assumption — not the data.

Our analysis highlights the necessity of accounting for the reaction of monetary policy to the government spending shock to properly study the relationship between monetary policy and the government spending multiplier. Failure to do so ignores the central bank’s ability to respond to shocks. This leads to a misrepresentation of how the multiplier depends on monetary policy. Our results indicate that to boost the impact of government spending on the economy, the central bank would need to tolerate inflation for long periods of time – for longer than two years, according to our estimates. However, this creates a dilemma because it would require the central bank to violate its main mandate of maintaining price stability, and would conflict with the Taylor principle.

The exception to our main result that the monetary policy regime itself responds to a government spending shock is the zero lower bound that began in December 2008. Analyzing the government spending multiplier when nominal interest rates are stuck at zero is critical for ongoing policy problems. When we apply our sign restriction method that allows for an immediate government (spending) response, the multiplier estimates exceed those previously reported by the (empirical) literature. This result suggests that recent fiscal stimulus packages, such as the American Recovery and Reinvestment Act of 2009 and the CARES Act of 2020, may have larger economic effects than the literature has previously suggested. Analyzing the dynamics of the real economy post-2008 is challenging due to the
scarcity of data: estimates will be imprecise. Overcoming this challenge should be the next main focus for empirical research.
Chapter 3

When is an Expectile the Best Linear Unbiased Estimator?

3.1 Introduction

People are not all alike. In this article, we show that expectile regression is an unbiased estimator of the mean regression line for individuals who are more likely than usual to be above or below that line, ceteris paribus. We find that expectile regression will be the best linear unbiased estimator (BLUE) under simple assumptions which nest the classical Gauss-Markov assumptions. Then, when heteroscedasticity is present, we find that a weighted expectile regression is the BLUE. All of these estimators have the usual GLS form that will be familiar to readers. We also consider the behavior of ancillary statistics, namely R-squared and the estimator of residual variance. Because expectiles are a type of generalized quantile, our results extend the Gauss-Markov theorem and GLS theorem to generalized quantile regression. In cases where standard quantile regression is difficult to apply—such as binary response models—expectile regression is the new best option.

Mortgage application denial is an important example. Historically, the supply of credit to minorities and other disadvantaged persons has been different from the supply of credit in the broader market. When minorities face credit constraints in the housing market, they become less able to build equity as well as less able to relocate to other neighborhoods. This mechanism reinforces poverty traps in low-income areas and contributes to intergenerational poverty. The Home Mortgage Disclosure Act (HMDA) requires financial institutions to publish certain data related to home mortgage applications. With these data, it is straightforward to show whether black and Hispanic applicants are denied at higher rates than white applicants in any given year. Using expectile regression, we show that the racial
disparity for the most-affected expectile of individuals is essentially twice as severe as has been previously reported. Individuals who are approximately four times as likely as average to be denied are the most affected. Then, we will answer whether those denials can be attributed to factors other than race: financial, employment, property characteristics, or others.

Expectile regression is also known as asymmetric least squares (ALS), or least asymmetrically-weighted least squares (LAWS). Recently, there has been disagreement regarding how expectiles can be interpreted. Expectiles are usually compared to quantiles, but that comparison has serious limitations. Expectiles are estimated by minimizing a weighted least squares problem with different weights on positive versus negative residuals so we rely on a more direct interpretation. Specifically, each expectile regression line would be the mean regression line if the probabilities of seeing observations above or below the regression line was reweighed in the same way that the weighted least-squares criterion is weighted. For mortgage denial, this represents cases where an individual is more likely than average to be denied, conditional on all observed covariates. Taking the expectile’s interpretation seriously leads to a new framework of analysis, which develop in three steps. They are (1) we show that expectile regression is an appropriate estimator for this empirical context, in the sense that it is an unbiased linear estimator. Our result is not limited to the binary response framework. Then, (2) we identify the most efficient linear estimator for this class of problem, which has the usual GLS form, and (3) we determine an appropriate goodness-of-fit criterion to represent the proportion of variation in the data that can be explained by the covariates in our regressions. These three steps turn expectile regression into a much more practical methodology than it has been in the past.

First, we show that expectile regression is an appropriate way to characterize individuals who are more or less likely than average to be denied a mortgage. The asymmetric least squares estimator is similar to ordinary least squares (OLS), but puts more (less) weight on observations that are above (or below) the regression line. In case it is not obvious that this is the correct estimator for our problem, we justify the expectile regression in the same way that ordinary least squares has traditionally been justified. Specifically, we generalize the Gauss-
Markov theorem to the asymmetric regression environment. Under simple assumptions, that theorem states that the ordinary least squares estimator of regression coefficients is the Best Linear Unbiased Estimator (BLUE) in the sense that it has the minimum variance of any such estimator. In this paper, we also identify two alternative scenarios where an expectile regression estimator is the BLUE. In one scenario, data are missing at random but with a probability that depends on the sign of the error term. In the other scenario, the desired regression line has an error term that is not zero on average.

Next, we consider the possibility that the regression’s residuals are heteroscedastic, which is obviously the case in the mortgage denial application. In a linear probability model, the response variable (whether the individual is denied or not) has a Bernoulli distribution, and its variance varies with the exogenous regressors. We find that it is possible to exploit the heteroscedasticity in order to reduce the variance of the estimator—exactly the same as with Generalized Least Squares (GLS). We show that a Generalized Expectile Regression (GER) is the best linear unbiased estimator under heteroscedasticity and that the optimal weights are proportional to the inverse of the residual’s variance. This result nests its classical counterpart in the GLS theorem of Aitken [1936]. Exact GLS weights are typically unknown, and thus the BLUE is infeasible. But we point out that the standard expectile weights are feasible in the binary response context, and that this result extends to nonlinear models, such as Logit and Probit models. Because expectile regression is a type of generalized quantile regression, our generalized expectile regression is also the best linear unbiased estimator of quantile coefficients in special cases where quantile and expectile coefficients are the same—but we point out that mortgage denial is not such a case.

This research establishes a new link between two major topics in the literature. On the one hand, classical linear models such as ordinary least squares (OLS) and generalized least squares are universally known and well-studied, but fail to consider the heterogeneity of treatment effects across the distribution of the dependent variable. On the other hand, quantile regression (QR) and generalized quantiles are designed specifically to explore heterogeneous treatment effects, but lacks the simplicity and elegance of the classical linear models. Linear estimators of quantiles and generalized quantiles were not seriously studied
until Waltrup et al. [2015] and Daouia et al. [2019], and the question of whether there is a Best Linear Unbiased Estimator for (generalized) quantiles does not appear to have been studied at all. For the first time, we present expectile regression in a framework that nests the classical linear models and the foundational theorems in modern regression analysis—the Gauss-Markov theorem as stated by Markov [1912], and the GLS theorem as stated by Aitken [1936]. This is a substantial breakthrough for binary response models, where quantile regression has been difficult to apply and heterogeneous treatment effects have been difficult to analyze. Our results can be compared to (or combined with) the sorted effects method proposed by Chernozhukov et al. [2018], who target the same problem but do not consider regressions other than the mean.

In concert, these individual contributions support a broader overall theme. Expectile regression is usually described as an alternative form of quantile regression or as being “similar to quantiles”. In our application, the sample quantiles and expectiles nothing in common. On the contrary, expectiles in binary response problems are as simple to estimate as generalized least squares, and they are as easy to interpret as linear probability models. We strongly suggest that expectile regression belongs in the classical linear regression framework, and that they should not be confined to the quantile regression’s sphere of influence.

The rest of the document is arranged as follows. Section 2 introduces expectiles formally and presents an overview of the expectile regression problem. Section 3 gives the expectile regression an elementary treatment in terms of non-central counterparts to the Gauss-Markov assumptions. That treatment is somewhat trivial, but novel in the literature and necessary for what follows. In section 4 we show that an expectile estimator (or predictor) may be the BLUE (or predictor) in very simple model misspecification problems. Section 6 is devoted to ancillary statistics (mean squared error and $R^2$) for regressions of this type and section 7 concludes.
3.2 Background

For a distribution $F$ with a finite mean, expectiles are the class of location parameters that contain the minimum (infimum) of $F$, the maximum (supremum) of $F$, and every point in between. Expectiles were introduced by Newey and Powell [1987], but have only recently become popular. Expectile regression is popular primarily due to its ability to elicit results not found using mean regression: see Waltrup et al. [2015], Kneib [2013], or Waldmann et al. [2018] for discussions. Because of that similarity to quantile regression, expectiles are sometimes classified as a type of generalized quantile Daouia et al. [2019], m-quantile Breckling and Chambers [1988], or $L_p$-quantile Chen [1996]. Under certain conditions—which are not satisfied in our application—these three families allow the original quantile regression of Koenker and Bassett [1978] to be estimated using alternative loss functions. For the location-scale model with Lebesgue measure, different $L_p$-quantiles produce the same sets of regression lines Yao and Tong [1996], Daouia et al. [2019], making a broad menu of estimators available for each such line. The virtues of the asymmetric class of estimators are widely extolled: see Koenker [2005], Koenker et al. [2017], Kneib [2013]. But—in contrast to the rest of that family and to quantiles in particular—expectiles are produced using an asymmetrically weighted least squares estimator, and expectile regression nests ordinary least squares (OLS) when the weights are constant. Furthermore, expectiles are the only generalized quantiles with a (piecewise) quadratic loss function and an interpretation as an expected value. Expectiles have other unique properties; see Ziegel [2016] or Bellini et al. [2014]. For an excellent overview of the mathematical properties of a sample expectile, see Holzmann and Klar [2016].

Because expectile estimators are the only linear estimators in the class of $L_p$-quantiles, and because they are so fundamentally similar to OLS and GLS estimators, an obvious methodological question is whether the Gauss-Markov theorem can be extended to expectile regression. It can. Others have suggested that expectiles may be the efficient estimator for the $m$-quantile class Stahlschmidt et al. [2014] without exploring the concept in detail. The efficient estimator will depend on the underlying distribution $F$, but the efficient estimator
from some limited classes of estimators can be studied directly. Because the expectile coefficients are produced by minimizing a quadratic (weighted least squares) function, the solution produces a linear estimator in the classical sense.\footnote{Generalized quantiles are defined as minimizers of asymmetrically weighted generalized loss functions, see Daouia et al. \cite{Daouia2019}. Only the quadratic subset of this class produces linear estimators. If we remove linearity as a requirement, other estimators may be MVUE depending on the context; see Koenker et al. \cite{Koenker2017}.} For the linear subset of $m$-quantiles—which is only expectiles—we find that the efficient estimator has a form similar to traditional GLS.

One critical impediment to the adoption of expectiles and expectile regression has been the lack of any elementary treatment of the subject. With apologies, we have done only a little to advance that cause. Hopefully, the reader will find our framework and our application to be more accessible than the typical treatment of this subject. Expectiles can be very easy to interpret, and some of our results are obvious in hindsight. Among these, we provide two examples where one of Newey and Powell’s expectile estimators (our GLS variant) is the BLUE in a mean regression. These two examples can be motivated as standard GLS problems, and the result (the estimator is the BLUE) follows from the GLS theorem. Our framework is built on the four assumptions from the classical linear model: the model is linear in parameters, regressors are strictly exogenous, full rank, and spherical variance-covariance. Exogeneity and spherical variance-covariance require modification in order to accommodate non-central estimators, i.e. estimators of parameters other than the mean. A fifth assumption is typically presented in the classical linear model–Gaussianity of the residuals–which is impossible to assume when the location parameter is not at the mean of the distribution. Accordingly, we make no such restriction on the shape of the distribution. However, quasi-likelihood models that elicit expectiles do exist: see Philipp [2021a], Waldmann et al. [2018]. The expectile regression is known to be the efficient estimator for one of these; see Philipp [2021a].

Economic disparity between groups is a perennial topic. As far as mortgage denial is concerned, Munnell et al. \cite{Munnell1996} find that minorities were more than twice as likely to be denied a mortgage as whites, though a large portion of that inequality could be explained by factors other than race. We find that a greater burden falls on those who were comparatively
disadvantaged—more likely than average to be denied—in the first place. The proportion of
the burden that can be explained by other factors is relatively constant. Labor mobility (or
lack thereof) has long been considered a major mechanism that sustains inequality across
generations—see van der Berg et al. [2011], Lemanski [2011], or Turok and Borel-Saladin
[2018]. In the United States, school funding is typically tied to local property taxes, and the
only way for parents to move their children from low-quality schools to high-quality schools
is by moving their residence to another district. But, in past decades, residential segregation
in the U.S. has decreased only little Wagmiller et al. [2017] and cannot be fully explained by
non-race variables, such as income Aliprantis et al. [2019]. Our proposed methodology will
allow others to analyze inequality in new data sets, such as mortgage applications, college
admissions, and other binary survey data where quantile regression is difficult to apply. But
expectile regression is not limited to these environments—its appeal should be universal.

3.3 Preliminaries

The context here is similar to a standard linear regression model. The vector $y$ is a stochastic
linear function of $X$:

$$y = X\beta + \epsilon.$$  \hfill (3.1)  

The data $y, X$ are observed as vectors of $n$ observations of random variables $Y, X$ in a
joint probability space with outcomes in $\mathbb{R}^{k+1}$. The elements of the linear model are the
repeated observations of the dependent variable $y_i$, its covariates $x_i$, the linear coefficients
$\beta$ and the vector of error terms $\epsilon$. Namely,
\(y = [y_1, \ldots, y_n]'\),
\(X = [x_1, \ldots, x_n]'\),
\(x_i = [1, x_{i,2}, \ldots, x_{i,k}]'\),
\(\beta = [\beta_1, \ldots, \beta_k]'\),
\(\epsilon = [\epsilon_1, \ldots, \epsilon_n]'\).

The conditional distribution \(F_{Y|X}\) is assumed to exist for all \(x\) in the support of \(X\). In this document, we will explicitly consider the case where \(\{y_i, x_i\}\) are jointly i.i.d. but \(F_{Y|X}\) need not be identically distributed. Expectile regression with serial correlation or other dependency structure has been studied recently: see for instance the working paper by Barry et al. Barry et al. [2018] or the related result in the paper by Philipps [2021a]. The model conforms to the following assumption:

**Assumption 0:** The process \((Y, X) := \{y_i, x_i : \Xi \to \mathbb{R}^{k+1}, i = 1, \ldots, n\}\) is defined on a joint probability space \((\Xi, \mathcal{F}, P)\) where \(\Xi\) is the universe, \(\mathcal{F}\) the corresponding sigma-algebra, and \(P : \mathcal{F} \to [0, 1]\) a corresponding probability measure. The conditional distribution of \(Y\) given \(X\), \(F_{Y|X}\), exists for all \(x\) in the support of \(X\) and has at least two finite moments. \(X\) has at least two finite moments.

We define the **generalized quantiles** of a distribution \(F\) as a set of summary statistics indexed by \(\tau \in (0, 1)\). They are the class of minimizers

\[
\theta_\tau(F) = \arg\min_\theta \int (\tau - I(y < \theta)) \rho(y - \theta) dF(y)
\]  

obtained by minimizing the expected value of some objective function \(\rho\) with respect to the distribution of \(Y\), but with asymmetric weights applied such that positive and negative errors are treated differentially. For an \(m\)-class objective function \(\rho\), these produce the standard \(m\)-statistic when \(\tau = .5\) and the \(\tau^{th}\) \(m\)-quantile more generally Breckling and Chambers [1988]. When \(\rho\) is an \(L_p\) loss function, \(\rho(y - \theta) = \|y - \theta\|_p^p\), these are the \(L_p\)-
quantiles of Chen [1996]. Generalized quantiles extend to the linear regression environment in Assumption 3.3 by replacing \( \theta \) with \( x'\beta \), and minimizing over possible vectors of linear coefficients:

\[
\theta_{\tau}(P) = \arg \min_{\beta} \int |\tau - I(y < x'\beta)|\rho(y - x'\beta)dP
\]

(3.3)

The standard \( L_1 \)– and \( L_2 \)-quantile objective functions are shown in Figure 3.1. The \( L_2 \)-quantile loss function is an asymmetrically weighted least squares criterion and produces expectiles, or expectile regression. Because these are the only linear estimators\(^2\) in the class of \( L_p \)-quantiles, they are the focus of the rest of this article.

\(^2\)A “linear” estimator can be written as \( \hat{\beta} = C'y \) for some matrix \( C \), i.e. it is linear in each \( y_i \). The estimators corresponding to 3.2 on the previous page have first-order conditions

\[
\frac{1}{n} \sum_{i=1}^{n} |\tau - I(y_i < \hat{\theta})|\psi(y_i - \hat{\theta}) = 0,
\]

which is linear in \( y_i \) only if \( \psi(y_i - \theta) = \frac{\partial}{\partial \theta} \rho(y_i - \theta) \) is linear for all \( y_i \). Clearly this implies that \( \rho \) is quadratic.
3.3.1 Expectiles

Expectiles are weighted averages indexed by \( \tau \in (0, 1) \) that range from the minimum of \( F \) to its maximum. The standard arithmetic mean is the most popular expectile, and occurs when \( \tau = .5 \). One definition of the \( \tau^{th} \) expectile of a distribution, which is a special case of equation 3.2 on page 71, is

\[
\mu_{\tau}(Y) := \arg\min_{\theta} \int \varsigma_{\tau}(y-\theta)dF(y).
\]  

(3.4)

Thus, Newey and Powell [1987] show that expectiles can be obtained by minimizing a particular “swoosh” function

\[
\varsigma_{\tau}(u) = \begin{cases} 
\tau u^2 & \text{if } u \geq 0 \\
(1-\tau)u^2 & \text{if } u < 0 
\end{cases}
\]  

(3.5)

which is a piecewise quadratic loss function comparable to the piecewise linear “check” function of Koenker and Bassett [1978]. As such, expectiles have been characterized primarily by their relationship to quantiles. However, the asymmetric least squares form of equation 3.5 makes evaluation of the \( \tau^{th} \) linear sample quantiles into a nearly-standard weighted least squares problem, using the empirical measure \( P_n \):

\[
\hat{\beta}_{\tau} := \arg\min_{\beta} \int \varsigma_{\tau}(y_i - \mathbf{x}_i\beta)dP_n \\
= \arg\min_{\beta} \sum_{i=1}^{n} w_i(y_i - \mathbf{x}_i\beta)^2
\]  

(3.6)

where the weights \( w_i = \tau \) if \( y_i \geq \mathbf{x}_i\beta \) and \( w_i = 1-\tau \) otherwise. The coefficients for the \( \tau^{th} \) linear regression expectile adopt the subscript \( \tau \), say \( \beta_{\tau} \), following the notation of Koenker [2005]. As a weighted least squares problem, the sum in equation 3.6 can be expressed as \( \hat{\epsilon}'W\hat{\epsilon} \), where \( W \) is the diagonal matrix \( [W]_{ii} = w_i \). Naturally, the least asymmetric sum of squares estimator is \( \hat{\beta}_{\tau} = (X'WX)^{-1}X'Wy \) which Newey and Powell [1987] show to be consistent and asymptotically normal under reasonably general conditions. See Holzmann 73
and Klar [2016] for more general asymptotics in the location model. Recent literature has added substantial context regarding the usefulness of these statistics for specialized purposes in economics and finance: see Waldmann et al. [2018], Waltrup et al. [2015], or Ziegel [2016].

As the minimizer of the “swoosh” function in 3.5 and 3.6, the $\tau$th expectile of the distribution $F$ has a dual interpretation. First, it can be expressed as a weighted average:

$$
\mu_\tau(F) = \int y \kappa w(y) \, dF(y)
$$

where

$$
w(y) = \begin{cases} 
\tau & y \geq \mu_\tau \\
1-\tau & y < \mu_\tau \end{cases}
$$

and $\kappa$ is some constant such that the weights integrate to one, $\kappa = \left(\int w(y) \, dF(y)\right)^{-1}$ or $\kappa^{-1} = E(w(Y))$. Then for proper\(^3\) weights as above, $\mu_\tau = \kappa E(w(Y) \times Y)$. Because the minimizer of the function in 3.5 does not vary over affine transforms of $\varsigma_\tau$, $\kappa$ will have a relatively small role in the remainder of this document.

Alternately, the weights may be incorporated into the distribution $F$ such that expectile may be interpreted as the (unweighted) expected value of the distribution $\tilde{F}$,

$$
\mu_\tau(F) = \int y \, d\tilde{F}(y)
$$

where $d\tilde{F} = \kappa w(y) \, dF(y)$. (3.7)

This can be attributed to Breckling and Chambers [1988], who make a similar point in relation to $m$-quantiles. Both the interpretation with respect to $F$ and the same with respect to $\tilde{F}$ will play a role in later sections. However, the interpretation in 3.7 is our default interpretation: the $\tau$th expectile of $F$ is the mean of $\tilde{F}$, or it is the mean of $F$ if observations above $\mu_\tau(F)$ were $\frac{1}{1-\tau}$ as likely to occur as they actually are, \textit{ceteris paribus}. Because the $\tau$th expectile is the expected value of a random variable drawn from $\tilde{F}$, we adopt

---

\(^3\) We say that the weights $w(y)$ are proper if $E(w(Y)) = 1$. The weights $\kappa w(Y)$ are proper by construction.
the notation $E_\tau$ for the $\tau^{th}$ expectile operator, which is a weighted expectations operator:

$$E_\tau(Y) = \int y \kappa w(y) \, dF(y) = \int y \, d\hat{F}(y) = \mu_\tau(F).$$  \hspace{1cm} (3.8)

We will use this notation extensively. These asymmetric expectations $E_\tau(\cdot)$ are clearly a special case of the usual expectations operator $E_F(\cdot)$ (and vice versa) and inherits all of its properties. In the case where $\tau = .5$, we have the usual expectations under $F$, so we suppress the subscript and simply write $E(\cdot)$.

In the regression environment, the sample expectile (predictor) for $y$ has the usual GLS form

$$\hat{y}_\tau = X\hat{\beta}_\tau$$

$$= X(X'WX)^{-1}X'Wy$$

$$= P_\tau y$$  \hspace{1cm} (3.9)

which is an asymmetrically weighted $L_2$ projection of $y$ onto the space spanned by $X$. The matrix $P_\tau = X(X'WX)^{-1}X'W$ is the weighted projection or “hat” matrix. We also obtain the vector of residuals $\hat{\epsilon} = y - \hat{y}$ or

$$\hat{\epsilon} = y - \hat{y}$$

$$= (I - X(X'WX)^{-1}X'W)y$$

$$= M_\tau y$$  \hspace{1cm} (3.10)

---

4The weighted expectation $E_\tau(Y)$ can also be expressed as

$$\mu_\tau = E_\tau(Y) = E \left[ \frac{\psi_\tau(Y - \mu_\tau)}{E(\psi_\tau(Y - \mu_\tau))} \times Y \right]$$

where $\psi_\tau(y) = \tau - I(y < 0)$, which is the derivative of Koenker and Bassett’s “check” function $\rho_\tau(u)$ Koenker and Bassett [1978], Koenker [2005, p. 36]. The absolute value of $\psi_\tau$ is sometimes called the check function, see Barry et al. [2018]. We have suppressed this notation in order to reduce the reader’s barrier to entry: our “proper” weights are given by $\kappa w(y) = \frac{\psi_\tau(y - \mu_\tau)}{E(\psi_\tau(Y - \mu_\tau))}$. In cases where it is not necessary to normalize the weights, it suffices to say that $w_i$ is proportional to $(1 - \tau)$ for negative errors and $\tau$ otherwise.

75
where the matrix $M_\tau$ is the expectile annihilator matrix. The projection and annihilator matrices $P_\tau$ and $M_\tau$ are idempotent but not symmetric, thus they are not orthogonal projections in the usual sense. Instead, the sample expectile can be considered an oblique projection Basilevsky [2013, p. 165] Zhang [2017, p. 578]. Oblique projections are rarely discussed per se in econometrics, but they have been studied explicitly in other applied sciences, such as signal processing [Behrens and Scharf, 1994]. Some details regarding $P_\tau$ and $M_\tau$ are given in the appendix and several of the implications of their structure are discussed in Section 3.6.

### 3.3.2 First Illustration: Mortgage Denial

The mortgage denial data of Munnell et al. [1996] are challenging to explore “beyond the mean” using traditional methodologies. Because every quantile of the response variable is either zero or one (accept or deny), a quantile regression reveals little about treatment effects. But, if some individuals are more or less likely to be denied, ceteris paribus, then an expectile regression is an obvious approach to that problem.

The two most important variables are $Deny$, which is 1 if the application was denied (zero otherwise), and $Black$ which is 1 if the applicant was black (zero otherwise). For the simple regression of $Deny$ on $Black$, expectiles are shown in Figure 3.2. The coefficient of $Black$ is shown as a function of $\tau$ in Figure 3.3.

Several results are apparent from these two figures. In Figure 3.2, the mean regression ($\tau = .5$) is shown in black, while expectiles above the mean are shown in red and expectiles below the mean are shown in blue. The slope is very low for expectiles on either extreme end of the spectrum: individuals who were extremely likely to be denied or extremely likely to be accepted see only a small effect due to race. But around $\tau = .8$, the slope is noticeably higher than for the mean. Figure 3.3 shows the coefficient of $Black$ as a function of $\tau$: the maximum is achieved around $\tau = .83$. Apparently, individuals with $\frac{\tau}{1-\tau} \approx 5$ are the most affected.

---

5 Quantile regression results for the simple regression are shown in Figure C.1 in Appendix C.8. The results are degenerate and standard asymptotics are invalid.
Figure 3.2: The simple regression of deny on black. All observations fall into one of the four corners, constraining a traditional quantile regression. Expectiles for \( \tau \) equal to all multiples of \( .1 \in (0,1) \) are given along with \( \tau = .99 \). The mean regression \( \tau = .5 \) is shown in black, and represents the probability of denial on average. Expectiles in red represent individuals more-likely-than-average to be denied; expectiles in blue represent the opposite.

3.1. We conclude that individuals who were already more likely to be denied (credit-disadvantaged) are the ones who suffer the greatest racial disparity. A subset of these regression coefficients is reported in Table 3.1. In this application, the regression coefficient is positive and statistically significant for every \( \tau \in (.01,.99) \), but that is not always the case. In the next subsection, we give a brief application where the sign of the coefficient changes.

### 3.3.3 Second Illustration: The Mexican Repatriation

In Figure 3.4 on page 79, a mean regression is compared to the expectile regression of Newey and Powell [1987]. The data are taken from the US Census and described in Cortes and Sant’Anna [2019]. These data represent the Mexican repatriation which occurred during the 1930’s, during the great depression. In the southwestern parts of the United States, organized labor and political groups harassed Mexican citizens (and Mexican-Americans)
Figure 3.3: In a linear probability model for the simple regression of Deny on Black, the coefficient of Black varies depending on which expectile is estimated. The maximum coefficient occurs around $\tau = .83$, suggesting individuals who are approximately five times more likely than usual to be denied (ceteris paribus) have the largest racial disparity. For that group, the effect of Black is roughly twice the size of the average effect. However, the effect is statistically significant at all levels.

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau$</td>
<td>.5</td>
<td>.6</td>
<td>.7</td>
<td>.8</td>
<td>.9</td>
</tr>
<tr>
<td>black</td>
<td>0.191***</td>
<td>0.239***</td>
<td>0.287***</td>
<td>0.323***</td>
<td>0.302***</td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
<td>(0.030)</td>
<td>(0.032)</td>
<td>(0.033)</td>
<td>(0.028)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.093***</td>
<td>0.133***</td>
<td>0.192***</td>
<td>0.290***</td>
<td>0.479***</td>
</tr>
<tr>
<td></td>
<td>(0.006)</td>
<td>(0.009)</td>
<td>(0.012)</td>
<td>(0.016)</td>
<td>(0.019)</td>
</tr>
<tr>
<td>Observations</td>
<td>2,380</td>
<td>2,380</td>
<td>2,380</td>
<td>2,380</td>
<td>2,380</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.042</td>
<td>0.053</td>
<td>0.063</td>
<td>0.071</td>
<td>0.067</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 3.1: Regression coefficients are shown for $\tau = .5, .6, .7, .8, .9$. In contrast to the OLS regression where individuals with $black = 1$ are 19% more likely to be denied, we see that the average is driven by individuals who were relatively more likely to be denied. For individuals four times more likely than average to be denied, the estimated effect of $black = 1$ is close to 32%. The effect of $black$ is larger for individuals who were already experiencing credit challenges in that sense.
Figure 3.4: Expectile regression with nine different values of $\tau$. Two major results are apparent that were not clear from the mean regression: first, the data are heteroscedastic with increasing variance as we move from left to right. Second, the bottom part of the distribution slopes *downwards* on average, moving in the opposite direction from the rest of the group.

Figure 3.5: On the vertical axis, the change in the repatriation intensity for a one-unit change in the proportion of Mexicans in 1930. In other words, the vertical axis represents the slope coefficient from the regression. The horizontal axis represents which expectile we are interested in. As you see, coefficients are positive in the center of the plot (at the mean regression) but the far-left dips below zero and is statistically significant.
<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity</td>
<td>Intensity</td>
<td>Intensity</td>
<td>Intensity</td>
<td>Intensity</td>
<td>Intensity</td>
</tr>
<tr>
<td>( \tau = .001 )</td>
<td>.067*</td>
<td>-0.070**</td>
<td>0.058</td>
<td>0.138**</td>
<td>0.267***</td>
</tr>
<tr>
<td></td>
<td>(0.038)</td>
<td>(0.031)</td>
<td>(0.060)</td>
<td>(0.054)</td>
<td>(0.036)</td>
</tr>
<tr>
<td>( \tau = .01 )</td>
<td>-0.003***</td>
<td>-0.001***</td>
<td>-0.000**</td>
<td>-0.000**</td>
<td>-0.000</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
<tr>
<td>Proportion of Mexicans</td>
<td>877</td>
<td>877</td>
<td>877</td>
<td>877</td>
<td>877</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.077</td>
<td>0.064</td>
<td>0.031</td>
<td>0.171</td>
<td>0.524</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
<th>(9)</th>
<th>(10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity</td>
<td>Intensity</td>
<td>Intensity</td>
<td>Intensity</td>
<td>Intensity</td>
<td>Intensity</td>
</tr>
<tr>
<td>( \tau = .5 )</td>
<td>.340***</td>
<td>.409***</td>
<td>.507***</td>
<td>.550***</td>
<td>.640***</td>
</tr>
<tr>
<td></td>
<td>(0.032)</td>
<td>(0.031)</td>
<td>(0.026)</td>
<td>(0.023)</td>
<td>(0.027)</td>
</tr>
<tr>
<td>( \tau = .7 )</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000**</td>
<td>0.000***</td>
<td>0.000***</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
<tr>
<td>( \tau = .9 )</td>
<td>877</td>
<td>877</td>
<td>877</td>
<td>877</td>
<td>877</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.664</td>
<td>0.747</td>
<td>0.842</td>
<td>0.874</td>
<td>0.893</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 3.2: Expectile coefficients from the regression of repatriation intensity on the proportion of Mexicans in a given locale in 1930. The OLS estimate from this sample is .34, which is very significant. Close to the maximum of the distribution, the regression coefficient reaches as high as .64, which is nearly twice the OLS estimate. At the bottom of the distribution, the estimated coefficient turns negative and (at \( \tau = .01 \)) is statistically significant. Standard errors are the sandwich estimator of Newey and Powell with the degrees of freedom adjustment \( \frac{n}{n-k} \).

80
living north of the border, and pushed them to leave the U.S and return to Mexico. This led to roughly 1.3 million persons leaving the U.S. over the course of ten years, between the 1930 and 1940 census. For an in-depth description of that dark period, see Balderrama and Rodríguez [2006].

Each observation in Figure 3.4 on page 79 represents a single municipality represented in the census, and the size of the circle represents the population of that municipality. On the horizontal axis is the proportion of Mexican citizens living in that municipality at the time of the 1930 census. The proportional change in the population of Mexican citizens from 1930-1940 is shown on the vertical axis. The mean regression reveals that the change was larger in places with a higher proportion of Mexican nationals in 1930–suggesting that the repatriation “push” was more focused on those areas. That result is not surprising.

However, the intensity of the repatriation did not increase with the proportion of Mexican nationals everywhere. On the contrary, the bottom part of the distribution is below zero on the vertical axis—suggesting that some areas saw an increase in their Mexican population over that ten-year period. In fact, the bottom of the distribution moves further downwards as the proportion of Mexicans in 1930 increases—this is evident from the expectile regression, where the lowest regression line is downwards-sloping. In fact, this suggests that some places—where the repatriation was the least intense—saw the inflows of Mexican nationals be larger if the initial population was larger. These municipalities moved against the tide, and acted as safe havens. It is possible that individuals fleeing other municipalities pursued safety in numbers and moved into these municipalities. The downward slope is indeed statistically significant, as seen in Table 3.2 on the previous page.

At the same time, the top of the distribution moves upwards much more quickly (almost twice as quickly) than the mean, suggesting that the most-racist cities had effectively twice the average repatriation intensity. The slope coefficients are shown in Table 3.2. A summary of these effects is represented in Figure 3.5 on page 79. As a function of τ, it becomes clear that the slope coefficient is negative in the lowest extremes of the distribution and positive (and large) in the highest extremes. The data also clearly feature heteroscedasticity, i.e. the variance increases with the Mexican proportion of the population in 1930. For the mean
regression at $\tau = .5$, ordinary least squares is not the best linear unbiased estimator. In the next section, we will determine which estimator is the best linear unbiased estimator for each of these expectile regression lines.

### 3.4 Expectile (Generalized) Least Squares

In this section, we will extend the Gauss-Markov framework [Markov, 1912] and the GLS framework [Aitken, 1936] to expectile regression. For a general linear model, the estimated regression coefficients have the form

$$ \hat{\beta}_{\tau} = (X'AX)^{-1}X'Ay $$

for some matrix $A$. In the next two subsections, we will show that the expectile regression estimator and its generalized counterpart are BLUE under simple conditions. For clarity, we will refer to the expectile regression estimator in equation 3.11 using whichever name indicates the classical estimator that would be produced in the central case when $\tau = .5$. Thus, the “Generalized” Expectile Regression or Expectile GLS nests the GLS estimator when $\tau = .5$. A Weighted Expectile Regression or Expectile WLS nests the classical weighted least squares when $\tau = .5$. The “ordinary” expectile regression of Newey and Powell [1987] nests Ordinary Least Squares when $\tau = .5$. Otherwise, “Expectile Regression” refers to the concept generally.

In each following subsection, we treat $\tau$ as known a priori or assume that multiple $\tau$ are of interest separately. In section 3.5, we consider the possibility that a particular $\tau$ is optimal for specific model designs.

#### 3.4.1 Expectile Gauss-Markov Assumptions

Here, we show that the expectile regression can be the best linear unbiased estimator. As in the previous section, we have a random sample of data pairs $y_i, x_i$ for $i = 1, ..., n$. Denote the column vector made from $\{y_i\}_{i=1}^n$ as $y$ and the matrix made from each $x_i'$ as $X$. The
weights $w_i$ and the asymmetric expectations operator $E_\tau$ are defined as in the previous section.

The following four assumptions nest the classical Gauss-Markov assumptions when $\tau = .5$ and the expectile regression nests Ordinary Least Squares. We will address that relationship more thoroughly in Section 3.5. An interesting result is that the first and third assumptions (linearity and full rank) are the same for all expectiles and require no modification at all, while the second and fourth vary depending on which expectile is of interest. The fourth assumption in particular (spherical variance-covariance) is usually not taken seriously: a more general heteroscedastic version is given in Section 3.4.2.

The first assumption is obvious: the model for the $\tau^{th}$ conditional expectile of $y_i$ given $x_i$ is linear.

**Assumption 1:** The model is linear in parameters.

$$y_i = \beta_{\tau,1}x_i,1 + \ldots + \beta_{\tau,k}x_i,k + \epsilon_i$$

$$= x_i'\beta + \epsilon_i$$  \hspace{1cm} (3.12)

**Assumption 2:** Weighted Strict Exogeneity:

$$E_\tau(\epsilon_i|X) = 0 \quad \text{or} \quad E(w_i\epsilon_i|X) = 0 \quad \forall i \in \{1,\ldots,n\} .$$  \hspace{1cm} (3.13)

**Assumption 3:** Full-Rank condition: The $n \times k$ matrix $X$ has full column rank.

**Assumption 4:** “Spherical” error variance or Asymmetric Homoscedasticity:

$$E_\tau(\epsilon_i^2|X,W) = \kappa \nu^2 \quad \text{or} \quad E(w_i\epsilon_i^2|X,W) = \nu^2 \quad \forall i .$$  \hspace{1cm} (3.14)

and

$$E_\tau(\epsilon_i\epsilon_j|X,W) = 0 \quad \text{or} \quad E(w_i\epsilon_i\epsilon_j|X,W) = 0 \quad \forall i \neq j .$$  \hspace{1cm} (3.15)

Each of these assumptions buys something of value. The first assumption buys us a
convenient interpretation of each coefficient as the partial derivative of the $\tau^{th}$ expectile of $y_i$ with respect to the corresponding covariate:

$$\frac{\partial E_\tau(y_i)}{\partial x_{i,j}} = \beta_{\tau,j} \forall j \in \{1, \ldots, k\}.$$ 

such that the $\tau^{th}$ expectile of $y_i$ increases by $\beta_{\tau,j}$ when $x_{i,j}$ increases by one, *ceteris paribus*. These are the “expectile treatment effects” discussed by Stahlschmidt et al. [2014], but they are also the “average treatment effect” in a causal model if values above the regression line were $\frac{\tau}{1-\tau}$ times as likely to occur as they actually are.

The second assumption assures that the weighted residual vector $\epsilon$ is orthogonal to the data $X$. This can be re-formulated, using $\epsilon_i = y_i - x_i'\beta_\tau$ from equation 3.12, as

$$E_\tau(y_i|X) = x_i'\beta_\tau \quad \text{or} \quad E(w_i y_i|X) = x_i'\beta_\tau \times E(w_i|X) \quad \forall i \in \{1, \ldots, n\}.$$  

It is convenient but not necessary to require $E(w_i|X) = 1$, conforming to our definition of “proper” weights in footnote 4. Either way, Assumption 2 assures that the regression line is indeed the $\tau^{th}$ expectile of $y_i$ given $X$. Equation 3.13 also implies some other trivial results. For instance,

$$E_\tau(x_{ij} \epsilon_i) = 0 \quad \text{or} \quad E(w_i x_{ij} \epsilon_i) = 0 \quad \forall i, j.$$  \hfill (3.16)

This follows directly from the tower rule.\(^6\) And the conditional moment in equation 3.13

\(^6\)We can write

$$E_\tau(x_{ij} \epsilon_i) = E(E_\tau(x_{ij} \epsilon_i|X)) = E(x_{ij} E_\tau(\epsilon_i|X)) = 0 \quad \forall i, j.$$  \hfill (3.17)

Alternatively, write

$$E(w_i x_{ij} \epsilon_i) = E(E(w_i x_{ij} \epsilon_i|X)) = E(x_{ij} E(w_i \epsilon_i|X)) = 0.$$ 

Note that the subscript $\tau$ appears only once on the right side of equation 3.17 because writing $E_\tau(E_\tau(\cdot))$ would imply applying the weights twice. The tower rule is not generally applicable to $E_\tau$ when $\tau \neq .5$; see a comment and a proof in [Philipps, 2020].
also implies the value of the unconditional moment by the tower rule:

\[ E_\tau(\epsilon_i) = E(w_i \epsilon_i) = E(E(w_i \epsilon_i | X)) = 0. \]

Assumption 3 is required for identification, and the same assumption is required for ordinary least squares and many other linear models. Notably, it does not matter which expectile we are modelling—the third assumption either holds for all expectiles or it holds for none. Compare this to the major result in Newey and Powell [1987], where the \( \tau^{th} \) expectile of the distribution exists if and only if its first moment exists. Compare it also to the asymptotic results in [Holzmann and Klar, 2016], who prove that the \( \tau^{th} \) sample expectile converges to the sample expectile under the same minimal sufficient conditions as the mean converges to the true mean—essentially, the strong law of large numbers applies equally to all expectiles if it applies to any of them. Some further remarks regarding assumption 3 are given in Appendix C.1.

The fourth assumption can also be written as \( E(W \epsilon \epsilon' | X, W) = \nu^2 I_n \). Assumption 4 implies that each \( \epsilon_i \) has the same weighted variance. Alternately, assumption 4 implies that the right and left conditional variances (given that the error term is positive or negative) vary by an a priori known ratio.\(^7\) Specifically, the fourth assumption implies that the ratio of variance for positive errors to variance for negative errors is \( \frac{1-\tau}{\tau} \). Because the \( \tau^{th} \) expectile is not the mean, the distribution of residuals is “skew” in precisely this way and is not zero on average.

It also implies that each pair of distinct \( \epsilon_i, \epsilon_j \) are uncorrelated. This can also be called “spherical” variance in the sense of a sphere defined under an asymmetric distance function.\(^8\) We would also point out that the weighted covariance in equation 3.15 uses weights which vary depending on \( \epsilon_i \); though \( \epsilon_i \) and \( \epsilon_j \) may be reversed and the statement remains true.

---

\(^7\)To assist the reader, we would like to draw attention to the relationship between \( \nu^2 \) and \( \sigma^2 = E(\epsilon_i^2) \). Clearly, they are the same if the weights are exactly one for all observations (ordinary least squares). A common consistent estimator for the weighted variance parameter \( E(w_i \epsilon_i^2) \) in weighted least squares problems is \( \frac{1}{n-k} \sum_{i=1}^{n} w_i \epsilon_i^2 \), which we will recommend as an estimator for \( \nu^2 \) when the variance of \( \beta \) or the predictor is of interest. In section 3.6 we will discuss estimators for \( \sigma^2 = E(\epsilon_i^2) \), which is appropriate when the variance of the residual (or of \( y_i \)) is of interest.

\(^8\)Asymmetric distance functions are also called quasi-metrics, and satisfy most of the standard properties of metrics: see [Matthews, 1994].
As with the first and second assumptions, comparisons to the standard symmetric variance are interesting and some of the usual shorthand formulas remain true in this environment. For instance, we can define the weighted variance of $\epsilon_i$ given $X$ as

$$WVar(\epsilon_i|X) := E_\tau ((\epsilon_i - E_\tau(\epsilon_i|X))^2|X) \tag{3.18}$$

$$= E_\tau (\epsilon_i^2 - 2\epsilon_iE_\tau(\epsilon_i|X) + E_\tau(\epsilon_i|X)^2|X)$$

$$= E_\tau(\epsilon_i^2|X) - E_\tau(\epsilon_i|X)^2.$$

This is identical to the usual expression given for variance; $Var(\epsilon_i|X) = E(\epsilon_i^2|X) - E(\epsilon_i|X)^2$, with the addition of weights. But that should come as no surprise—this value is the variance of the weighted distribution $\tilde{F}$, and thus the usual properties of variance should hold for $\tilde{F}$. Also, with weighted strict exogeneity (Assumption 2) we have $E_\tau(\epsilon_i|X)^2 = 0$, so $WVar(\epsilon_i|X) = E_\tau(\epsilon_i^2|X)$. Again, the $\tau^{th}$ expectile of $F$ is the mean of $\tilde{F}$, so it inherits the usual properties of the mean when $\tilde{F}$ is treated as the true distribution. See section 3.6 for further discussion of the differing definitions of “variance” available in this context.

Using Assumptions 1-4, we can motivate the construction of expectile regression and determine whether it is the “best” linear unbiased estimator in the sense of having the minimum possible variance. We can also collapse these assumptions to represent the usual first four Gauss-Markov assumptions (for Ordinary Least Squares).

**Proposition 1.** Let the weights $w_i$ be constant ($\tau = .5$) and proper ($E(w_i) = 1$) such that $W = I_n$ uniquely. Then assumptions 1-4 are the first four Gauss-Markov assumptions.

This proposition can serve as a formal statement of the Gauss-Markov assumptions. The proposition is actually true even when weights are constant but improper (not equal to one), but the previous statements simplify in the most elegant manner when $w_i = 1$. Under these four assumptions, the following is a “standard” result.

**Proposition 2.** Let assumptions 1-4 be correct. Then the expectile regression estimator
\( \hat{\beta}_\tau = (X'WX)^{-1}X'Wy \) has the following properties:

\[
E(\hat{\beta}_\tau|X, W) = \beta_
\]
\[
Var(\hat{\beta}_\tau|X, W) = \nu^2 (X'WX)^{-1}
\]

Thus, the estimator is unbiased with variance \( \nu^2 (X'WX)^{-1} \). A full derivation of this result is given in Appendix C.2: The Linear Unbiased Estimator, but it is identical to the standard result for GLS estimators. See Greene [2003], for instance. Next, we will show that the estimator \( \hat{\beta}_\tau \) is the BLUE.

### 3.4.2 The “Best” Linear Unbiased Estimator

We say that the “Best” Linear Unbiased Estimator (BLUE) is the linear unbiased estimator with the least variance. The Gauss-Markov theorem presented by Markov [1912] states that ordinary least squares is the BLUE under assumptions 1-4 with \( \tau = .5 \). When residuals are heteroscedastic, the Generalized Least Squares estimator (GLS) of Aitken [1936] applies.

**With Spherical Variance-Covariance**

Here, we will show that any linear and unbiased estimator under assumptions 1-4 for \( \tau \in (0, 1) \) will have at least as much variance as \( \hat{\beta}_\tau = (X'WX)^{-1}X'Wy \). A proof of that fact begins with the definition of a linear estimator and then proceeds as follows.

By definition, a linear estimator is an estimator that can be written as a linear function of \( y \); say \( \tilde{\beta}_\tau = Cy \) for some choice of nonrandom matrix C. For our estimator \( \tilde{\beta}_\tau \), we have

\[
E(\tilde{\beta}_\tau|X, W) = E(Cy|X, W)
\]
\[
= E(CX\beta_\tau + C\epsilon|X, W)
\]
\[
= E(CX\beta_\tau|X, W) + E(C\epsilon|X, W)
\]
\[
= CX\beta_\tau.
\]
Here, \( E(C \epsilon | X, W) \) is zero because \( C \epsilon \) is an unbiased linear estimator of the \( \tau^{th} \) expectile of \( \epsilon \), which is uniquely zero under assumption 2. But the fact that \( \tilde{\beta}_\tau \) is unbiased also implies \( CX \beta_\tau = \beta_\tau \), which requires that the \( k \times k \) matrix \( CX = I \). Also, we can always write \( C = D + (X'WX)^{-1}X'W \) for some \( D \). Doing this, we see that

\[
I = CX = DX + (X'WX)^{-1}X'WX
\]

(3.19)

\[
= DX + I.
\]

Then \( DX \) is uniquely zero. Now write the conditional variance of \( \tilde{\beta}_\tau = Cy \) using this same decomposition:

\[
\text{Var}(\tilde{\beta}_\tau | X, W) = E \left( (\tilde{\beta}_\tau - \beta_\tau)(\tilde{\beta}_\tau - \beta_\tau)' | X, W \right)
\]

\[
= E (C\epsilon\epsilon'C'|X, W)
\]

\[
= CE (\epsilon\epsilon' | X, W) C'
\]

\[
= \left( D + (X'WX)^{-1}X'W \right) \nu^2W^{-1} \left( D + (X'WX)^{-1}X'W \right)'
\]

\[
= \nu^2 \left( DW^{-1}D' + (X'WX)^{-1} + DW^{-1}WX (X'WX)^{-1} + (X'WX)^{-1}X'WW^{-1}D \right)
\]

(3.20)

\[
= \nu^2 \left( DW^{-1}D' + (X'WX)^{-1} \right).
\]

This follows using the fact that \( DX = 0 \). But the matrix \( DW^{-1}D' \) is positive definite, so

\[
\text{Var}(\tilde{\beta}_\tau | X, W) = \nu^2 \left( DW^{-1}D' + (X'WX)^{-1} \right)
\]

\[
\geq \nu^2 (X'WX)^{-1} = \text{Var}(\tilde{\beta}_\tau | X, W).
\]

That is, any unbiased linear estimator has at least as much variance as \( \tilde{\beta}_\tau \). With these steps, we have proven the proposition below.

**Proposition 3.** Under assumptions 1 through 4, the expectile regression estimator

\[
\hat{\beta}_\tau = (X'WX)^{-1}X'Wy
\]

is uniquely zero because \( C \epsilon \) is an unbiased linear estimator of the \( \tau^{th} \) expectile of \( \epsilon \), which is uniquely zero under assumption 2. But the fact that \( \tilde{\beta}_\tau \) is unbiased also implies \( CX \beta_\tau = \beta_\tau \), which requires that the \( k \times k \) matrix \( CX = I \). Also, we can always write \( C = D + (X'WX)^{-1}X'W \) for some \( D \). Doing this, we see that

\[
I = CX = DX + (X'WX)^{-1}X'WX
\]

(3.19)

\[
= DX + I.
\]

Then \( DX \) is uniquely zero. Now write the conditional variance of \( \tilde{\beta}_\tau = Cy \) using this same decomposition:

\[
\text{Var}(\tilde{\beta}_\tau | X, W) = E \left( (\tilde{\beta}_\tau - \beta_\tau)(\tilde{\beta}_\tau - \beta_\tau)' | X, W \right)
\]

\[
= E (C\epsilon\epsilon'C'|X, W)
\]

\[
= CE (\epsilon\epsilon' | X, W) C'
\]

\[
= \left( D + (X'WX)^{-1}X'W \right) \nu^2W^{-1} \left( D + (X'WX)^{-1}X'W \right)'
\]

\[
= \nu^2 \left( DW^{-1}D' + (X'WX)^{-1} + DW^{-1}WX (X'WX)^{-1} + (X'WX)^{-1}X'WW^{-1}D \right)
\]

(3.20)

\[
= \nu^2 \left( DW^{-1}D' + (X'WX)^{-1} \right).
\]

This follows using the fact that \( DX = 0 \). But the matrix \( DW^{-1}D' \) is positive definite, so

\[
\text{Var}(\tilde{\beta}_\tau | X, W) = \nu^2 \left( DW^{-1}D' + (X'WX)^{-1} \right)
\]

\[
\geq \nu^2 (X'WX)^{-1} = \text{Var}(\tilde{\beta}_\tau | X, W).
\]

That is, any unbiased linear estimator has at least as much variance as \( \tilde{\beta}_\tau \). With these steps, we have proven the proposition below.

**Proposition 3.** Under assumptions 1 through 4, the expectile regression estimator

\[
\hat{\beta}_\tau = (X'WX)^{-1}X'Wy
\]
is the best linear unbiased estimator in the sense that it has the least possible variance.

Next, we will ask whether the estimator $\hat{\beta}_\tau$ is useful to construct a predictor of $y$ or of any (known) linear function of $y$; say $Ay$ for $A$ possibly but not necessarily $I_n$. The following two propositions can be proven in a nearly identical manner, so the proof of proposition 4 is given in Appendix C.3

**Proposition 4.** Under assumptions 1 through 4, $AX\hat{\beta}_\tau$ is the best linear unbiased predictor of $Ay$, given $X$.

In particular, we might note that $A$ could be any one of the elementary basis vectors such that $Ay = y_i$ and $AX\beta = x_i'\hat{\beta}_\tau$. Then we have also proven the next proposition.

**Proposition 5.** Under assumptions 1 through 4, $x_i'\hat{\beta}_\tau$ is the best linear unbiased predictor of $y_i$.

And, lastly, proposition 4 also implies proposition 3. Let $A$ be $(X'WX)^{-1}X'W$; such that $AX = I_k$. The result follows.

Together, these three propositions assure that (i) $\hat{\beta}_\tau$ is the BLUE for the expectile treatment effects of $X$ on $Y$; (ii) the vector $AX\hat{\beta}_\tau$ is the BLUE of any linear function of the dependent variable $y$; and (iii) in particular, $x_i'\hat{\beta}_\tau$ is the best linear unbiased estimator (predictor) of the $i^{th}$ observation, given its covariates $x_i$. So expectile regression is the optimal estimator whether the coefficients or the predictor are of interest, and whether the entire sample or a particular individual are of interest.

Expectiles are not usually an unbiased estimator (nor are generalized quantiles), but assumption 2 ensures that this property holds if the weights $w_i$ are known a priori. We will show that this estimator is sometimes “feasible” in section B.3.3. For our binary response problem, the weights are known a priori—they are either $\tau$ or $1 - \tau$ depending on whether the dependent variable is 1 or 0. This differs from traditional GLS, where the optimal weights must be estimated and the “best” estimator is impossible to obtain in a finite sample.

Assumption 4 ensures that the estimator is the most efficient, but there is little reason to believe that it is a valid assumption. In the next subsection, we will relax assumption 4 and find the optimal estimator under general variance-covariance structure.
With Heteroscedasticity

When the residual vector $\epsilon$ is heteroscedastic, assumption 4 is violated. One strategy to mitigate that problem is to transform the data into a form such that the Gauss-Markov assumptions are applicable. Let assumptions 1 through 3 hold, but let the variance of $\epsilon$ be something other than before, say $E(W^{1/2}\epsilon'W^{1/2}|X, W) \neq \nu^2 I_n$. Instead, let

$$E(\epsilon'X, W) = \nu^2 \Sigma = \nu^2 W^{-1/2}\Omega W^{-1/2}$$  (3.21)

for some $\Sigma$ and some corresponding $\Omega$, possibly a function of $X$. In the case where $\Omega$ is diagonal (implied by Assumption 0) we have $E(\epsilon'X, W) = \nu^2 W^{-1}\Omega$ which obviously nests assumption 4 if $\Omega = I_n$. Otherwise, we have a symmetric positive definite variance of the residual vector with $\frac{1}{1+\tau}$ times the variance for negative errors as for positive errors, but some nontrivial covariance structure. This is the same as the interpretation in the previous subsection: the distribution of errors remains “skew”, though they may have different variances.

Because $\Sigma$ is symmetric and positive definite, we have some invertible matrix $V$ such that $V'V = W^{1/2}\Omega^{-1}W^{1/2} = \Sigma^{-1}$. Without loss of generality, suppose that $V$ is the Cholesky decomposition of $\Sigma^{-1}$. The traditional way of finding the best estimator applies here: left-multiply the entire model by $V$ and say

$$Vy = VX\beta + V\epsilon \equiv \tilde{y} = \tilde{X}\beta + \tilde{\epsilon}.$$

It is trivial to show that assumptions 1 and 3 apply to this transformed model. Clearly,

---

$^9$Notice $E(w_i\epsilon_i^2|X, W) = \nu^2 \Omega_{ii}$; the elements of $\Omega$ are proportional to the expected squared error.
the model is linear and $\tilde{X}$ has full rank. But also,

$$
E(\tilde{\epsilon} \tilde{\epsilon} | \tilde{X}, W) = E(V \epsilon' V' | X, W) \\
= V E(\epsilon \epsilon' | X, W)V' \\
= \nu^2 V \Sigma V' \\
= \nu^2 I_n.
$$

That is, assumption 4 will apply to the transformed model. It is also straightforwards to show that the OLS estimator using $\tilde{y}, \tilde{X}$ is an unbiased estimator, under assumption 2. Because we have already incorporated the expectile weights in equation 3.21, this has become a perfectly standard GLS problem. Thus, the following proposition indicates the best linear unbiased estimator for the $\tau^{th}$ expectile coefficients if the data are heteroscedastic.

**Proposition 6.** Let assumptions 1-3 hold and let $E(\epsilon \epsilon' | X, W) = \nu^2 \Sigma = \nu^2 W^{-1/2} \Omega W^{-1/2}$ for some known symmetric positive definite $\Omega$, possibly a function of $X$. Then the expectile GLS estimator

$$
\hat{\beta}_{\tau, GLS} = (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} y
$$

is the “best” linear unbiased estimator in the sense that is has the minimum variance in that class of estimators.

A proof is given in Appendix C.3. Obviously, we have the corollary result that the linear predictor $AX\hat{\beta}_{\tau, GLS}$ is the best unbiased linear predictor of $Ay$ and that $x_i' \hat{\beta}_{\tau, GLS}$ is the best unbiased linear predictor of $y_i$ given covariates $x_i$. We leave that proof to the reader. The critical piece of this puzzle is that assumption 2 implies that the $\tau^{th}$ expectile is the regression line that we are interested in. Whatever $\tau$ may be, the best linear unbiased estimator is trivial to find. As such, the classical result extends to any expectile with just a bit of additional notation.

This result will also play a major role in the next section, where we consider the alternative justifications for assumption 2. To conclude this section, we restate the two main points. The “ordinary” expectile regression is BLUE under an asymmetric spherical
variance-covariance assumption where the ratio of positive to negative variance is $\frac{1-\tau}{\tau}$. The “generalized” expectile regression nesting the estimator of Aitken [1936] is BLUE under the assumption of heteroscedasticity of known form when the ratio of variance for positive and negative errors is $\frac{1-\tau}{\tau}$.

3.4.3 A Best Linear Unbiased Estimator for Quantiles

Recently, others have become interested in alternative estimators for quantile regression, or for generalized quantiles. In the special case where these different types of regression lines coincide, estimators based on different loss functions have different finite-sample and asymptotic properties. For instance, the mean and median of the Gaussian distribution are the same value but the sample mean has lower variance. At the same time, the sample median is more robust to outliers [Huber, 1997]. There are other less-known considerations in regression models, such as the quantile crossing problem discussed by Waltrup et al. [2015], who suggest the possibility that sample expectiles may be a superior estimator of quantile regression lines.

The fact that quantile and expectile regression can produce the same set of regression lines was introduced by Yao and Tong [1996], and has been developed further. Under conditions such that the generalized quantile or m-quantile coefficients exist,

$$\theta_\tau(P) = \arg\min_\beta \int |\tau - I(y < x'\beta)|\rho(y - x'\beta)dP$$

(3.23)

an expectile regression could be used to estimate $\theta_\tau$ if there is some $\beta_{h(\tau)}$ such that $\theta_\tau = \beta_{h(\tau)}$ [Daouia et al., 2019]. In Yao and Tong’s example [1996], $\rho(u)$ was the $L_1$ absolute value loss function and $\theta_\tau$ were standard regression quantiles of Koenker and Bassett [1978]. If the loss function $\rho(u)$ is convex and $P$ has continuous density on the support, a location-scale model is sufficient for that result:

$$y_i = x_i'\theta_\tau + x_i'\gamma \epsilon_i$$

$$\epsilon_i \sim i.i.d F$$
where the coefficients $\gamma$ determine how the variance of the distribution increases with $x_i$. In that case, it is straightforward to show that there exists a function $h(\tau) : (0, 1) \mapsto (0, 1)$ such that the coefficients $\theta_\tau$ are equal to $\beta_{h(\tau)}$. Specifically,

$$h(\tau) = \frac{\int |y - x_i^\tau \theta_\tau| I(y < x_i^\tau \theta_\tau) dP}{\int |y - x_i^\tau \theta_\tau| dP}.$$  \hspace{1cm} (3.24)

In that case, every $\theta_\tau$ is equal to some expectile coefficients. But this could be true for a particular line under much more general conditions. Regardless, if some expectile coefficients $\beta_{h(\tau)}$ exist such that $\beta_{h(\tau)} = \theta_\tau$, then $\hat{\beta}_{h(\tau),GLS}$ is also the BLUE for $\theta_\tau$.

**Proposition 7.** Let assumptions 1,3 hold with the generalized quantile $x_i^\tau \theta_\tau$ the desired regression line. Let there be some expectile regression line such that $\beta_{h(\tau)} = \theta_\tau$. Then $\hat{\beta}_{h(\tau),GLS}$ is the BLUE for $\theta_\tau$ and $x_i^\tau \hat{\beta}_{h(\tau),GLS}$ is the best linear unbiased predictor of $x_i^\tau \theta_\tau$.

This proposition follows immediately from the result in proposition 6 on page 91. It also supports the result of Waltrup et al. [2015], who suggest using the full set of regression expectiles as a substitute for the full set of regression quantiles. However, we would caution the reader that there is no guarantee that $\beta_{h(\tau)} = \theta_\tau$ in general. For example, our mortgage application data 3.5.4 clearly has no co-located quantiles and expectiles except in the degenerate cases where $\tau = 0, 1$. In that example, expectiles seem to be preferable on their own merits.

### 3.5 Expectiles in Misspecified Mean Regressions

As stated in Proposition 1, the four assumptions for expectile regression nest the four Gauss-Markov assumptions when $\tau = .5$. We have shown that the expectile regression coefficients $\hat{\beta}_\tau = (X'WX)^{-1} X'Wy$ are the best linear unbiased estimator under those four expectile assumptions, which also implies that OLS is the BLUE when $\tau = .5$. When heteroscedasticity is present, the generalized expectile regression is the BLUE. Those results do not rely on any particular interpretation of the $\tau^{th}$ expectile coefficients. As such, we can extend the general reasoning to more specific applications.
In this section, we show that the (generalized) expectile regression coefficients are the BLUE when we have a standard model, with the four traditional Gauss-Markov assumptions, but these assumptions are violated in particular ways. The following three subsections consider three separate cases where Gauss-Markov assumptions are violated and a generalized expectile regression is the BLUE. Essentially, these justify the selection of the $\tau^{th}$ expectile as the line of greatest interest.

First, let us restate the Gauss-Markov assumptions as they appear in the context of a mean regression. Because we are only estimating the $\tau = .5^{th}$ expectile and the weights are symmetric, we drop the superfluous $\tau$ subscript from $\beta$.

**G-M Assumption 1:** The model is linear.

$$ y_i = \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + \epsilon_i = x_i' \beta + \epsilon_i $$

(3.25)

**G-M Assumption 2:** Strict Exogeneity:

$$ E(\epsilon_i | X) = 0 \quad \text{or} \quad \forall i \in \{1, ..., n\} . $$

(3.26)

**G-M Assumption 3:** Full Rank: $X$ is of full column rank i.e. $rank(X) = k$.

**G-M Assumption 4:** Spherical variance-covariance

$$ E(\epsilon \epsilon' | X) = \sigma^2 I_n. $$

(3.27)

We have already shown that the first and third assumptions do not change depending on which expectile we want to estimate. The fourth assumption is of little consequence, as we have already determined that we can find the BLUE with or without that assumption (see equation C.4). But the second assumption—strict exogeneity—will be violated in the following three examples. If the fourth is also violated, we may replace the ordinary expectile regression with the generalized expectile regression without loss of generality.
3.5.1 Expectiles for Subsample Contingency Analysis

A major overarching theme for expectile regression methods is the idea that heterogeneity exists within the data. This is true, at the very least, because the error terms $\epsilon_i$ are not uniquely zero and not all observations are “alike”. It is possible that the unknown and unobservable conditional distributions of $\epsilon_i$’s vary. The source of these unpredictable innovations $\epsilon_i$ is usually attributed to omitted variables, neglected complexity in the “true” model, or some other lack of understanding as to how $y_i$ came into being. The underlying factors are of interest, but difficult to analyze.

For subsamples whose $\epsilon_i$’s are not identical in distribution to the overall sample distribution, we may improve prediction by using tailored estimators—some form of contingency analysis. Suppose that the OLS model is not misspecified and Gauss-Markov assumptions 1-4 (or 1-3) are valid. In that case, OLS (or GLS) is BLUE per all the usual arguments. But it is not strictly true that OLS is the best estimator or produces the best predictor for all subsamples, given some additional information. We may construct contingency predictors and contingency estimators based on information beyond what is contained in $y$ and $X$. Expectiles are especially useful for this purpose.

Under Gauss-Markov assumptions 1-4, it is well known that

$$E(\hat{\beta}_{OLS}|X) = \beta$$
$$Var(\hat{\beta}_{OLS}|X) = \sigma^2(X'X)^{-1}. \quad (3.28)$$

Supposing we have a vector of covariates $x_i$, real or hypothetical, and we wish to predict $y_i$, we have the OLS predictor $x'_i\hat{\beta}_{OLS}$ with variance

$$Var(x'_i\hat{\beta}_{OLS}|X) = x'_iVar(\hat{\beta}_{OLS}|X)x_i$$
$$= \sigma^2x'_i(X'X)^{-1}x_i. \quad (3.29)$$

This is also the best linear unbiased predictor of $y_i$, which follows from our proposition 4 and from equation 3.20 in particular. However, if we know that an individual observation
with covariates $\mathbf{z}'$ has $\alpha \neq 1$ times the usual odds of a positive error term $\epsilon^* \geq 0$, relative to the unconditional distribution$^{10}$ of $\epsilon$, 

$$
\frac{\Pr(\epsilon^* \geq 0)}{\Pr(\epsilon^* < 0)} = \alpha \frac{\Pr(\epsilon_i \geq 0)}{\Pr(\epsilon_i < 0)}
$$

(3.30)

and we do not change the shape of the conditional distribution if the residual is positive or negative, then we have cause to doubt the predictor $\mathbf{z}' \hat{\beta}_{\text{OLS}}$. This new information about the error term violates two Gauss-Markov assumptions: strict exogeneity and spherical variance (assumptions 2 and 4)$^{11}$, at least for the coefficients $\hat{\beta}_{\text{OLS}}$, which must therefore be biased. Is there a different set of coefficients we should consider, which incorporates the information in equation 3.30? Yes, and one of the sample expectiles is the optimal linear predictor.$^{12}$ The following proposition states the result.

**Proposition 8.** Let Gauss-Markov assumptions 1-3 hold for the data $\mathbf{y}, \mathbf{X}$. For a given observation with covariates $\mathbf{z}$ and an atypical residual distribution as in equation 3.30, the optimal linear predictor is $\mathbf{z}' \tilde{\beta}_\tau$ where $\tilde{\beta}_\tau$ is the GLS-type expectile estimator, $\hat{\beta}_{\text{OLS},\text{GLS}} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1} \mathbf{X}'\Sigma^{-1}\mathbf{y}$, $\nu^2 \Sigma = E((\mathbf{y} - \mathbf{X}\beta_\tau)(\mathbf{y} - \mathbf{X}\beta_\tau)'|\mathbf{X}, W)$.

The proof in this case is only slightly more detailed than previous proofs, and it is given in Appendix C.3. We can also show that the best linear unbiased estimator of $\beta_\tau$ is $\hat{\beta}_{\text{OLS},\text{GLS}}$.

**Proposition 9.** Let Gauss-Markov assumptions 1-3 hold for the data $\mathbf{y}, \mathbf{X}$. For any

---

$^{10}$Both in-sample and out-of-sample prediction typically assume that the distribution of errors for the predicted observations is similar to that of the data. Relaxing that assumption can be achieved in a variety of ways, including the simple way given here.

$^{11}$Strict exogeneity is violated by construction. If the distribution of $\epsilon_i$ is $F$, then the distribution of the residual $\epsilon^*$ is the $\tilde{F}$ from equation 3.7 with $\frac{1 - \tau}{\tau} = \alpha$ (their definitions are the same). Let $\tau > .5$ without loss of generality. $E(\epsilon_i|\mathbf{X}) = E(w_i\epsilon_i|\mathbf{X}) = 0$ implies $$.5E(\epsilon_i I(\epsilon_i \geq 0)|\mathbf{X}) + .5E(\epsilon_i I(\epsilon_i < 0)|\mathbf{X}) = \tau E(\epsilon_i I(\epsilon_i \geq 0)|\mathbf{X}) + (1 - \tau)E(\epsilon_i I(\epsilon_i < 0)|\mathbf{X}) = 0,$

so, subtracting leaves:

$$
\begin{align*}
(5 - \tau)E(\epsilon_i I(\epsilon_i \geq 0)|\mathbf{X}) + (5 - 1 + \tau)E(\epsilon_i I(\epsilon_i < 0)|\mathbf{X}) &= 0
\end{align*}
$$

which is impossible so long as $\epsilon_i$ is not uniquely zero. The variance of a vector of these atypical observations may still be spherical, but the coefficient $\sigma^2$ must change because we have proven that the OLS predictor has the minimum possible variance! Notice that this also implies the condition in equation 3.33.

$^{12}$A straightforward calculation shows that $\frac{1 - \tau}{\tau} = \alpha$. For a specific, known contingency, we may use a specific $\tau$ in equation C.7. More generally, it is reasonable to perform the estimation using many different $\tau$ in order to survey the full spectrum of possible variation.
observation with a misspecified residual distribution as in equation 3.30, the best linear unbiased estimator of the coefficient vector $\beta_\tau$ is $\hat{\beta}_{\tau,\text{GLS}} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$, $\nu^2\Sigma = \Var(y - X\beta_\tau|X,W)$.

Following proposition 8, the proof of proposition 9 is simple. We have already shown that unbiasedness requires the predictor to use an unbiased estimator for $\beta_\tau$ and that the optimal linear predictor is $z'\hat{\beta}_{\tau,\text{GLS}}$ for any $z$, so it must hold in particular when $z$ is an elementary basis vector and thus the optimal linear predictor of $[\beta_\tau]_j$ is $[\hat{\beta}_{\tau,\text{GLS}}]_j$ for any $j$ in $1,...,k$. A proof that $\hat{\beta}_{\tau,\text{GLS}}$ has the minimum variance of any unbiased estimator for $\beta_\tau$ is exactly that given in equation C.6. Under propositions 8-9, the expectile coefficients $\beta_\tau$ retain the interpretation as partial derivatives of the expected value of the data with respect to $z_i$. That is, for the atypical observation $y^*$,

$$\frac{\partial E(y^*)}{\partial z_j} = \beta_{\tau,j} \forall j \in \{1,...,k\}. \tag{3.31}$$

and the expectile coefficients $\hat{\beta}_{\tau,\text{GLS}}$ are the optimal estimators of these average treatment effects for the atypical observation.

In other words, if a hypothetical individual were twice as likely as others to be above the regression line–*ceteris paribus*–then the $\tau = \frac{2}{3}$ *sample* expectile regression line would be the mean regression line for that individual. This interpretation allows us to explore the mortgage denial data–a binary response problem–where quantile regression is not applicable. In those data, some individuals are more or less likely to be denied a mortgage, even conditional on other covariates. Average treatment effects for the full spectrum of contingencies are considered in section 3.5.4 on page 102. The following two subsections develop two additional motivations for the expectile, but we consider the interpretation given in this subsection to be the primary motivation. If the reader wishes to see these estimators in action, and the results that follow, skip directly to section 3.5.4 on page 102.
3.5.2 Expectiles for Missing Data

Similar reasoning is applicable when data are missing not-at-random, but asymmetrically. Suppose that the “true” data generating process \( y_i^\ast = x_i^\ast \beta + \epsilon_i^\ast \) has an unknown distribution of errors \( \tilde{F} \) as in equation 3.7, but the data is not perfectly observed and observations are missing. If we only know that positive error terms \( \epsilon^* \geq 0 \) are \( \alpha \) times as likely to go missing as negative error terms, we have

\[
\frac{\Pr(\epsilon_i \geq 0)}{\Pr(\epsilon_i < 0)} = \frac{1}{\alpha} \frac{\Pr(\epsilon^* \geq 0)}{\Pr(\epsilon^* < 0)} \tag{3.32}
\]

where the observed data have \( \frac{1}{\alpha} \) times as many positive error terms. Clearly, this is the same as equation 3.30. So we can incorporate this information as in equation C.7 to produce a useful estimator. If \( E(\epsilon^*|X^*) = 0 \) for the true data generating process but equation 3.32 is true for the observed data, then \( E(w_i \epsilon_i|X) = 0 \) is true for the sample as in equation C.7.

Then we may solve for the optimal estimator by the method in the previous section.

**Proposition 10.** Let Gauss-Markov assumptions 1-3 hold for the true data generating process of \( y^*, X^* \). For data \( y, X \) with missing observations as in 3.32, the best linear unbiased estimator for the true DGP is the GLS expectile estimator, \( \hat{\beta}_{\tau,GLS} = (X\Sigma^{-1}X)^{-1}X\Sigma^{-1}y, \nu^2\Sigma = E(\epsilon\epsilon'|X,W) \).

The proof of this proposition is exactly the same as before. There is some \( \tau \in (0,1) \) such that \( E_\tau(\epsilon_i|X) = 0 \). Then any unbiased linear estimator can be written \( \hat{\beta} = Cy \) with \( E_\tau(\hat{\beta}|X) = X\beta_\tau \), where \( \beta_\tau \) is the true \( \tau^{th} \) expectile coefficient vector for the observed data. Then equation C.6 shows that \( \hat{\beta}_{\tau,GLS} \) is the BLUE in this class of estimators. Alternately, we might characterise this example as the same as in Section 3.5.1 on page 95, except with the caveat that *every* observation comes from the same atypical distribution. Then the same proofs are applicable.

The choice of the optimal linear predictor for some covariates \( z \) depends on whether we are interested in data before or after observations go missing. If we are interested in the observed data only, then mean regression methods would still be appropriate. If we are interested in prediction from the underlying data generating process, the optimal predictor
Proposition 11. Let Gauss-Markov assumptions 1-3 hold for the true data generating process of $y^*, X^*$, with observations missing as in 3.32. For an observation with covariates $z$, the best linear unbiased predictor for the true data generating process is $z' \hat{\beta}_{\tau, GLS}$ with 

$$\hat{\beta}_{\tau, GLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y,$$

$$\nu^2 \Sigma = E(\epsilon\epsilon' | X, W).$$

Missing data problems are common in many disciplines, and it is often possible that missing data are skewed in one direction or the other. In job training programs, for instance, individuals who fail to complete the training program are probably less likely than average to get a job. Because those data are not missing-at-random, OLS estimates of the probability of finding a job would be biased, but another expectile may be unbiased. Importantly, other asymmetric model designs (frontier regression, for instance) may motivate the use of asymmetric weights as in this section and the previous section. In such cases, the design of the estimator and its optimality among linear estimators could be shown using similar principles.

3.5.3 Expectiles: Relaxed Exogeneity

The most obvious interpretation of the $\tau$th expectile is made clear in subsections 3.5.1 and 3.5.2. But these are not the only cases where expectiles are the BLUE. To step away from the usual mean regression while maintaining the rest of the regression problem’s structure, we relax assumption 2 in the smallest way possible:

$$E(\epsilon_i | X) = c \forall i. \quad (3.33)$$

We have merely altered our assumption so that the average error term is not zero, but some other constant $c \in \mathbb{R}$. In financial applications, for example, it is typical to prefer assets with positive (non-zero) expected returns. But in any case, the mean regression line is not the only line of interest. And, in heteroscedastic models, we that the rest of the regression coefficients $\beta_2, ..., \beta_k$ vary with the estimated constant. This was shown in Figure 3.4 on page 79.
This is a major motivation for quantile-type methods: we are interested in regression lines that pass through different levels of the dependent variable, i.e. we wish to estimate models with multiple different $c$ in equation 3.33. The non-zero exogeneity condition implies the following about our model:

**Proposition 12.** Let Assumptions 1,3,4 hold, let $E(\epsilon_i|X) = c$ with $c \in \mathbb{R}$ a (finite) constant, and $0 < \Pr(\epsilon_i \geq 0) < 1$. Then there exist expectile weights $w_i$ of the form

$$
w_i = \begin{cases} 
\tau & \text{if } \epsilon_i \geq 0 \\
1 - \tau & \text{if } \epsilon_i < 0 
\end{cases}
$$

such that $E(w_i \epsilon_i|X) = 0$, for some $\tau \in [0,1]$

The proposition above leads to an interesting outcome. If strict exogeneity fails as in equation 3.33 on the preceding page, and $E(\epsilon_i|X) = c$, then there must be a $\tau \in (0,1)$ such that expectile weighted exogeneity holds\textsuperscript{13}.

\[ E(\epsilon_i|X) \neq 0, \quad \text{but} \quad E_\tau(\epsilon_i|X) = E(w_i \epsilon_i|X) = 0. \quad (3.34) \]

Obviously, this violates the Gauss-Markov assumptions and the OLS estimator is not the BLUE—it is biased. However, there is some $\tau^{th}$ expectile such that the $\tau^{th}$ expectile regression estimator is unbiased. An explicit formula for that $\tau$ is given in C.19 in the appendix. Then the following two propositions are in place. First, the OLS estimator is biased. Second, an expectile regression is unbiased.

**Proposition 13.** Let Gauss-Markov assumptions 1,3,4 hold and $E(\epsilon_i|X) = c$ for all $i = 1,...,n$. The OLS estimator is biased if $c \neq 0$.

\textsuperscript{13}This, as in the proposition, excludes the degenerate case where all $\epsilon_i$ are positive or that where they are all negative. That is implied by the condition $0 < \Pr(\epsilon_i \geq 0) < 1$. Obviously, there are few examples where the desired regression line is outside the range of the data as implied by these two cases.
Proof. The proof is very obvious. From the definition, we have

\[ E(\hat{\beta}_{OLS}|X) = E((X'X)^{-1}X'y|X) \]
\[ = E((X'X)^{-1}X'(X\beta + \epsilon)|X) \]
\[ = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'E(\epsilon|X) \]
\[ = \beta + (X'X)^{-1}X'1_n c \]

and for \( c \neq 0 \), the latter term is not zero unless \( X = 0_{n \times k} \) as well, which violates Assumption 3.

The unbiasedness of the expectile coefficients \( \hat{\beta}_\tau = (X'WX)^{-1}X'Wy \) (with \( \tau \) that satisfies equation 3.34) is simple to prove. From the fact that \( E_\tau(\epsilon_i|X) = 0 \), the proof is the same as that given in equation C.3 on page 212. We have the following proposition.

**Proposition 14.** Let Gauss-Markov assumptions 1, 3, 4 hold and \( E(\epsilon_i|X) = c \). Then the expectile regression estimator \( \hat{\beta}_\tau = (X'WX)^{-1}X'Wy \) with \( \tau \) given in equation C.19 is an unbiased estimator of \( \beta \).

Because \( \hat{\beta}_\tau = (X'WX)^{-1}X'Wy \) is unbiased, the variance is identical to that given in equation C.4 on page 212 for general \( E(\epsilon\epsilon'|X, W) = \Sigma \).

Naturally, this extends to predictors of \( AX\beta \), including \( x_i\beta \). The OLS predictor is biased, but the expectile predictor \( AX\hat{\beta}_\tau \) is unbiased. Under heteroscedasticity, we may rely on the sandwich variance formula or use the GLS estimator discussed in the previous section.

**Proposition 15.** Let Gauss-Markov assumptions 1, 3 hold. Let \( E(\epsilon_i|X) = c \), and let \( E(\epsilon\epsilon'|X, W) = \nu^2\Sigma = \nu^2W^{-1/2}\Omega W^{-1/2} \) for some diagonal positive definite \( \Omega \). Then the expectile regression estimator \( \hat{\beta}_\tau = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \) with \( \tau \) given in equation C.19 is the best linear unbiased estimator of \( \beta \).

The proof in section 3.4.2 on page 90 is applicable because \( E_\tau(\epsilon_i|X) = 0 \). Specifically, equation C.6 shows that this estimator must be the BLUE. As you can see, the expectile is a natural generalization of the mean regression to a non-central regression design implied
by equation 3.33. Under general variance conditions, our expectile GLS estimator is the BLUE.

3.5.4 Application: Mortgage Denial

For the mortgage denial data of Munnell et al. [1996], the interpretation in section 3.5.1 on page 95 is directly applicable. The $\tau$th expectile of $Deny$ is the average for individuals who are $\frac{\tau}{1-\tau}$ times as likely as average to be denied, ceteris paribus. And because the model is a linear probability model,

$$E(Deny_i|x_i) = Pr(Deny_i = 1|x_i)$$

the expectile weights for any $\tau \in (0,1)$ are known a priori. Individuals who are denied are above the regression line and receive a weight of $\tau$, while the others receive a weight of $1 - \tau$. Thus, the standard expectile regression is perfectly feasible (as is OLS) in this application. The GLS-type weights are still infeasible. A more formal discussion of feasibility is given in Appendix C.4.

The research question is whether there is a statistically significant difference in mortgage application denial based on race. There is. In the simple regression of $Deny$ on $Black$, (Figure 3.2 on page 77) we show that black applicants were 19% more likely to be denied on average, but a credit-disadvantaged individual at the .83 expectile is as much as 33% more likely to be denied if they are black. Does this result persist when additional explanatory variables are added? Table 3.3 on the following page shows the answer. With a full slate of additional covariates used by Munnell et al. [1996], the size of the racial disparity decreases by about half—but it cannot be eliminated. At the $\tau = .9$ expectile, the coefficient on $Black$ is .15.

Figures 3.6 and 3.7 show the shape of the effect of each variable on $Deny$, as a function of $\tau$. In the first figure, we see that the effect of $Black$ on $Deny$ has exactly the same shape as it had in figure 3.3, although the effect has decreased by about half. Figure 3.7 shows that most other variables have the same shape. The payment/income ratio $pi\_rat$, the
<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>black</td>
<td>0.017***</td>
<td>0.049***</td>
<td>0.084***</td>
<td>0.124***</td>
<td>0.150***</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.014)</td>
<td>(0.023)</td>
<td>(0.032)</td>
<td>(0.038)</td>
</tr>
<tr>
<td>pi_rat</td>
<td>0.140***</td>
<td>0.299***</td>
<td>0.449***</td>
<td>0.655***</td>
<td>0.895***</td>
</tr>
<tr>
<td></td>
<td>(0.044)</td>
<td>(0.079)</td>
<td>(0.114)</td>
<td>(0.157)</td>
<td>(0.194)</td>
</tr>
<tr>
<td>hse_inc</td>
<td>0.038</td>
<td>0.025</td>
<td>-0.048</td>
<td>-0.166</td>
<td>-0.288</td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td>(0.069)</td>
<td>(0.110)</td>
<td>(0.167)</td>
<td>(0.237)</td>
</tr>
<tr>
<td>ltv_med</td>
<td>0.008**</td>
<td>0.018**</td>
<td>0.031**</td>
<td>0.051**</td>
<td>0.073**</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.007)</td>
<td>(0.013)</td>
<td>(0.021)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>ltv_high</td>
<td>0.039**</td>
<td>0.115***</td>
<td>0.189***</td>
<td>0.250***</td>
<td>0.268***</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td>(0.037)</td>
<td>(0.050)</td>
<td>(0.057)</td>
<td>(0.062)</td>
</tr>
<tr>
<td>ccred</td>
<td>0.006***</td>
<td>0.017***</td>
<td>0.031***</td>
<td>0.048***</td>
<td>0.062***</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.003)</td>
<td>(0.005)</td>
<td>(0.007)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>mcred</td>
<td>0.007**</td>
<td>0.014*</td>
<td>0.021*</td>
<td>0.034*</td>
<td>0.055*</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.007)</td>
<td>(0.011)</td>
<td>(0.018)</td>
<td>(0.030)</td>
</tr>
<tr>
<td>pubrec</td>
<td>0.050***</td>
<td>0.131***</td>
<td>0.197***</td>
<td>0.237***</td>
<td>0.195***</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.025)</td>
<td>(0.035)</td>
<td>(0.040)</td>
<td>(0.039)</td>
</tr>
<tr>
<td>denpmi</td>
<td>0.520***</td>
<td>0.708***</td>
<td>0.702***</td>
<td>0.669***</td>
<td>0.560***</td>
</tr>
<tr>
<td></td>
<td>(0.126)</td>
<td>(0.069)</td>
<td>(0.045)</td>
<td>(0.038)</td>
<td>(0.036)</td>
</tr>
<tr>
<td>selfemp</td>
<td>0.014***</td>
<td>0.036***</td>
<td>0.060***</td>
<td>0.095***</td>
<td>0.145***</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.012)</td>
<td>(0.021)</td>
<td>(0.032)</td>
<td>(0.047)</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.081***</td>
<td>-0.146***</td>
<td>-0.183***</td>
<td>-0.215***</td>
<td>-0.124*</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.024)</td>
<td>(0.028)</td>
<td>(0.039)</td>
<td>(0.072)</td>
</tr>
<tr>
<td>Observations</td>
<td>2,380</td>
<td>2,380</td>
<td>2,380</td>
<td>2,380</td>
<td>2,380</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.107</td>
<td>0.210</td>
<td>0.266</td>
<td>0.301</td>
<td>0.243</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 3.3: Regression coefficients for the full spectrum: $\tau = .1, .3, .5, .7, .9$, with many covariates added. The effect of black shrinks slightly but is statistically significant at all expectiles shown even with nine additional explanatory variables in the model. In this case, the largest effect of black occurs at $\tau = .9$ and is roughly 15%. Unsurprisingly, the coefficients for most variables increase with $\tau$: it is the individuals with a high probability of denial who also experience the largest marginal effects.
Figure 3.6: As in the simple regression, the effect of \textit{Black} on \textit{Deny} is larger for $\tau > .5$, individuals who are more likely than average to be denied. The estimated effect reaches its maximum of .15 at $\tau = .89$. This is nearly double the OLS coefficient (.84).

Figure 3.7: Similar to \textit{Black}, most variables have a larger effect on individuals who were already credit-disadvantaged, i.e. $\tau > .5$. The notable exception is \textit{denpmi}, which has a mostly flat effect but is slightly larger for $\tau < .5$. 
<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>τ</td>
<td>.1</td>
<td>.3</td>
<td>.5</td>
<td>.7</td>
<td>.9</td>
</tr>
<tr>
<td>black</td>
<td>0.012*** (0.003)</td>
<td>0.041*** (0.012)</td>
<td>0.081*** (0.022)</td>
<td>0.137*** (0.033)</td>
<td>0.108*** (0.025)</td>
</tr>
<tr>
<td>pi_rat</td>
<td>0.047*** (0.008)</td>
<td>0.174*** (0.029)</td>
<td>0.377*** (0.061)</td>
<td>0.617*** (0.115)</td>
<td>0.871*** (0.142)</td>
</tr>
<tr>
<td>hse_inc</td>
<td>0.008 (0.010)</td>
<td>0.033 (0.020)</td>
<td>0.063 (0.077)</td>
<td>0.110 (0.122)</td>
<td>-0.262 (0.162)</td>
</tr>
<tr>
<td>ltv_med</td>
<td>0.002*** (0.001)</td>
<td>0.008** (0.004)</td>
<td>0.019*** (0.005)</td>
<td>0.044** (0.017)</td>
<td>0.076** (0.031)</td>
</tr>
<tr>
<td>ltv_high</td>
<td>0.032*** (0.011)</td>
<td>0.095*** (0.034)</td>
<td>0.152*** (0.053)</td>
<td>0.256*** (0.046)</td>
<td>0.130*** (0.036)</td>
</tr>
<tr>
<td>ccred</td>
<td>0.004*** (0.001)</td>
<td>0.013*** (0.002)</td>
<td>0.027*** (0.004)</td>
<td>0.051*** (0.006)</td>
<td>0.063*** (0.007)</td>
</tr>
<tr>
<td>mcred</td>
<td>0.002*** (0.001)</td>
<td>0.009*** (0.002)</td>
<td>0.019*** (0.006)</td>
<td>0.029*** (0.011)</td>
<td>0.070*** (0.025)</td>
</tr>
<tr>
<td>pubrec</td>
<td>0.041*** (0.008)</td>
<td>0.129*** (0.024)</td>
<td>0.211*** (0.033)</td>
<td>0.239*** (0.038)</td>
<td>0.126*** (0.027)</td>
</tr>
<tr>
<td>denpmi</td>
<td>0.525*** (0.128)</td>
<td>0.735*** (0.037)</td>
<td>0.716*** (0.029)</td>
<td>0.838*** (0.020)</td>
<td>0.593*** (0.029)</td>
</tr>
<tr>
<td>selfemp</td>
<td>0.008*** (0.002)</td>
<td>0.030*** (0.007)</td>
<td>0.064*** (0.017)</td>
<td>0.115*** (0.028)</td>
<td>0.165*** (0.038)</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.022*** (0.002)</td>
<td>-0.083*** (0.010)</td>
<td>-0.175*** (0.018)</td>
<td>-0.282*** (0.034)</td>
<td>-0.136*** (0.064)</td>
</tr>
<tr>
<td>Observations</td>
<td>2,133</td>
<td>2,130</td>
<td>2,132</td>
<td>2,153</td>
<td>2,331</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.017</td>
<td>0.214</td>
<td>0.472</td>
<td>0.584</td>
<td>0.547</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3.4: Regression coefficients for the full spectrum: τ = .1, .3, .5, .7, .9, with many covariates added. The GLS estimator is used with ω_i = (x_i′β(1 − x_i′β))^{-1}. Results largely resemble the previous table, with the coefficient on black reaching a value of approximately 14%.
total housing expense to income ratio $hse_{inc}$, the loan-to-value ratio indicators $ltv_{med}$ and $ltv_{high}$, the consumer’s credit score $cred$, the mortgage credit score $mcred$, and the public bad credit record indicator $pubrec$ all have this shape, as does the self-employed indicator $selfemp$. The only obvious exception is $denpmi$, an indicator for individuals who were denied private mortgage insurance. This seems to have a relatively constant—and large—effect across the board, though the effect on $\tau < .5$ is slightly larger than higher expectiles.

For linear probability models, Goldberger et al. [1964] suggested using generalized least squares to improve the efficiency of the estimator. Because the response variable is binary (for any $\tau \in (0, 1)$), we have

$$Var(\epsilon_i|X) = x_i^\prime \beta_\tau (1 - x_i^\prime \beta_\tau).$$

(3.36)

Setting weights equal to the inverse of this expression produces consistent estimates [McGillivray, 1970] and has been adopted widely for linear probability models [Greene, 2003, p. 727]. We incorporate that approach for our expectile linear probability model and show results in Table 3.4. These are similar to those in Table 3.3, with the maximum coefficient for black reaching approximately 14%. In the GLS case, however, the maximum coefficient is obtained at $\tau = .7$, while the OLS coefficient is maximized at $\tau = .9$. When we consider every centile of the unit interval for $\tau$, we find that GLS coefficient on Black is .174 and occurs at $\tau = .83$. With GLS, the number of observations in the GLS results varies because observations with fitted values outside the unit interval have undefined variance per 3.36. As a result, these observations dropped as the model iterates to convergence. The expectile coefficient plots with GLS reproduce the main results from Figures 3.6 and 3.7 albeit with a partial loss of smoothness for high $\tau$. These are given in B.3.3.

In each figure each table, the economic result is the same. A standard OLS estimator does not represent the full spectrum of possibilities. Rather, it reports only the unconditional average. In atypical cases, the difference between black and non-black distributions is smaller (for very low $\tau$ or extremely high $\tau$) or larger (for $\tau$ between .6 and .9, say) than usually reported. This type of heterogeneity has been previously unexplored.
3.6 Ancillary Statistics

Neither the $R^2$ statistic nor the estimated variance and mean squared error of expectile residuals have been studied thoroughly. In this section, we develop those statistics and give some remarks. As an example, assume that $\epsilon \mid X \sim (0, \sigma^2 \Sigma)$ so that $\beta$ is the “true” mean regression coefficient vector. Then

$$Var(y \mid X) = E((y - X\beta)(y - X\beta)\mid X) = E(\epsilon\epsilon \mid X) = \sigma^2 \Sigma \quad (3.37)$$

We have a shape parameter $\Sigma$ and a scale parameter $\sigma^2$. The shape parameter $\Sigma$ requires further assumptions to identify: see Magnus [1978]. Tools are available for the mean regression, as in White et al. [1980]. These tools can be adapted to expectile regression, but this subject is large and situationally dependent. We cannot address it properly here. As in previous sections, we will restrict our attention to the independent case where $E(\epsilon\epsilon \mid X)$ is a diagonal matrix.

Note also that the expectile variance parameter $\sigma^2_\tau$ is not the mean variance $\sigma^2$ nor is it the $\sigma^2_i$ of the location-scale model which is popular in the generalized quantile literature [Aragon et al., 2005, Bianchi et al., 2018], which varies with $i$ for a given $\tau$. Instead, we presume that multiple expectiles are of interest and note that they do not have the same mean squared error, even under simple conditions.

Moreover, the problem of estimated variance of the residuals is bifurcated when the location parameter is non-central. We have three different statistics of interest:

1. The variance of residuals, $Var(\epsilon_i \mid X)$.
2. The weighted variance or mean squared error (under the weighted distribution $\tilde{F}$ in 3.7).
3. The mean squared error of residuals, $E(\epsilon_i^2 \mid X)$.

Because the residuals are not zero on average except when $\tau = .5$, the variance of the
residuals will no longer be equal to the mean squared error. However, under the weighted distribution $\tilde{F}$ both are the same: this was proven in equation 3.18 on page 86.

The expected error itself is also interesting, but it is merely the difference between the estimated expectile and the estimated mean regression:

$$E(\epsilon_i|X) = E(y_i - x_i'\beta_\tau + x_i'\beta_{.5} - x_i'\beta_{.5}|X)$$
$$= E(y_i - x_i'\beta_{.5}|X) - x_i'\beta_\tau + x_i'\beta_{.5}$$
$$= x_i'\beta_{.5} - x_i'\beta_\tau. \quad (3.38)$$

Thus, the variance of residuals $Var(\epsilon_i|X)$ can be obtained as

$$Var(\epsilon_i|X) = Var(y_i - x_i'\beta_\tau - E(\epsilon_i|X)|X)$$
$$= Var(y_i - x_i'\beta_{.5}|X) \quad (3.39)$$

which is merely the OLS or GLS residuals. This requires no special treatment\textsuperscript{14}.

Likewise, the weighted variance is equal to the weighted mean squared error:

$$WVar(\epsilon_i|X) = E_\tau(\epsilon_i^2|X), \quad (3.40)$$

see equation 3.18 on page 86. Thus, when we take the interpretations from section 3.5 on page 93 seriously, we may employ estimators such as the standard weighted variance estimator

$$\frac{n}{n-k} \left( \sum_{i=1}^{n} w_i \right)^{-1} \sum_{i=1}^{n} w_i \hat{\epsilon}_i^2. \quad (3.41)$$

In contrast, we use an un-weighted estimator to estimate the expectile mean squared error. The typical estimator

$$s^2 = \frac{\hat{\epsilon}' \hat{\epsilon}}{n-k} \quad (3.42)$$

is usable but not necessarily optimal. This is discussed in the next section.

\textsuperscript{14}The fact that $Var(y_i - x_i'\beta_\tau|X) = Var(y_i - x_i'\beta_{.5}|X)$ for all $\tau$ in the location-scale model suggests using the estimated GLS weights $\hat{\Omega}$ from a standard GLS problem for all expectiles. Others may wish to explore the practicality of this approach.
As an aside, the scale parameter $\sigma^2_\tau$ for the mean squared error of $(y - X\beta_\tau)$ will vary across $\tau$. This means that the residuals $\epsilon_i$ can become extremely large as the predictor $X\beta_\tau$ moves far into the tail of an unbounded distribution. Even if $y_i|X$ are conditionally i.i.d., we have $E(\epsilon\epsilon'|X) = \sigma^2_\tau I_n$ where the constant $\sigma^2_\tau$ will vary depending on $\tau$. From the classic literature, we know that $\sigma^2_\tau$ is minimized when $\tau = .5$. The following subsections are devoted to estimating $\sigma^2_\tau$.

### 3.6.1 Estimated Mean Squared Error

The “usual” OLS estimators for the residual mean squared error can be adapted to the expectile regression environment. For the case where $\tau = .5$, estimators for $\sigma^2$ include the Gaussian MLE $\hat{\sigma}^2$ and the “unbiased” moment-based estimator $s^2$. These can be used without modification for non-central expectile MSE by employing the expectile residuals $\hat{\epsilon} = y - X\hat{\beta}_\tau$. Then they are

$$\hat{\sigma}^2_\tau = \frac{\hat{\epsilon}'\hat{\epsilon}}{n}, \quad s^2_\tau = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-k}. \tag{3.43}$$

For $\tau = .5$, the latter was shown to be unbiased by Gauss [1823]. That classical result was reformulated by Aitken [1936] using matrix algebra and is found in nearly all elementary econometrics textbooks today. We will address these traditional estimators jointly by focusing first on the inner product of residuals, $\hat{\epsilon}'\hat{\epsilon}$. With the assumption that errors are i.i.d., the covariance matrix will be diagonal and we have, effectively, $n$ observations of the same error distribution. And

$$E(\epsilon\epsilon'|X) = E((y - X\beta_\tau)(y - X\beta_\tau)'|X)) = \text{diag}(\sigma^2_\tau). \tag{3.44}$$

However, the properties of estimators in equation 3.43 that are well-known in the OLS case do not extend to expectile regression. This is shown below. Taking the sum of squared residuals $\hat{\epsilon}'\hat{\epsilon}$, we derive its expected value as a function of $\sigma^2_\tau$. This is a standard way to prove unbiasedness of $s^2$ for the mean regression. Notice: with the annihilator matrix $M_\tau$
we have
\[
\hat{\epsilon} = y - \hat{X}\hat{\beta} = M_\tau y = M_\tau (X\beta_\tau + \epsilon) = M_\tau \epsilon. \tag{3.45}
\]

So we may write the sum of squared residuals as
\[
E(\hat{\epsilon}'\hat{\epsilon}|X, W) = E(\epsilon'M_\tau' M_\tau \epsilon|X, W)
\]
\[
= \sum_{i=1}^{n} \sum_{j=1}^{n} [M_\tau' M_\tau]_{ij} E(\epsilon_i \epsilon_j|X, W)
\]
\[
= \sum_{i=1}^{n} [M_\tau' M_\tau]_{ii} \sigma^2_\tau. \tag{3.46}
\]

This follows from the independence of \(\epsilon_i\)'s and the definition of matrix multiplication.

Nearly the same result was given by Aitken [1936] where the last line will reduce to \(\sigma^2(n - k)\) in the special case \(\tau = .5\). But for any other expectile, the oblique annihilator matrix \(M_\tau\) is not symmetric and \(M_\tau'M_\tau\) does not simplify. Of course, the last expression above is \(\sigma^2_\tau \times \text{trace}(M_\tau'M_\tau)\) where

\[
\text{trace}(M_\tau'M_\tau) = \text{trace}((I - P_\tau)'(I - P_\tau))
\]
\[
= \text{trace}(I - P_\tau - P_\tau' + P_\tau'P_\tau). \tag{3.47}
\]

We will address this piecewise. The trace of \(I\) is \(n\). Next,

\[
\text{trace}(P_\tau) = \text{trace}(X(X'WX)^{-1}X'W)
\]
\[
= \text{trace}((X'WX)^{-1}X'WX)
\]
\[
= k. \tag{3.48}
\]
and the same for $\text{trace}(P'_{\tau}P_{\tau})$. Then

\[
\text{trace}(P'_{\tau}P_{\tau}) = \text{trace}(WX(X'WX)^{-1}X'X(X'WX)^{-1}X'W)
\]
\[
= \text{trace}((X'WX)^{-1}X'X(X'WX)^{-1}X'WX)
\]
\[
= \text{trace}\left(\left(\sum_{i=1}^{n} w_{i}x_{i}x_{i}'\right)^{-1}\left(\sum_{i=1}^{n} x_{i}x_{i}'\right)\left(\sum_{i=1}^{n} w_{i}x_{i}x_{i}'\right)^{-1}\left(\sum_{i=1}^{n} w_{i}^{2}x_{i}x_{i}'\right)\right).
\]

(3.49)

As you see, the trace of $P'_{\tau}P_{\tau}$ is a random variable that will depend on the data generating process. But, conditional on $X$ and $W$, we have the following lemma.

**Lemma 16.** Let $y, X$ be as above and $\hat{\epsilon} = y - X\hat{\beta}_{\tau}$ with $\hat{\beta}_{\tau} = (X'WX)^{-1}X'Wy$. Then

\[
E(\hat{\epsilon}'\hat{\epsilon}|X, W) = \sigma_{\tau}^2(n - 2k + \text{trace}(P'_{\tau}P_{\tau})).
\]

(3.50)

So, neither variance estimator in equation 3.43 will be unbiased except in a special case where $\text{trace}(P'_{\tau}P_{\tau}) = k$ or $\text{trace}(P'_{\tau}P_{\tau}) = 2k$. The classic result that $s^2$ is unbiased and $\hat{\sigma}^2$ is biased downwards does not hold except for $\tau = .5$. Rather, $s^2$ is unbiased if and only if $\text{trace}(P'_{\tau}P_{\tau}) = k$, which is if and only if $\tau = .5$. For all other values of $\tau$, the estimator $s^2$ is biased upwards. But there are special cases where $\text{trace}(P'_{\tau}P_{\tau}) = 2k$, and the standard MLE estimator $\hat{\sigma}^2$ is unbiased. In practice, neither estimator is especially desirable for unbiasedness. Moreover, the value of $\text{trace}(P'_{\tau}P_{\tau})$ can exceed $2k$ for extreme expectiles (see Figure 3.8) in very skew distributions, which indicates that both variables may be biased upwards under these circumstances. The limiting value $\lim_{n \to \infty} \text{trace}(P'_{\tau}P_{\tau})$ of the trace of the inner product of the two projection matrix is shown in Figures 3.8 and 3.9. In the former, $k = 1$. In the latter, $k = 5$. We will discuss the properties of this function in 3.6.2 on page 113. We will also show that the sampling distribution for small $n$ can differ significantly from is limiting distribution; see Figure 3.10 on page 116.

Notice that equation 3.50 implies the following result.

**Lemma 17.** Let $y, X$ be as above and $\hat{\epsilon} = y - X\hat{\beta}_{\tau}$ with $\hat{\beta}_{\tau} = (X'WX)^{-1}X'Wy$. Then the
Figure 3.8: The trace of the inner product of the expectile projection matrix $P_\tau$ with itself: \( \text{trace}(P_\tau^t P_\tau) \). The trace is shown as a function of $\tau$ and of $F(\mu_\tau)$ i.e. what proportion of observations are given weight $w_i$ equal to $1 - \tau$. In this case, $X = 1_n$, so $k = 1$. The minimum, $\text{trace}(P_\tau^t P_\tau) = k = 1$, occurs wherever $\tau = .5$ and not otherwise.

Figure 3.9: The trace of the inner product of the expectile projection matrix $P_\tau$ with itself: \( \text{trace}(P_\tau^t P_\tau) \). The trace is shown as a function of $\tau$ and of $F(\mu_\tau)$ i.e. what proportion of observations are given weight $w_i$ equal to $1 - \tau$. In this case, $k = 5$. The minimum, $\text{trace}(P_\tau^t P_\tau) = k = 5$, occurs wherever $\tau = .5$ and not otherwise.
estimator

\[
\hat{\sigma}_{\tau,n}^2 = \left( \frac{y - X\hat{\beta}_\tau}{n - 2k + \text{trace}(P_\tau'P_\tau)} \right)'
\]

is unbiased; i.e. \( E(\hat{\sigma}_{\tau,n}^2 | X, W) = \sigma_{\tau}^2 \).

The proof is obvious, given the preceding lemma where \( E(\epsilon'\epsilon | X, W) \) is shown. As a practical matter it should make little difference: the usual estimators in 3.43 will be extremely close to each other and to the revised estimator in 3.51 when \( n \) is sufficiently large and the skewness of the distribution in question is not extreme. However, the difference may be noticeable in small samples. To employ the revised estimator, evaluation of the trace of \( P_\tau'P_\tau \) numerically is only slightly more difficult than evaluation of the estimator \( \hat{\beta}_\tau \) itself. In total, \( P_\tau'P_\tau \) is of comparable complexity to the sandwich variance in equation C.4.

If evaluating \( \text{trace}(P_\tau'P_\tau) \) may be costly for very large or high-dimensional data, a further simplified estimator with desirable properties is \( \frac{n\sum w_i^2}{(\sum w_i)^2} \times k \). This becomes clear in the next section.

### 3.6.2 Consistency of MSE Estimators

Results relating to the asymptotic performance of the estimated expectile regression coefficients \( \hat{\beta}_\tau \) can be found in the literature. Newey and Powell [1987] provide broad conditions for consistency and asymptotic normality of the estimator in the linear regression case. Working papers by Barry et al. [2018] and Philipps [2021a] present asymptotic results for the weighted regression coefficients. Holzmann and Klar [2016] investigate the asymptotic properties of the expectile more thoroughly for the location model.

Here, we show that the proposed estimator in equation 3.51 is consistent. We also show that the estimators in equation 3.43 are consistent. Under the assumption that the sequence \( \{y_i, x_i\} \) is independently drawn from a location-scale model, we have \( \text{Pr}(w_i = \tau) \) constant and \( W, X \) are independent. Then as \( n \to \infty \) we have the following.

**Lemma 18.** Let \( W, X \) be independent and \( n^{-1}X'X \xrightarrow{p} Q_X \) for some matrix \( Q_X \) as \( n \to \infty \). Then
\[ n^{-1}X'WX = \frac{1}{n} \sum_{i=1}^{n} w_i x_i' \xrightarrow{p} E(w_i x_i') = E(w_i)E(x_i'). \] (3.52)

The above statement follows from independence and the definition of expectations. The location-scale model (with independence of \(W, X\)) is not required for consistency of \(\hat{\beta}_\tau\) under misspecification, but it is required for the simple result below.

**Lemma 19.** Let \(n^{-1}X'WX \xrightarrow{p} E(w_i)E(x_i')\) as \(n \to \infty\) and \(E(x_i')\) have rank \(k\). Then

\[
\text{trace}(P_i'P_i) = \text{trace} \left( \left( \sum_{i=1}^{n} w_i x_i x_i' \right)^{-1} \left( \sum_{i=1}^{n} w_i x_i x_i' \right) \right).
\]

\[ \xrightarrow{p} \text{trace} \left( (nE(w_i)Q_X)^{-1} (nQ_X) (nE(w_i)Q_X)^{-1} (nE(w_i^2)Q_X) \right) \]

\[ = \frac{n^2E(w_i^2)}{n^2E(w_i)^2} \text{trace} \left( Q_X^{-1}Q_X Q_X^{-1}Q_X \right) \]

\[ = \frac{E(w_i^2)}{E(w_i)^2} \times k \] (3.53)

The result follows from independence and the continuous mapping theorem. The obvious estimator for the ratio in the last line is \(\frac{n \sum w_i^2}{\sum w_i^2}\), which is \(O_p(1)\) as both its numerator and denominator are \(O_p(n^2)\). The value of this ratio (both equation 3.53 and the obvious estimator) varies depending on only two factors: \(\tau\) itself, and the proportion of observations such that \(w_i = \tau\), which is simply equal to \(\Pr(y_i \geq x_i' \beta_\tau)\). It is interesting that the expectile variance estimate is influenced by which quantile the \(\tau^{th}\) quantile happens to approximate.

**Proposition 20.** Let \(\sigma_\tau^2 = E(\epsilon_i^2)\), \(\text{rank}(E(X'WX)) = k\), \(E(X'\epsilon) = 0\), and let \(\hat{\beta}_\tau\) be a consistent estimator of \(\beta_\tau\), \(\hat{\beta}_\tau \xrightarrow{p} \beta_\tau\). The estimator

\[
\sigma_{\tau,n}^2 = \frac{(y - X\hat{\beta}_\tau)'(y - X\hat{\beta}_\tau)}{n - 2k + \text{trace}(P_i'P_i)} \xrightarrow{p} \sigma_\tau^2 \] (3.54)

i.e., \(\sigma_{\tau,n}^2\) is a consistent estimator.
Proof. The proof is almost standard. Of course 
\[ \hat{\epsilon}_i = \epsilon_i - x'_i(\hat{\beta}_\tau - \beta), \]
so
\[ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 - 2 \left( \frac{1}{n} \sum_{i=1}^{n} \epsilon_i w_i x'_i \right) (\hat{\beta}_\tau - \beta) + (\hat{\beta}_\tau - \beta)' \left( \frac{1}{n} \sum_{i=1}^{n} w_i x_i x'_i \right) (\hat{\beta}_\tau - \beta) \]
\[ \xrightarrow{p} \sigma^2 \]

by the weak law of large numbers, the continuous mapping theorem, and the convergence in probability of \( \hat{\beta}_\tau \) to \( \beta_\tau \). Then
\[ \tilde{\sigma}^2_{\tau,n} = \frac{n}{n - 2k + \text{trace}(P'_\tau P_\tau)} \hat{\sigma}^2 \xrightarrow{p} \sigma^2 \]

by the continuous mapping theorem, together with the fact that \( \text{trace}(P'_\tau P_\tau) = O_p(1) \), which implies \( \frac{n}{n - 2k + \text{trace}(P'_\tau P_\tau)} \xrightarrow{p} 1 \). Notice that we have proven consistency for \( \tilde{\sigma}^2_{\tau,n} \) by first proving consistency of \( \hat{\sigma}^2 \), but \( \frac{n}{n-k} \xrightarrow{p} 1 \) also. Then \( s^2_\tau \) is also consistent. \( \square \)

The difference between the three consistent estimators will be small in data sets of any reasonable size, particularly in cases where \( \text{trace}(P'_\tau P_\tau) \) is close to \( n \) or \( k \). In Figure 3.10, we provide some simulation evidence that the sampling properties of \( \text{trace}(P'_\tau P_\tau) \) are not a major nuisance. For five small-to-medium sample sizes (\( n = 10, 25, 250, \) or 1000) with a fixed \( \Pr(y_i \geq x'_i \hat{\beta}_\tau) \), the 70% confidence intervals of the \( \text{trace}(P'_\tau P_\tau) \) are shown. The simulated data \( X \) are \( n \times 5 \) and uniformly distributed. These sampling distributions converge rapidly towards the limiting value as \( n \) increases, but they do not vary substantially when \( n \) is small. The difference between the sample median at \( n = 10 \) is never more than 3, which is less than \( k = 5 \) in this case. Even so, the difference suggests that there is value in using the calculated \( \text{trace}(P'_\tau P_\tau) \) for the degrees of freedom adjustment in small samples. Note that only \( F(\mu_\tau) \geq .5 \) are reported because the problem is symmetric: \( \tau = .2, F = .1 \) is the same as \( \tau = .8, F = .9 \), which is shown.
Figure 3.10: The median and 70% confidence intervals for \( \text{trace}(P'P) \) for \( k = 5 \) and \( n = 10, 25, 250, \) or 1000. \([X]_{ij}\) is distributed uniformly and the distribution of \( y \) is not specified, but \( \Pr(y_i \geq x_i'\beta_T) = F(\mu_T) \) is taken as fixed.
3.6.3 Asymmetric Conditional MSE

It may be useful to estimate the conditional mean squared error $E(\epsilon_i^2|\epsilon_i \geq 0)$ for two reasons. First, the use of expectile predictors for $\tau \neq .5$ eliminates any possibility that the distribution of errors is symmetric for nondegenerate cases. It is trivial to show that, if the distribution were symmetric, the mean would be zero—which we know to be false. This makes symmetric confidence intervals (of the usual $\mu \pm 1.96\sigma$ form, for instance) problematic\(^{15}\). Second, it may be useful to compare the estimated ratio of conditional mean squared errors to the assumed ratio $\frac{1-\tau}{\tau}$ as a test of assumption 4. Others may wish to explore the development of such a test.

As such, it is desirable to estimate the conditional mean squared error of the residual for the two cases where it is positive or negative. Denote the number of $\epsilon_i \geq 0$ as $n_1$ and the number of $\epsilon_i < 0$ as $n_2$ such that $n_1 + n_2 = n$. Obvious estimators are

\begin{align*}
\hat{\sigma}_{\tau}^2|\epsilon_i \geq 0 &= \frac{1}{n_2} \sum_{i=1}^{n} \hat{\epsilon}_i^2 I(\hat{\epsilon}_i \geq 0) \quad (3.55) \\
\hat{\sigma}_{\tau}^2|\epsilon_i < 0 &= \frac{1}{n_1} \sum_{i=1}^{n} \hat{\epsilon}_i^2 I(\hat{\epsilon}_i < 0) \quad (3.56)
\end{align*}

but it is less obvious how to partition the “degrees of freedom” penalty to create an unbiased estimator. This is important, because either $n_1$ or $n_2$ may be very small for extreme quantiles ($\tau$ close to 0 or 1). A general result is not obvious, but for the i.i.d. case where $I(\epsilon_i \geq 0)$ is independent of $\epsilon_i^2$, we would have

\begin{align*}
E(\sum_{i=1}^{n} \hat{\epsilon}_i^2 I(\hat{\epsilon}_i \geq 0)|X, W) &= nE(\hat{\epsilon}_i^2 I(\hat{\epsilon}_i \geq 0)|X, W) \\
&= E(\hat{\epsilon}_i^2|\hat{\epsilon}_i \geq 0) \times Pr(\hat{\epsilon}_i \geq 0) \quad (3.57)
\end{align*}

The expectation in the last line can be estimated by its sample moment and the

\(^{15}\)Under correct model assumptions (expectile assumption 4), we have $\frac{1-\tau}{\tau}$ times the variance for positive errors relative to negative errors, so it is possible to solve for the assumed relative variances as a function of $\sigma^2$. However, we prefer to consider the possibility that this assumption may be violated.
probability can be estimated using the empirical CDF, so

\[ \sigma^2_\epsilon \mid \epsilon_i \geq 0 := \frac{1}{n - 2k + \text{trace}(P^*_\tau P^*_\tau)} \left( \sum_{i=1}^{n} \hat{\epsilon}_i^2 I(\epsilon_i \geq 0) \right) \times \frac{n}{n^2} \quad (3.58) \]

is a reasonable estimator for equation 3.55 with a degrees of freedom penalty. Clearly this estimator\(^{16}\) is consistent in the same way as equation 3.54 and a similar estimator can be made for the mean squared error of negative error terms.

### 3.6.4 Expectile Adjusted \( R^2 \)

The \( R^2 \) statistic for expectile regression (expectile \( R^2 \)) can be found in only a few places in the literature. Our results from the previous two sections cast some light on the construction of that statistic, so we provide a comment.

The \( R^2 \) statistic is widely used as a measure of goodness of fit for regression lines, but has the undesirable property that it improves even when irrelevant regressors are added to the model specification. This fact has a long history in the literature for mean regression. For expectiles, \( R^2 \) appears to have been introduced by Aragon et al. [2005]. The \( R^2 \) statistic for generalized \( m \)-quantiles was introduced only very recently [Bianchi et al., 2018]. Depending on its construction, a statistic of this type may also suffer from the overfitting problem which is so well studied in the special case where \( \tau = .5 \); so the degrees of freedom penalty is important. See the famous paper by Cramer [1987] for a discussion of the bias of \( R^2 \) or see chapter 8 of the famous text by Maddala [1977] for a theoretical overview. The un-adjusted \( R^2 \) statistic for non-central estimators is similar to the usual\(^{17} \) weighted least squares \( R^2 \), such as the version given by Kvålseth [1985] or the suggested variation (pseudo-\( R^2 \)) by

\(^{16}\) In the expectile GLS case where additional weights \( \{\omega_i\}_{i=1}^{n} \) are desired, replace \( \sum_{i=1}^{n} \hat{\epsilon}_i^2 I(\epsilon_i \geq 0) \) with

\[ \left( \sum_{i=1}^{n} \omega_i I(\epsilon_i \geq 0) \right)^{-1} \sum_{i=1}^{n} \omega_i \hat{\epsilon}_i^2 I(\epsilon_i \geq 0) \]

as in the usual WLS variance estimator.

\(^{17}\) The pseudo R-squared statistic for weighted regression may be used for expectile regression:

\[ R^2_{ER} = 1 - \frac{WSSE}{WSS} \quad (3.59) \]

but suffers from the usual criticisms as far as model selection is concerned. For additional discussion regarding model selection in expectile regression models, see Spiegel et al. [2017] or Zhao and Zhang [2018].

118
Willett and Singer [1988]. The generalized version proposed by Anderson-Sprecher [1994] can also be used. As with generalized least squares models, there are many options.

The adjusted $R^2$ statistic, denoted $\bar{R}^2$, was proposed by Theil [1961] and has become a standard tool for mean regression. The adjusted $R^2$ for expectiles proposed by Aragon et al. [2005] is

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^{n} \varsigma_r(y_i - x_i' \hat{\beta}_r)/(n - \nu)}{\sum_{i=1}^{n} \varsigma_r(y_i - \bar{y}_r)/(n - 1)}$$

(3.60)

where the numerator on the right incorporates the weighted sum of squared errors and the denominator incorporates the weighted total sum of squares. These can also be written as

$$WSSE = \sum_{i=1}^{n} \varsigma_r(y_i - x_i' \hat{\beta}_r) = \sum_{i=1}^{n} \hat{w}_i(y_i - \hat{y}_i)^2$$

$$WSST = \sum_{i=1}^{n} \varsigma_r(y_i - \bar{y}_r) = \sum_{i=1}^{n} \hat{w}_i(y_i - \hat{y}_r)^2.$$  

(3.61)

Note that the estimated weights in the $WSSE$ and $WSST$ need not be the same; see footnote 18. Here, $\hat{w}_i$ is equal to $\tau$ if $y_i \geq \bar{y}_r$ and $1 - \tau$ otherwise. A degrees-of-freedom adjustment is already in place in equation 3.60, where the numerator is divided by $n - \nu$ and the denominator by $n - 1$. The choice of $\nu$ is not obvious. In the mean regression case, the purpose of this degrees-of-freedom adjustment—as stated by Thiel in the original example—is

Notice that the formulation given by Willett

$$= 1 - \frac{(y - X \hat{\beta}_e)'W(y - X \hat{\beta}_e)}{y'Wy - n\bar{y}_e^2}$$

extends to expectiles only if the weights we prefer for the denominator are the weights that produce $\bar{y}_e$; we may prefer to keep the expression in the form

$$= 1 - \frac{(y - X \hat{\beta}_e)'W(y - X \hat{\beta}_e)}{(y - \bar{y}_e 1_n)'W(y - \bar{y}_e 1_n)}$$

18 A difficulty occurs where $WSST = \sum_{i=1}^{n} \hat{w}_i(y_i - \hat{y}_e)^2$. Anderson and Sprecher suggest comparing the residual sum of squares from the full model with a constant-only model;

$$R^2 = 1 - \frac{RSS(\text{Full})}{RSS(\text{Reduced})}.$$

For regression expectiles, the two models may not produce the same weights. This corresponds to the formulation in equations 3.60 and 3.61, where the function $\varsigma_r$ is included explicitly. Willet’s pseudo-$R^2$ is the same only if the weights from the two models are the same; such as in the well-specified binary response case. They will not be the same in general.
to create an unbiased estimator of the residual variance in the numerator and an unbiased estimator of the variance of $y_i$ in the denominator. That interpretation is obfuscated by the weights in 3.60, but the role of the penalty against overfitting is still clear. Importantly, this $\bar{R}^2$ does nest Thiel’s adjusted $R^2$ when $\tau = .5$.

In Appendix C.7, it is shown that the expected value of $E(\epsilon' W \epsilon | X, W)$ is

$$E(\epsilon' W \epsilon | X, W) = (\text{trace}(W) - \text{trace}(WP_\tau))\sigma_\tau^2,$$

and the optimal degrees of freedom correction can be obtained from this. When we follow Thiel’s argument in favor of the revised statistic, then an unbiased estimator should be used for the estimated variances of $\epsilon_i$ and $y_i$, respectively. With that purpose in mind, we would suggest further modifying as

$$\tilde{R}^2 = 1 - \frac{\sum_{i=1}^{n} \varsigma_\tau (y_i - X_i' \tilde{\beta}_\tau) / (\text{trace}(\tilde{W}) - \text{trace}(\tilde{W}P_\tau))}{\sum_{i=1}^{n} \varsigma_\tau (y_i - \tilde{y}_\tau) / (\text{trace}(W) - \text{trace}(WP_1^1))}. \quad (3.62)$$

The expected values of $WSST$ are also made clear in Appendix C.7. Thus, this may be the ideal formulation of Thiel’s adjusted $R^2$ for expectiles. The more glaring issue is whether the weights in the numerator should be the same as the denominator, as discussed in footnote 18. The meaning of the statistic would vary depending on that choice.

As a measure of goodness of fit under an asymmetric loss function, the formulation in 3.62 uses that loss function $\varsigma_\tau$ and is entirely appropriate. The other possibility is based on pseudo-$R^2$ statistics, such as from Willett and Singer [1988]. The fundamental question is whether we prefer a measure of goodness of fit per the loss function $\varsigma_\tau$ or whether we are trying to replicate the adjusted $R^2$ statistic for a latent model, such as the model produced by the distribution $\tilde{F}$ rather than $F$. In the latter case,

$$\tilde{R}^2 = 1 - \frac{\sum_{i=1}^{n} \hat{w}_i (y_i - X_i' \hat{\beta}_\tau) / (\text{trace}(\hat{W}) - \text{trace}(\hat{W}P_\tau))}{\sum_{i=1}^{n} \hat{w}_i (y_i - \hat{y}_\tau) / (\text{trace}(\hat{W}) - \text{trace}(\hat{W}P_1^1))}$$

would seem to be the preferable statistic: the weights in both the numerator and denominator should be taken to be the best available estimation of the weights corresponding
to $\hat{F}$, which comes from the full model. The weights in $P_\tau$ and $P_1^\tau$ would also be the estimated weights from the full model. This reduces to the pseudo-$R^2$ of Willett and Singer [1988] when the degrees of freedom adjustment is omitted and has the same interpretation as the generalized $R^2$ of Anderson-Sprecher [1994] when the estimated weights $\hat{w}_i$ as in equation C.22 are taken seriously. This is the same as taking the interpretation in section 3.5.1 seriously—then our belief is that the modeled individuals are more (or less) likely than average to be above the regression line.

### 3.7 Conclusions

Expectile regression can reveal information that mean regression and quantile regression will never find. Under modified assumptions, the Gauss-Markov theorem and the GLS theorem each extend to expectile regression. Sample expectiles can be the best linear unbiased estimator of the true expectile, and the true expectile regression line is the mean regression line for an individual $\frac{\tau}{1-\tau}$ times as likely as average to be above the regression line. Interestingly, the first Gauss-Markov assumption (linearity) and third assumption (full rank) from the Gauss-Markov framework require no modification. For the location-scale model where generalized quantiles attained by different loss functions produce the same sets of regression lines, the $\tau^{th}$ expectile GLS estimator is also the BLUE for whichever regression line (under any loss function) corresponds to the $\tau^{th}$ expectile—thus, expectile regression coefficients are also the BLUE for quantile and $m$-quantile regression coefficients in special cases.

We would direct the reader’s attention to the binary response environment and our expectile linear probability model. For linear probability models, expectile regression reveals information that neither mean regression nor quantile regression can reveal. Furthermore, the interpretation of the $\tau^{th}$ expectile could not be simpler—it represents the mean regression for an individual who was $\frac{\tau}{1-\tau}$ times as likely as average to be above the regression line, ceteris paribus. Our application in Section 3.5.4 shows that individuals who were relatively likely to be denied a mortgage application have larger racial disparities. The racial difference
nearly doubles for individuals who are 5 times more likely than average to be denied. This is true in the simple regression, and in a multivariate expectile regression, and when GLS weights were used. A similar pattern is found for some of the other covariates—not only for the racial variable.

This work sheds light on some new ideas, but leaves many exciting possibilities unexplored. Expectile regression has become popular only in the last few years, and a surprising number of simple results (such as ours) have not been studied seriously. For example, standard models for data with serial correlation or other dependency structure are relatively unexplored in the expectile regression context. The binary response example that we employ is another such underserved subject: standard Logit and Probit models, or other binary response models, have not been adapted to expectile regression at all. We hope that these contributions yet to be made are conspicuous by their absence, and we invite others to consider them vigorously.

Expectile regression is a useful tool for the social sciences, and expectiles fit directly into the classical framework. Most researchers will find these estimators relatively easy to use and tremendously useful. We can only lament the fact that they did not become popular sooner. Expectiles can be used anywhere that a mean regression, such as OLS, is used. For linear probability models where quantile regression is not available, these estimators should be the new default option.
Chapter 4

The MLE of Aigner, Amemiya, and Poirier is not the Expectile MLE

4.1 Introduction

Expectiles are a continuum of summary statistics that range from the minimum to the maximum of a given distribution and can be used to characterize the total shape of that distribution. For example, the standard arithmetic mean is one special expectile. Since Kneib [2013], expectiles and expectile regression have gained popularity in the statistical and econometric literature. And since Ziegel [2016], expectile-based risk measures such as Expected Shortfall have become standard in quantitative finance. The rapid growth of this specialized topic has led to confusion, which can be separated along two dimensions. First, there is disagreement regarding which paper was the first to introduce expectiles. Second, there is debate regarding whether expectiles can be estimated by maximum likelihood and which likelihoods can be used for that purpose. This article will help to resolve those issues.

Currently in the literature, the origin of expectile regression is attributed alternately to Newey and Powell [1987] (NP) or to Aigner et al. [1976] (AAP). The asymmetrically weighted least-squares estimation that elicits expectiles is originally studied (briefly) in the paper by Aigner, Amemiya, and Poirier [1976], but the concept of expectiles belongs to Newey and Powell [1987]. Prior to that publication, the word “expectile” did not exist. In addition, the context in these two papers differs dramatically. The likelihood estimator proposed by AAP estimates a single parameter—not a continuum of parameters as with expectile regression. There is little—if any—discussion in the literature of the AAP model in

---

1This chapter was published under the same title at Econometric Reviews, Volume 40, No. 2. See Philipp (2021b). The copyright holder has granted permission for the work to be republished in this dissertation.
In their paper, Aigner et al. [1976] produce an asymmetrically weighted Gaussian likelihood model and discuss its maximum likelihood estimator (MLE). Their MLE does not produce an asymmetrically weighted least squares estimator—at least not an obvious one. But those authors did not intend to develop expectiles. On the contrary, they briefly discuss asymmetrically weighted least squares and demonstrate that what we now know as the elicitable scoring function for expectiles is an inconsistent estimator of their location parameter. This is in spite of the fact that their location parameter is an expectile—vacuously—as every quantile of a continuous distribution with a finite first moment is also an expectile; see Waltrup et al. [2015] or Yao and Tong [1996].

In contrast, Newey and Powell [1987] introduce expectiles as a continuum of semi-parametric statistics that can be interpreted as weighted averages. These expectiles are estimated using asymmetrically weighted least squares (expectile regression) and fit into the broader family of generalized quantiles or m-quantiles [Breckling and Chambers, 1988]. More recently, expectiles have gained interest for their explanatory power [Waltrup et al., 2015], as an alternative method of estimating quantiles [Daouia et al., 2019], and for their usefulness as a risk measure [Ziegel, 2016, Bellini and Di Bernardino, 2017]. Likelihood models that can elicit these useful statistics have emerged only in the last few years [Waldmann et al., 2018], and the model of Aigner et al. [1976] has not been studied in this context at all. This article sheds some light on that subject. A number of recent publications have claimed that expectile regression is obtained using the MLE of Aigner et al. [1976], which we discuss and show to be false.

The asymmetrically weighted Gaussian likelihood estimation found in Aigner et al. [1976] was an attempt to estimate a production possibilities frontier or other frontier, rather than any measure of central tendency. Their model is very natural; it partially preserves the quadratic objective function that has been so widely employed but extends it away from the mean. Those authors also argue in favor of their new model as a way to explore asymmetry in the loss function for a predictor, though they do not focus on this application. Either
way, we have a concept that shares certain features with expectile (or quantile) regression. The overall similarities are notable, (1) they do not study expectiles and (2) their model is not the correct likelihood model to estimate expectiles. We will demonstrate this in the following sections.

Why do these misconceptions persist? Apparently, there has been no effort in the literature to compare the AAP model with the asymmetric normal distribution of Kato et al. [2002] or its reparametrization used by Waldmann et al. [2017], which does elicit expectiles. Under scrutiny, the Quasi-maximum-likelihood estimation (QMLE) for expectiles used by Waldmann et al. [2018, 2017] shares interesting similarities and differences with the likelihood model and MLE of Aigner et al. [1976]. Both models use asymmetrically weighted Gaussian distributions, and both can be constructed by combining two half-normal densities. However, the models’ weights and the shape of their asymmetries are not the same. Furthermore, only the asymmetric normal density discussed by Waldmann et al. [2018] elicits the expectile in a quasi-likelihood setting. The only case where the two densities are the same is the one special case that they share: both models nest the usual Gaussian distribution. Because neither of the likelihood models discussed here are well known, we will provide a summary and some commentary for each of them. In juxtaposition, the differences become clear.

The AAP estimator does not seem to be useful for estimating expectiles. But, ironically, expectile regression may be useful to estimate the AAP location parameter if the AAP model is true. Unlike the maximum likelihood estimator, the expectile regression has standard asymptotics—specifically, it achieves asymptotic normality. Recently, it has become popular to explore Least-Squares-based estimators of the median and other quantiles [Waltrup et al., 2015, Daouia et al., 2019], but we do not find the AAP estimator to be appealing for that purpose.

In the following subsection, we briefly state the nature of the expectile and the context for the remaining discussion. Section 2 describes the expectile likelihood and its (Q)MLE. Section 3 describes the likelihood model proposed by Aigner et al. [1976] and its corresponding estimator. Comparisons are given in Section 4 and conclusions are given in Section 5.
4.1.1 Preliminaries

Let \( \{y_i \in \mathbb{R}, x_i \in \mathbb{R}^k\}_{i=1}^n \) be a set of i.i.d. draws of the random variables \( Y \) and \( X \) from some data-generating process whose joint distribution has support on \( \mathbb{R}^{k+1} \). Denote the unconditional marginal distribution of \( y \) as \( F \). For \( \tau \in (0,1) \), the \( \tau^{th} \) expectile of the distribution of \( Y \) is denoted by \( \mu_\tau \) and is defined as

\[
\mu_\tau := \arg \min_\mu \int \varsigma_\tau(y-\mu) dF(y)
\]

(4.1)

where the “swoosh” function \( \varsigma_\tau(\cdot) \) is given by

\[
\varsigma_\tau(u) = |\tau - I(u<0)|u^2.
\]

(4.2)

This concept was introduced and explored substantially in the article by Newey and Powell [1987]. One interpretation is available from the first-order condition, where

\[
\tau \int_{\mu_\tau}^\infty (y-\mu_\tau) dF(y) = -(1-\tau) \int_{-\infty}^{\mu_\tau} (y-\mu_\tau) dF(y).
\]

That is, the \( \tau^{th} \) expectile has \( \tau/(1-\tau) \) times the left-partial moment \( \int_{\mu_\tau}^\infty |y-\mu_\tau| dF(y) \) as its right-partial moment \( \int_{-\infty}^{\mu_\tau} |y-\mu_\tau| dF(y) \). The standard arithmetic mean is the \( \tau = .5 \) expectile, and has a well-known interpretation as the point of balance where the integral to the right and left are equal. Non-central expectiles for \( \tau \neq .5 \) are the points which do not balance, but have a particular ratio of tail integrals. Because they are indexed by \( \tau \in (0,1) \) and range from the minimum to the maximum of \( Y \), expectiles have frequently been compared to quantiles. In a broader mathematical context, both expectiles and quantiles are members of the family of \( m \)-quantiles [Breckling and Chambers, 1988], \( L_p \)-

\[\text{To enrich the physics analogy, the first moment is the point of balance of a lever where the torque from the right and left sides of the fulcrum are equal. The } \tau^{th} \text{ expectile is the point of imbalance where the torque from the left of the fulcrum is } \tau/(1-\tau) \text{ times the torque on the right.} \]

\[\text{For non-degenerate } F, \text{ Newey and Powell [1987, Theorem 1] prove that the expectile relation } \mu(\tau) : (0,1) \rightarrow \{y|0 < F(y) < 1\} \text{ is strictly monotone increasing in } \tau \text{ and injective. Add the convention that } \mu(0) := \inf(Y) \text{ and } \mu(1) := \sup(Y), \text{ and } \mu(\tau) \text{ is a surjective and injective mapping from the closed unit interval to } \text{S}_Y \in \mathbb{R}, \text{ the convex hull of the support of } Y. \text{ Thus, every value in the interval } [\inf(Y), \sup(Y)] \text{ is an expectile of } Y. \]

126
quantiles [Chen, 1996], or generalized quantiles [Daouia et al., 2019]. In the linear regression case where $\mu_\tau$ is replaced with $x_i'\beta_\tau$, true regression quantiles and expectiles may be the same set of fit lines: see Waltrup et al. [2015], Yao and Tong [1996], or Kneib [2013]. For distributions with a first absolute moment, every quantile is also an expectile though the converse is not necessarily true.

Several other interpretations and motivations for the expectile are available in more specific contexts. See Waldmann et al. [2018] for some discussion. For explanation of why expectiles are currently important in quantitative finance, see Ziegel [2016] or Bellini and Di Bernardino [2017].

### 4.2 Expectile Likelihood Models

In this section, we introduce the Asymmetric Normal Distribution that elicits the $\tau$th expectile as its QMLE. To help distinguish this from the model of Aigner et al. [1976], we denote the density as $f_\tau$ and the location parameter $\mu_\tau$ with the subscript $\tau$ corresponding to the $\tau$th expectile, as in the original article by Newey and Powell. This also helps to emphasize that the shape of the distribution varies depending on $\tau$, which is interpreted as a fixed hyperparameter. In the context of expectile regression, each $\tau$ is indexed to a different location parameter $\mu_\tau$ and several different $\tau$ may be used to estimate multiple expectiles simultaneously.

#### 4.2.1 The “Expectile” Distribution

The most obvious likelihood model for expectile regression can be found in the article by Waldmann et al. [2018]. Those authors suggest the Asymmetric Normal Distribution (AND) and provide its probability density function explicitly. Unfortunately, the model used by Aigner et al. can also be called an asymmetric normal distribution. Because the model and parametrization given below is specifically adapted to expectile regression,
we will avoid confusion by referring to the following probability model as the “expectile” distribution.

For a random variable $Y$ with location parameter (mode) $\mu_\tau$, scale parameter $\sigma^2_\tau$, and asymmetry parameter $\tau$, we have:

$$f_\tau(y|\mu_\tau, \sigma^2_\tau; \tau) = \frac{2\sqrt{\tau(1-\tau)}}{\sqrt{\pi}\sigma^2_\tau(\sqrt{\tau}-\sqrt{1-\tau})} \exp \left\{ - \frac{w(y)}{\sigma^2_\tau} (y - \mu_\tau)^2 \right\} \quad (4.3)$$

where the weight function $w(y)$ is

$$w(y) = \begin{cases} \tau & y \geq \mu_\tau \\ 1 - \tau & \text{if } y < \mu_\tau \end{cases}$$

In this case, we can say that the residual $\epsilon_i = y_i - \mu_\tau$ is distributed as

$$\epsilon_i \sim \begin{cases} HN(0, \sigma^2_\tau \tau) & \text{w.p. } \frac{\sqrt{\tau}}{\sqrt{\tau}+\sqrt{1-\tau}} \\ -HN(0, \sigma^2_\tau (1-\tau)) & \text{w.p. } \frac{\sqrt{1-\tau}}{\sqrt{\tau}+\sqrt{1-\tau}} \end{cases} \quad (4.4)$$

where $HN(\cdot)$ is the half-normal distribution with mode zero and scale parameter $\sigma^2_\tau \tau$ or $\sigma^2_\tau (1-\tau)$.Clearly, the probability of positive and negative errors from (4.3) are not equal. Rather, the CDF $F_\tau(x) = \int_{-\infty}^{x} f_\tau(x)dx$ has $F_\tau(\mu_\tau) \geq .5$ for $\tau \leq .5$.

An unfortunate attribute of this model parametrization (compared to the same model as written by Kato et al. [2002]) is the complexity of the multiplicative constant. The likelihood for (4.3) is

$$p(y|\mu_\tau, \sigma^2_\tau; \tau) = \prod_{i=1}^{n} \frac{2\sqrt{\tau(1-\tau)}}{\sqrt{\pi}\sigma^2_\tau(\sqrt{\tau}-\sqrt{1-\tau})} \left[ \exp\left\{ - \frac{\tau}{\sigma^2_\tau} (y_i - \mu_\tau)^2 \right\} I(y_i \geq \mu_\tau) \right. \times \left[ \exp\left\{ - \frac{1-\tau}{\sigma^2_\tau} (y_i - \mu_\tau)^2 \right\} I(y_i < 0) \right] \quad (4.5)$$
and the corresponding log-likelihood is

\[
l_{r}(\mu_{r}, \sigma_{r}^{2} ; \tau | y) = \sum_{i=1}^{n} \left( \frac{1}{2} \ln(4\tau(1 - \tau)) - \frac{1}{2} \ln \pi \sigma_{r}^{2} - \ln(\sqrt{\tau} - \sqrt{1 - \tau}) - \frac{\tau}{\sigma_{r}^{2}} \epsilon_{i}^{2} \right) I(\epsilon_{i} \geq 0) \\
+ \sum_{i=1}^{n} \left( \frac{1}{2} \ln(4\tau(1 - \tau)) - \frac{1}{2} \ln \pi \sigma_{r}^{2} - \ln(\sqrt{\tau} - \sqrt{1 - \tau}) - \frac{1 - \tau}{\sigma_{r}^{2}} \epsilon_{i}^{2} \right) I(\epsilon_{i} < 0).
\]

(4.6)

As messy as the constant terms may be, they are irrelevant to a maximization procedure. And clearly, each of the expressions above will simplify to a standard Gaussian model when \( \tau = .5 \). To show that this model’s MLE is identical to the semiparametric estimator of Newey and Powell [1987], the log likelihood can also be written as

\[
l_{r}(\mu_{r}, \sigma_{r}^{2} ; \tau | y) = n \ln \left( \frac{2\sqrt{\tau(1 - \tau)}}{\sqrt{\pi} \sigma_{r}^{2}(\sqrt{\tau} - \sqrt{1 - \tau})} \right) - \frac{1}{\sigma_{r}^{2}} \sum_{i=1}^{n} \rho_{r}(y_{i} - \mu_{r})
\]

(4.7)

where \( \rho_{r}(u) \) is the asymmetrically weighted least-squares scoring function of Newey and Powell [1987]. Gneiting [2011] provides conditions for the uniqueness of that scoring function for eliciting expectiles. Those conditions are satisfied here: namely, the support of \( Y \) is \( \mathbb{R} \) and our log likelihood is strictly-concave.

### 4.2.2 The Expectile MLE

The estimator obtained by maximizing the log-likelihood in (4.6) is quite simple. First derivatives for the problem are

\[
\frac{\partial l_{r}(\mu_{r}, \sigma_{r}^{2} ; \tau | y)}{\partial \mu_{r}} = \frac{2}{\sigma_{r}^{2}} \sum_{i=1}^{n} |\tau - I(y_{i} < \mu_{r})|(y_{i} - \mu_{r}) \\
\frac{\partial l_{r}(\mu_{r}, \sigma_{r}^{2} ; \tau | y)}{\partial \sigma_{r}^{2}} = -\frac{n}{\sigma_{r}^{2}} + \frac{1}{\sigma_{r}^{4}} \sum_{i=1}^{n} |\tau - I(y_{i} < \mu_{r})|(y_{i} - \mu_{r})^{2}.
\]

(4.8)
Setting both equal to zero and solving, we have

\[
\hat{\mu}_{\tau, MLE} = \frac{\sum_{i=1}^{n} |\tau - I(y_i < \mu_{\tau})|y_i}{\sum_{i=1}^{n} |\tau - I(y_i < \mu_{\tau})|} \quad (4.9)
\]

\[
\hat{\sigma}^2_{\tau, MLE} = \frac{2}{n} \sum_{i=1}^{n} |\tau - I(y_i < \mu_{\tau})|(y_i - \mu_{\tau})^2. \quad (4.10)
\]

Several observations are available from this result. The estimator for \(\mu_{\tau}\) can also be written as \((\sum_{i=1}^{n} w_i)^{-1} \sum_{i=1}^{n} w_i y_i\), a standard estimator for a weighted average. The estimator for the scale parameter \(\sigma^2_{\tau}\) can also be written as \(\frac{2}{n} \sum_{i=1}^{n} \rho_{\tau}(y_i - \mu_{\tau})\) and is clearly equal to the usual Gaussian variance MLE when \(\tau = .5\). To show that the solution for the location parameter is not reliant on \(\sigma^2_{\tau}\), we may also write the concentrated likelihood using \(\hat{\sigma}^2_{\tau, MLE}\) as

\[
\begin{align*}
ll^{C}_{\tau}(\mu_{\tau}; \tau | y) &= \frac{n}{2} \ln(4\tau(1-\tau)) - \frac{n}{2} \ln \pi - n \ln(\sqrt{\tau} - \sqrt{1-\tau}) \\
&\quad - \frac{n}{2} \ln \left(\frac{2}{n} \sum_{i=1}^{n} |\tau - I(y_i < \mu_{\tau})|(y_i - \mu_{\tau})^2\right) - \frac{n}{2}. \quad (4.11)
\end{align*}
\]

Then, the first order condition would give

\[
\frac{\partial ll^{C}_{\tau}(\mu_{\tau}; \tau | y)}{\partial \mu_{\tau}} = n \frac{\sum_{i=1}^{n} |\tau - I(y_i < \mu_{\tau})|(y_i - \mu_{\tau})}{\sum_{i=1}^{n} |\tau - I(y_i < \mu_{\tau})|(y_i - \mu_{\tau})^2} = 0.
\]

So long as the empirical distribution is not degenerate with \(n\) identical observations for \(y_i\), the first-order condition leaves the same solution for \(\hat{\mu}_{\tau, MLE}\) as previously given in equation 4.9. An easy calculation will also verify the second-order condition, using either the original likelihood or the concentrated likelihood. And, because \(\rho_{\tau}\) is strictly convex, it is clear that the log likelihood representation in (4.7) is strictly concave and thus has a globally unique maximum.

The same holds true in the regression model where we replace \(\mu_{\tau}\) with \(x'_{i}\beta_{\tau}\). Specifically,
the first order conditions leave

\[
\frac{\partial l(\beta, \sigma^2; \tau \mid y)}{\partial \beta} = \frac{2}{\sigma^2} \sum_{i=1}^{n} x_i |\tau - I(y_i < x_i' \beta)| (y_i - x_i' \beta) = 0 \]

\[
\frac{\partial l(\mu, \sigma^2; \tau \mid y)}{\partial \sigma^2} = -n \sigma^2 + \frac{1}{\sigma^4} \sum_{i=1}^{n} |\tau - I(y_i < x_i' \beta)| (y_i - x_i' \beta) = 0,
\]

so the solution for \( \hat{\beta}_\tau \) is

\[
\hat{\beta}_\tau = \left( \sum |\tau - I(y_i < x_i' \beta)| x_i x_i' \right)^{-1} \left( \sum |\tau - I(y_i < x_i' \beta)| x_i y_i \right).
\]

(4.12)

Then \( \hat{\beta}_\tau \) is identical to the estimator given in Newey and Powell [1987]. See that citation for proofs of consistency and asymptotic normality under relatively general conditions. Those conditions require the existence of slightly more than four moments for \( y \), which is trivial to show here. Simply remark that the \( r \text{th} \) “central” moment of this distribution around \( \mu \) is bounded above by the \( r \text{th} \) moment of \( N(0, \sigma^2 \tau^2) \) or by \( N(0, \sigma^2 (1-\tau)) \), whichever is larger. The normal distribution has an \( r \text{th} \) moment for all \( r \in \mathbb{N} \), thus, the \( r \text{th} \) moment of the expectile distribution exists and is finite for all \( \tau \in (0, 1) \). In the heteroscedastic case where \( \sigma^2 \) varies among observations, perhaps as a function of \( x_i \), consistency and asymptotic normality of the corresponding weighted estimator is available if each \( \sigma^2_{\tau,i} \) can be estimated consistently. See Barry et al. [2018] for that result. More general asymptotics for expectiles can be found in Holzmann and Klar [2016]. In the location model where data are i.i.d. with distribution \( F \), a first absolute moment is sufficient for consistency and two moments are sufficient for asymptotic normality.

\[\text{\footnote{The estimator in (4.12) has the standard linear form typical of a generalized least squares estimator, } \hat{\beta}_\tau = (X'WX)^{-1}X'Wy.}\]

The matrix \( W \) is a diagonal matrix of weights \([W]_{ii} = w_i\), with \( w_i = \tau \) if \( y_i \geq x_i' \beta \) and \( w_i = 1 - \tau \) otherwise.
4.3 The AAP Likelihood

Next, we present the likelihood model given by Aigner et al. [1976] with some comments. In the context of the original article, we presume that the likelihood model is the true data generating process and has true parameters $\beta_0$ (or $\mu_0$ in the location model) and $\sigma^2_0$. Thus, these parameters $\beta, \mu, \sigma^2$ are not functions of a pre-specified asymmetry parameter $\theta$ and do not receive a subscript $\theta$ in our notation. Several other differences between the models are obvious; these will be discussed in Section 4.4.

4.3.1 The AAP Distribution

Aigner et al. [1976] begin with a linear regression model

$$ y = X\beta + \epsilon $$

(4.13)

for the vector $y$ of $n$ observations of the dependent variable, $X$ an $n \times k$ matrix of corresponding covariates, $\beta$ the coefficient vector, and $\epsilon$ the vector of residuals. Let the error terms $\epsilon_i$ be distributed as

$$ \epsilon_i = w_i^{-1/2} N(0, \sigma^2) $$

(4.14)

where

$$ w_i = \begin{cases} 
1 - \theta & \text{if } \epsilon_i > 0 \\
\theta & \text{if } \epsilon_i \leq 0.
\end{cases} $$

(4.15)

This is conceptually close to the formulation of the expectile distribution in Section 4.2, in the sense that we could re-parameterize the inverse weights given here. The coefficient of asymmetry $\theta$ is also reversed in the sense that $\theta = 1 - \tau$ in the weights $w_i$ from the expectile
model. But the weights are effectively inverted as well, so AAP write

\[ \epsilon_i = \begin{cases} 
\frac{1}{\sqrt{1 - \theta}} \epsilon_i^* & \text{if } \epsilon_i^* > 0 \\
\frac{1}{\sqrt{\theta}} \epsilon_i^* & \text{if } \epsilon_i^* \leq 0 
\end{cases} \] (4.16)

with \( \epsilon_i^* \sim i.i.d. \ N(0, \sigma^2) \), which is convenient for comparison with (4.4) on page 128.\(^6\)

The density in this case is also easily represented casewise,

\[ f_\theta(\epsilon|\beta, \theta, \sigma^2) = \begin{cases} 
\frac{\sqrt{(1-\theta)}}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(1-\theta)\epsilon_i^2}{2\sigma^2}\right] & \text{if } \epsilon > 0 \\
\frac{\sqrt{\theta}}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\theta \epsilon_i^2}{2\sigma^2}\right] & \text{if } \epsilon \leq 0.
\] (4.17)

If we take the model seriously, we could simply ask how often \( \epsilon_i \) is positive versus negative. Then, because \( \epsilon_i \) is i.i.d. with median zero, we might estimate the linear coefficients \( \beta \) by least absolute deviations and we would not need to know \( \theta \) at all. This follows from the fact that the latent error process \( \epsilon_i^* \) has median zero and quantiles are invariant under monotone transformations. But, when choosing a likelihood model for purposes of estimation, it seems that the applicability of the model should be influenced by which sort of summary statistics the practitioner wishes to estimate. In this model, the median colocates with the mode but not with the mean.

Working towards the maximum likelihood estimator, we write the likelihood for the model as

\[ p(y|\beta, \theta, \sigma^2) = \prod_{i=1}^{n} \left[ \frac{\sqrt{(1-\theta)}}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(1-\theta)\epsilon_i^2}{2\sigma^2}\right] I(\epsilon_i > 0) \right] \times \left[ \frac{\sqrt{\theta}}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\theta \epsilon_i^2}{2\sigma^2}\right] I(\epsilon_i \leq 0) \right]. \] (4.17)

\(^6\)Alternately, we might make the comparison between equation 4.16 and equation 4.4 even clearer by writing

\[ \epsilon_i = \begin{cases} 
HN(0, \frac{\sigma^2}{1 - \theta}) & \text{w.p. } \frac{1}{2} \\
-HN(0, \frac{\sigma^2}{\theta}) & \text{w.p. } \frac{1}{2}.
\]
and the log-likelihood as

\[
l_{\theta}(\beta, \sigma^2 | y) = \sum_{i=1}^{n} \left( -\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 + \frac{1}{2} \ln (1 - \theta) - \frac{(1 - \theta)}{2\sigma^2} \epsilon_i^2 \right) I(\epsilon_i > 0) \\
+ \sum_{i=1}^{n} \left( -\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 + \frac{1}{2} \ln \theta - \frac{\theta}{2\sigma^2} \epsilon_i^2 \right) I(\epsilon_i \leq 0). \tag{4.18}
\]

This is made somewhat more complicated by the fact that there are three variable parameters, rather than only two. However, we can concentrate the likelihood to simplify our analysis. Finding the first-order conditions to maximize with respect to \(\sigma^2\), we have

\[
\frac{\partial l_{\theta}(\beta, \theta, \sigma^2 | y)}{\partial \sigma^2} = \sum_{i=1}^{n} \left( -\frac{1}{2\sigma^4} + \frac{(1 - \theta)}{2\sigma^4} \epsilon_i^2 I(\epsilon_i > 0) + \frac{\theta}{2\sigma^4} \epsilon_i^2 I(\epsilon_i \leq 0) \right) = 0
\]

\[\Rightarrow \hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (1 - \theta) \epsilon_i^2 I(\epsilon_i > 0) + \theta \epsilon_i^2 I(\epsilon_i \leq 0) \]

\[= \frac{1}{n} \sum_{i=1}^{n} w_i \epsilon_i^2 \tag{4.19}
\]

which is elegant and helps us concentrate or profile the log-likelihood as

\[
l_{\theta}(\beta, \theta | y) = \sum_{i=1}^{n} \left( \frac{1}{2} \ln (1 - \theta) I(\epsilon_i > 0) + \frac{1}{2} \ln \theta I(\epsilon_i \leq 0) \right) \\
- \frac{n}{2} \ln \frac{1}{n} \sum_{i=1}^{n} \left( (1 - \theta) \epsilon_i^2 I(\epsilon_i > 0) + \theta \epsilon_i^2 I(\epsilon_i \leq 0) \right) \tag{4.20}
\]

\[= \sum_{i=1}^{n} \frac{1}{2} \ln w_i - \frac{n}{2} \ln \frac{1}{n} \sum_{i=1}^{n} w_i \epsilon_i^2 \tag{4.21}
\]

This is equivalent to the representations given in equations (5) or (7) in the original article. It is also deceptively simple.

Compare the result in (4.20) with its counterpart in (4.11) on page 130. Both contain an expression similar to \(\frac{n}{2} \ln \frac{1}{n} \sum_{i=1}^{n} w_i \epsilon_i^2\), which suggests that we could estimate the maximum of the likelihood using asymmetric least squares if we take \(\theta\) to be known a priori. That would be consistent with the interpretation from the expectile distribution in the previous section. However, in this model, \(w_i\) are not constant—the maximum likelihood estimator \(\hat{\theta}_{MLE}\) varies with \(x_i' \beta\). Even when \(\theta\) is taken to be known a priori, \(w_i\) does not correspond
to the inverse weights given in (4.16) but to their inverse square. This causes some minor
difficulty if we attempt to estimate \( \mu_o \) by asymmetric least squares, as discussed in Section
4.4.

### 4.3.2 The AAP Estimator

Here, we briefly outline the proof of consistency of the MLE for this model. Importantly,
the limiting condition is nonstandard due to the discontinuity in the density. This precludes
the use of standard asymptotics to obtain the limiting distribution of the estimator. That
fact should serve as a sufficient warning not to confuse this model with the expectile model
in section 4.2 on page 127.

It is simplest to explore the consistency of the AAP estimator in the location model
given by

\[
y_i = \mu_o + \epsilon_i
\]

which is nested in the regression model when \( x_i = 1 \) and \( \beta = \mu_o \). We will denote
\( \hat{\mu}_{MLE} \), \( \hat{\sigma^2}_{MLE} \), and \( \hat{\theta}_{MLE} \) as the maximum likelihood estimators for the “true” parameters
\( \mu_o, \sigma_o^2, \) and \( \theta_o \). Aigner et al. [1976] warn us against trying to maximize the likelihood for a
sample by minimizing its additive inverse less the term \( \sum_{i=1}^{n} \frac{1}{2} \ln w_i - \)

\[
\hat{\mu} = \arg \min_{\mu \in \mathbb{R}} Q_n(\mu) = \arg \min_{\mu} \ln \frac{1}{n} \sum_{i=1}^{n} w_i (y_i - \mu)^2
\]

–because each weight \( w_i \) is a function \( w_i = w(y_i, \mu, \theta) \) of the other parameters. This estimator \( \hat{\mu} \) has the standard form of asymmetric least squares, but it is inappropriate in
the AAP model because the asymmetric weights \( \theta \) and \( 1 - \theta \) are not constant with respect
to \( \mu \) almost everywhere. In particular, the value of \( \theta \) that maximizes the likelihood will vary
with \( \mu \). Accordingly, those authors suggest first estimating \( \theta_o \) by maximum likelihood,
substituting into the previously concentrated likelihood, and further concentrating the
likelihood. That procedure will leave

\[
\ell_{\theta}(\mu|y)_{\theta=\hat{\theta}_{MLE}} = -\frac{n_1}{2} \log \frac{\sum(y_i - \mu)^2 I(y_i \leq \mu)}{n_1} - \frac{n_2}{2} \log \frac{\sum(y_i - \mu)^2 I(y_i > \mu)}{n_2}
\]
where $n_1 = \sum_{i=1}^{n} I(y_i \leq \mu)$ and $n_2 = n - n_1$. Following that step, we can show consistency of $\hat{\mu}_{MLE}$ by showing that the minimum of $Q_n$ converges as $n \to \infty$ to the minimum of the “true” objective function. Then

$$
p \lim_{n \to \infty} Q_n = F_\theta(\mu) \ln E \left( (y_i - \mu)^2 | y_i \leq \mu \right) + \left(1 - F_\theta(\mu) \right) \ln E \left( (y_i - \mu)^2 | y_i > \mu \right).
$$

(4.25)

In the limit, the expression above should have $F_\theta(\hat{\mu}_{MLE}) = .5$ if $F_\theta$ as given in (4.16) is the true distribution. But in a finite sample, we see that the c.d.f. $F_\theta$ must be evaluated piecewise for any value $\mu > \mu_o$. After several steps and some considerable algebra [Aigner et al., 1976, p. 391], those authors successfully show that

$$
\lim_{\mu \to \mu_o-} \left[ \frac{\partial}{\partial \mu} p \lim_{n \to \infty} Q_n \right] = -\frac{2\sqrt{\theta}}{\sqrt{2\pi \sigma_o^2}} \left[ A - \log A - 1 \right]
$$

$$
\lim_{\mu \to \mu_o+} \left[ \frac{\partial}{\partial \mu} p \lim_{n \to \infty} Q_n \right] = \frac{2\sqrt{1 - \theta}}{\sqrt{2\pi \sigma_o^2}} \left[ B - \log B - 1 \right]
$$

where $A = \sqrt{(1 - \theta)/\theta}$ and $B = \sqrt{\theta/(1 - \theta)}$. The function $x - \log x - 1$ is nonnegative for positive $x$, so we see that the left-derivative of the function is nonpositive while the right derivative is nonnegative. Then $p \lim_{n \to \infty} Q_n(\mu)$ does achieve a minimum at $\mu_o$, despite the lack of a first-order condition.

All of the extra effort–this nonstandard proof of consistency–is necessary because of the discontinuity of the p.d.f. at the location parameter. A more general approach to consistency and asymptotic normality–relying on stochastic equicontinuity and differentiability in mean square–is not applicable. While Newey [1991] points out that convergence in probability is a sufficient condition for consistency, the density of this model fails to be differentiable in quadratic mean at $\mu_o$, preventing asymptotic normality.\footnote{For $\theta \neq .5$ and arbitrary $\sigma^2$, a simple calculation will show that

$$
n \int \left( f_\theta(y|\mu = \mu_o)^{1/2} - f_\theta(y|\mu = \mu_o + h/\sqrt{n})^{1/2} \right)^2 dy \to \infty
$$

which violates the result in Lemma 12.2.1 in Lehmann and Romano [2006]. See also van der Vaart [2000, p. 64]. This demonstrates that the AAP distribution is not differentiable in quadratic mean.} To make matters worse,
Hirano and Porter [2012] show that it may be impossible to find locally asymptotically normal estimators for discontinuous functionals under fairly general conditions. Indeed, the limiting distribution of the AAP estimator has proven to be elusive until recently.

Fortunately, the estimation of parameters at discontinuities in density is a known problem, and some recent developments are directly applicable to the AAP estimator. Borovkov [2018] uses empirical process arguments based on the change-point problem to study the behaviors of a new class of estimators for location parameters at discontinuities in the density. Importantly, he finds that one of them is asymptotically equivalent to the MLE and shows that its limiting form is asymmetric with exponential tails. Following Borovkov’s result, we conclude that the AAP MLE will not be asymptotically normal and thus cannot be used with standard testing.

Moving on to related topics, Aigner et al. [1976] also consider the possibility that their likelihood could be used with a known or predetermined asymmetry coefficient, $\theta$. Because the true distribution $F_\theta(y)$ of $y_i$’s has Lebesgue measure, we can see that the weights $w_i$ evaluated using any $\mu$ would then be constant almost surely (constant with respect to $\mu$ almost everywhere). Then clearly, the estimator in (4.23) is also the estimator given by Newey and Powell to estimate the $\tau = (1-\theta)^{th}$ expectile. But this “oracle” estimator, with $\theta$ known, is still an inconsistent estimator of the AAP location parameter. In the following section, we will explain why this is the case.

### 4.4 Comparisons

Both the AAP model and the expectile distribution nest the Gaussian distribution when $\tau$ or $\theta$ are equal to .5. Both distributions are somewhat more complex than the typical symmetric Gaussian distribution in terms of their parametric form. Between the two, the question of which is simpler is subjective, though this author would suggest that Aigner’s distribution is easier to propose a priori by writing (4.14), while the expectile distribution has a parametric form that is easier to work with, especially as far as estimators are concerned.

---

8Hirano and Porter [2012] also require differentiability in quadratic mean, which is not applicable to our discontinuous density problem.
As mentioned briefly in the previous section, we would point out the similarity between the two distributions by drawing attention to equations (4.4) and (4.16). If not for the square root operator in the denominator of (4.16) and the reversal of the unit interval (τ corresponds roughly to 1 − θ and vice versa), the two would have identical parametrizations for the positive and negative conditional distributions of errors. The second (equally significant) difference arises in the choice of how often positive and negative errors occur. In the Aigner et al. case, positive and negative errors occur with equal probability such that the location parameter is at the median. For the expectile distribution, they do not.

Aside from the special case where τ = θ = .5, the two distributions are never equivalent. They produce radically different estimators, with standard asymptotic results available only for the expectile distribution’s MLE. As far as the conceptual presentation of each model is concerned, their interpretations also have limited commonality.

4.4.1 The AAP Location Parameter is not the Expectile

The distribution from AAP has some unusual characteristics, but can it be used to estimate expectiles? We see the following. While discussing alternative estimators for their model, AAP point out that the semi-parametric “oracle” estimator given by

$$
\hat{\mu} = \arg \min_{\mu} \sum_{i=1}^{n} (\theta_0(y_i - \mu)^2 I(y_i \leq \mu) + (1 - \theta_0)(y_i - \mu)^2 I(y_i > \mu))
$$

(4.26)

$$
= \arg \min_{\mu} \sum_{i=1}^{n} w_i(y_i - \mu)^2
$$

is not a consistent estimator of their \( \mu_0 \). This represents an early appearance of the expectile scoring function (asymmetric least squares) in the literature, but it also violates any supposition that \( \mu_0 \) is the \( \tau = (1 - \theta_0)^{th} \) expectile or that \( \hat{\mu}_{MLE} \) elicits the \( (1 - \tau)^{th} \) expectile. Instead, they show that the estimator

$$
\hat{\mu} = \arg \min_{\mu} \sum_{i=1}^{n} \left( \sqrt{\theta_0}(y_i - \mu)^2 I(y_i \leq \mu) + \sqrt{1 - \theta_0}(y_i - \mu)^2 I(y_i > \mu) \right)
$$

is a consistent estimator of their location parameter. This will be equivalent to the elicitable
score function for the $\tau^{th}$ expectile where

$$\tau = \frac{\sqrt{1 - \theta_o}}{\sqrt{1 - \theta_o} + \sqrt{\theta_o}},$$

(4.27)

which is clearly not the same. That is a nice result, but Yao and Tong [1996] have proven that any location parameter can be written as the $\tau^{th}$ expectile for some $\tau$, so the existence of such a $\tau$ should be no surprise in retrospect (see Footnote 3). Importantly, use of the relation in (4.27) hinges on the assumptions that $\theta_o$ is known and the AAP model is the true data generating process, which are both nontrivial conditions.

Following Yao and Tong [1996], we may easily prove that the mode of the distribution given by (4.16) is not the $(1 - \theta_o)^{th}$ expectile. The $\tau^{th}$ expectile $\mu_{\tau}$ of any distribution satisfies

$$\tau E(|y - \mu_{\tau}| I(y > \mu_{\tau})) = (1 - \tau)E(|y - \mu_{\tau}| I(y \leq \mu_{\tau})).$$

(4.28)

For the AAP density, we may evaluate the expectations on the left and right of the location parameter in order to which expectile inhabits that point. The truncated normal distribution (for $\epsilon_i$) has conditional moments

$$E(y|y > \mu_o) = \mu_o + \frac{2\sigma_o}{\sqrt{2\pi}\sqrt{1 - \theta_o}}$$

$$E(y|y \leq \mu_o) = \mu_o - \frac{2\sigma_o}{\sqrt{2\pi}\sqrt{\theta_o}}.$$

Thus, we can see directly that (4.28) requires

$$\tau \left(\frac{2\sigma_o}{\sqrt{2\pi}\sqrt{1 - \theta_o}}\right) = (1 - \tau) \left(\frac{2\sigma_o}{\sqrt{2\pi}\sqrt{\theta_o}}\right)$$

or

$$\tau \sqrt{\theta_o} = (1 - \tau) \sqrt{1 - \theta_o}.$$

This leaves the solution for $\tau$ given in (4.27). But this result is specific to the AAP model and will not hold in general. We also need not employ any of these facts to estimate
expectiles, which we can do quite simply by other methods.\(^9\) In general, our knowledge of the density plays no role in estimation of the sample expectile [Holzmann and Klar, 2016].

For the AAP model, an interesting possibility would be to use the \(\tau\)^th sample expectile with \(\tau\) as in equation 4.27 as an estimator of \(\mu_o\) (with \(\hat{\theta}_{\text{MLE}}\) as an estimator of \(\theta_o\)). This estimator is known to be consistent and asymptotically normal if the distribution has at least two moments and its CDF is continuous at the \(\tau\)^th expectile.\(^10\) Both of these conditions are satisfied in the AAP case. This implies a tradeoff—the AAP MLE is more efficient with \(O_p(n^{-2})\) variance, but it is not asymptotically normal. The expectile estimator based on equation 4.27 is asymptotically normal, but not as efficient with \(O_p(n^{-1})\) variance. If we are willing to forgo the efficiency of the MLE, we can utilize standard testing and vice-versa.\(^11\)

Going one step further, we can compare the AAP location parameter to expectiles in the quasi-likelihood environment where the AAP distribution is not assumed to be the true distribution \(F\). To compare with equation 4.1, the AAP location parameter \(\mu_o\) (conditional on \(\theta\) and \(\sigma^2\)) can be defined as the value that maximizes the expected (quasi) likelihood function for a single observation.

\[
\mu_o|\theta, \sigma^2 := \arg \max_{\mu} \int l_\theta(\mu|\theta, \sigma^2; y) dF(y)
\]
\[
= \arg \max_{\mu} \int \left( \ln \frac{\sqrt{(1 - \theta)}}{\sqrt{2\pi\sigma^2}} - \frac{(1 - \theta)(y - \mu)^2}{2\sigma^2} \right) I(y > \mu)
\]
\[
+ \left[ \ln \frac{\sqrt{\theta}}{\sqrt{2\pi\sigma^2}} - \frac{\theta(y - \mu)^2}{2\sigma^2} \right] I(y \leq \mu) dF(y)
\]

This functional representation of the parameter \(\mu_o\) makes it clear that the indicator function

\(^9\)As an aside, it is also worth noting that the difference between the two distributions persists in the "external" case of \(\tau = 0\) or \(\tau = 1\), where we have

\[
\lim_{\theta \to 0} F_\theta(y|\mu, \theta, \sigma^2) = \frac{1}{2} \mathcal{U}(-\infty, \mu) + \frac{1}{2} \delta(\mu)
\]
\[
\lim_{\tau \to 0} F_\tau(y|\mu, \tau, \sigma^2) = \mathcal{U}(-\infty, \mu_\tau).
\]

Here, \(\delta\) is the Dirac delta function.

\(^10\)Write \(\hat{\tau} = \frac{\sqrt{1 - \hat{\theta}_{\text{MLE}}}}{\sqrt{1 - \hat{\theta}_{\text{MLE}}} + \sqrt{\hat{\theta}_{\text{MLE}}}}\), and it is trivial to show that \(\hat{\mu}_\tau \overset{\text{D}}{\to} \hat{\mu}_\tau\).

\(^11\)We would quickly point out that this tradeoff will exist for other parametric models with discontinuous density at the location parameter, so long as that parameter is not the minimum or maximum of the distribution. If the model has two moments, (1) the location parameter will be an expectile \(\mu_\tau\) for some \(\tau \in (0, 1)\) and (2) the sample expectile \(\hat{\mu}_\tau\) will be consistent and asymptotically normal [Holzmann and Klar, 2016].
and the parameter $\theta$ play a significant role in determining $\mu_0$. But it is also clear that this definition is not the same as the functional representation of expectiles in equation 4.1 for any value of $\tau$: the AAP location parameter is something else, even if we take $\theta$ as a predetermined hyperparameter. Gneiting [2011] shows that the “swoosh” function in 4.2 is the unique elicitable scoring functions for expectiles if $F$ has support on $\mathbb{R}$. This means that equation 4.29 cannot elicit the same expectile for every $F$. Thus it is impossible to know which expectile the AAP estimator will elicit without information about $F$.

To summarize this subsection, the AAP likelihood model is not useful for estimating expectiles. First, the AAP estimator does not estimate the $\theta^{\text{th}}$ (or $(1 - \theta^{\text{th}})$) expectile even when $\theta$ is known. Second, the assumed distribution is discontinuous at the location parameter of interest, which makes the limiting distribution nonstandard. Third, it makes little sense to use a quasi-likelihood estimator with a discontinuity at the mode unless we believe the true distribution has such a property; it may not be clear what the estimator is estimating. The likelihood model and estimator given by AAP may be quite useful, but their uses seem to be unrelated to estimating expectiles.

### 4.4.2 Illustrations

A graphical comparison makes the differences between our two distributions clear. See Figures 4.1, 4.2, 4.3, 4.4. Many of the properties discussed in the previous subsection (and elsewhere) can be gleaned from these illustrations.

In Figure 4.1 on the next page, it is obvious that the AAP distribution has discontinuous density at its mode, while the opposite is the case for the expectile density in Figure 4.2. The same fact is clear from the cumulative distributions in Figure 4.3 on page 143 and Figure 4.4 on page 143, where the AAP distribution has a “kink” at $F_0(y) = .5$ and the expectile distribution does not. Also, the AAP distribution can be seen to pass through its median at $\mu_0 = 0$ in this example, which is true for any $\theta \in (0, 1)$. This is not true for the expectile distribution for any $\tau \in (0, 1)$ except $\tau = .5$.

Other results pertaining to maximum likelihood estimators are also clear from the figures. The discontinuous AAP density achieves its maximum (mode) at the median, where the
Figure 4.1: The probability density of the AAP distribution with $\theta = .2$ in red; $\theta = .1$ in blue. Here, $\sigma^2 = 1$ but the variance of the two half-normal distributions is $\frac{\sigma^2}{\theta}$ on the left and $\frac{\sigma^2}{1-\theta}$ on the right.

Figure 4.2: The probability density of the expectile distribution with $\tau$ equal to $.8$, and $.9$. The location parameter $\mu_{\tau} = 0$ and the scale parameter $\sigma_{\tau}^2 = 1$ in both cases.
Figure 4.3: The probability CDF of the AAP distribution with $\theta = 2$ in red; $\theta = .1$ in blue. The median is clearly visible at $y = 0$.

Figure 4.4: The probability CDF of the expectile distribution with $\tau = .8$ in red; $\tau = .9$ in blue. In both cases, $\mu_\tau = 0$ and $\sigma_\tau^2 = 1$. 

143
distribution is not differentiable. That point does not correspond to the mean or the \( \theta^{th} \) expectile, though we can solve for the corresponding expectile under a correctly specified distribution as in (4.27). Because the mode collocates with the median, those authors also suggest a standard \( L_1 \) estimator (Least Absolute Deviations). The expectile distribution attains its mean and median away from the location parameter, which is its mode but also its \( \tau^{th} \) expectile. In that case, the mode collocates only with the \( \tau^{th} \) expectile (not with the mean or median), which leaves the asymmetrically weighted least squares regression of Newey and Powell [1987] and the maximum likelihood estimator shown here as known consistent estimators. Of course, those two estimators are equivalent. Gneiting [2011] shows that this asymmetric squared error loss function \( \varsigma \) contained in these methods is the unique elicitable scoring function for expectiles of random variables with support on \( \mathbb{R} \), as is the case here for both models.

### 4.5 Conclusions

The relationship between the distribution proposed by Aigner et al. [1976] and that discussed by Waldmann et al. [2018] is interesting in several regards. Both are quite simple, at least insofar as they produce an asymmetric form of the Gaussian distribution by applying constant weights to positive and negative errors. The selection of a discontinuous density at the mode (constant probability of positive and negative errors) in the AAP model is the source of the difference between the two. If not for that fact, we could reparameterize either model to become the other. But we cannot—the two distributions are markedly different, their estimators share little if any commonality, and only one seems useful for estimating expectiles by likelihood. If the AAP model is true, its MLE and an expectile estimator are both useful, though their asymptotic properties differ. In that framework, the MLE is more efficient but not asymptotically normal, while the expectile estimator is less efficient but asymptotically normal. In a quasi-likelihood framework where the AAP model is not assumed to be true, the AAP estimator does not elicit any particular expectile.

These two models fit into an interesting niche in probability research. There has been
a long history of development regarding asymmetric versions of the Gaussian distribution. See Azzalini and Valle [1996], for example. Many different variations exist and no particular specification has gained much traction. Even so, the need to extend our likelihood-based estimation abilities beyond the measures of central tendency is obvious; see Kneib [2013]. Asymmetric distributions are the key to achieving that purpose.

These models also fit an interesting historical narrative. The level of public interest in generalized quantile regression methods has increased substantially in the last several years [Waldmann et al., 2018], but dates back several decades. The asymmetrically weighted estimation in Aigner et al. [1976] is an interesting example from the same era as the quantile regression of Koenker and Bassett [1978]. It is also notable that the asymmetric least squares estimator was suggested prior to Newey and Powell [1987] if only in passing and not as a maximum likelihood estimator for the AAP model. The AAP MLE does not carry any obvious interpretation as a quantile or generalized quantile regression for three reasons. First, they suggest estimating the coefficient of asymmetry $\theta$, and do not entertain the notion that multiple $\theta$ could be employed to characterize the entire distribution. Second, their location parameter inhabits the same quantile of the AAP distribution regardless of the choice of asymmetry parameter $\theta$. Third, the concept of quantile regression was not mature at that time in the way that it is today, and there simply was no such suggestion in their paper.

As far as the modern literature is concerned, the distinction between these two papers (Aigner et al. [1976] versus Newey and Powell [1987]) should be clear. AAP did not invent expectiles, give them that name, nor discuss them in any way. The concept of the expectile as a generalized notion of the arithmetic mean originates in Newey and Powell [1987]. That paper places faith in the expectile as a means to describe the full distribution using averages, and explicitly suggests using multiple coefficients of asymmetry $\tau$ for this purpose. Likelihood-based models for expectile regression do not occur in the literature prior to Newey and Powell and apparently not until much later [Waldmann et al., 2017, for example].

Hopefully, we have dispelled some of the confusion regarding this issue. The correct citation for expectiles is Newey and Powell [1987], though it is fair to say that Aigner et al.
[1976] is the origin of asymmetric least squares. The correct likelihood model to elicit expectiles is that found in Waldmann et al. [2018], and the likelihood model of Aigner et al. [1976] is not related to that purpose. Likelihood-based expectile regression is still new to the literature and has considerable potential for further development. As always, any flaws in this work are the sole responsibility of the author.


Stephan Stahlschmidt, Matthias Eckardt, and Wolfgang K Härdle. Expectile treatment effects: An efficient alternative to compute the distribution of treatment effects. 2014.


Appendix A: Appendix to “Does the Government Spending Multiplier Depend On the Business Cycle?”

A.1 Technical Appendix

A.1.1 Model

We use the smooth-transition VAR model employed by Auerbach and Gorodnichenko [2012], Caggiano et al. [2015] and others to study whether the government spending multiplier depends on the business cycle. Equations A.1 to A.4 represent the model

\[ X_t = (1 - F(z_{t-1}))\Pi_E(\ell)X_{t-1} + F(z_{t-1})\Pi_R(\ell)X_{t-1} + u_t \]  \hspace{1cm} (A.1)

\[ u_t \sim N(0, \Omega_t) \]  \hspace{1cm} (A.2)

\[ \Omega_t = (1 - F(z_{t-1}))\Omega_E + F(z_{t-1})\Omega_R \]  \hspace{1cm} (A.3)

\[ F(z_t) = \frac{exp(-\gamma(z_t - c))}{1 + exp(-\gamma(z_t - c))} \]  \hspace{1cm} (A.4)

where \( X_t \) is a \( n \times 1 \) vector of macroeconomic variables, \( \Pi_E \) and \( \Pi_R \) are two sets of \( np \times n \) slope coefficient related to state \( E \) and \( R \), respectively. The \( n \times 1 \) vector of reduced form
errors, \( u_t \), is assumed to be normally distributed with zero mean and covariance matrix \( \Omega_t \). In contrast to linear VAR models, \( \Omega_t \) is not constant but evolves over time as a combination of \( \Omega_E \) and \( \Omega_R \) which are the covariance matrices related to state \( E \) and \( R \), respectively.

The different states are weighted by the function \( F \). \( F \) is a continuous and monotone logistic sigmoid function, bounded between 0 and 1. \( F \) depends on three ingredients. First, \( z_t \) is the state determining variable. We follow Auerbach and Gorodnichenko [2012] and choose the 7-month moving average of the real GDP’s growth rate as \( z_t \). Second, \( c \) is the threshold parameter that governs the dynamics of the model. If \( z_t \) moves from below \( c \) to above \( c \), the dynamics of the model transitions smoothly from state \( R \) to state \( E \). Finally, \( \gamma \) is the transition variable that governs the speed of transition from one state to another. If \( \gamma = 0 \), then \( F(z_t) = 0.5 \): the model collapses to a linear VAR model and no transition occurs. On the other hand, if \( \gamma \to \infty \), the model becomes a threshold-VAR model: the economy jumps from from one extreme state or another.

A.1.2 Multi-move Gibbs Sampling Algorithm

In contrast to Auerbach and Gorodnichenko [2012] and Caggiano et al. [2015], we employ Bayesian estimation techniques to fully characterize the posterior distribution of the model. Bayesian methods require the definitions of prior distributions for all model parameters. In the ST-VAR model, we must choose prior distributions for \( \Pi_E, \Pi_R, \Omega_E, \Omega_R, \gamma \) and \( c \).

Sample \( \Pi_s \) and \( \Omega_s \)

For \( \pi_s = vec(\Pi_s) \) and \( \Omega_s, \; s \in \{ E, R \} \), we choose the Minnesota prior

\[
p(\pi_s) \sim N(\pi_0, V_{s,0})
\]

and

\[
p(\Omega_s) \sim IW(S_0, v_0).
\]

The Minnesota prior assumes that \( X_t \) follows a multivariate random walk process: \( \pi_0 \) is a vector of zeros except for the elements that correspond to the variable’s first own lag. These
elements are equal to one. In this way, the Minnesota prior shrinks many parameters of \( \pi_s \) towards zero. In addition, \( V_{s,0} \) has the following form

\[
V(i, j)_{s,0} = \begin{cases} 
\theta_1 \ell^2 & \text{for parameter on own lags} \\
\frac{\theta_2 \sigma^2_s}{\ell \sigma^2_s} & \text{for parameter on foreign lags}
\end{cases}
\] (A.5)

where \( \ell \) is the lag length, and \( \sigma^2_{s,i} \) is the variance of an AR\((p)\) process for variable \( i \) conditional on state \( s \). \( \theta_1 \) and \( \theta_2 \) are hyperparameters. The Minnesota prior is a conjugate prior, having the same form as the likelihood function of the data. Thus the corresponding marginal posterior distribution are from the same family of distributions. They have the following forms

\[
p(\pi_s | \Omega_s, \gamma, c) \sim N(\bar{\pi}_s, \bar{V}_s)
\]

and

\[
p(\Omega_s | \pi_s, \gamma, c) \sim IW(S_{s,T}, v_T)
\]

with

\[
\bar{V}_s = \left( (X'_{s,t-1} X_{s,t-1} \otimes \Omega_s^{-1}) + V_{s,0}^{-1} \right)^{-1}
\] (A.6)

\[
\bar{\pi}_s = \bar{V}_s \left( (X'_{s,t-1} X_{s,t-1} \otimes \Omega_s^{-1}) \hat{\pi}_s + V_{s,0}^{-1} \pi_0 \right)
\] (A.7)

\[
S_{s,T} = (X_t - \Pi_s X_{s,t-1})' (X_t - \Pi_s X_{s,t-1})
\] (A.8)

\[
v_T = T - p
\] (A.9)

where \( \hat{\pi}_s \) is the conditional MLE estimate of \( \pi_s \) whose derivation is shown below. Lastly,

\[
X_{s,t} = \begin{cases} 
(1 - F(z_t))X_t & \text{if } s=E \\
F(z_t)X_t & \text{if } s=R
\end{cases}
\] (A.10)
Sample $\gamma$

In contrast to $\Pi_s$ and $\Omega_s$, $\gamma$ does not have a closed-form for its marginal posterior distribution. Hence, we rely on MCMC methods to approximate the marginal posterior distribution. We follow Gefang and Strachan [2010] and use the Metropolis-Hastings algorithm proposed by Chib and Greenberg [1995]. Suppose $\gamma^*$ is a draw from a proposal density $q(\gamma^*|\gamma^{k-1})$ and $\gamma^{k-1}$ is the draw of $\gamma$ from the previous iteration. Then, we set $\gamma^k = \gamma^*$ with probability

$$\alpha(\gamma^*|\gamma^{k-1}) = \min\{1, \frac{p(\gamma^*|\cdot)/q(\gamma^*|\gamma^{k-1})}{p(\gamma^{k-1}|\cdot)/q(\gamma^{k-1}|\gamma^*)}\} \tag{A.11}$$

where $p(\gamma^*|\cdot)$ and $p(\gamma^{k-1}|\cdot)$ are the posterior distributions–conditional on $X$ and the other parameters–of the candidate value $\gamma^*$ and the incumbent value $\gamma^{k-1}$, respectively. In addition, $q(\gamma^*|\gamma^{k-1})$ and $q(\gamma^{k-1}|\gamma^*)$ are the corresponding proposal densities.

We assume a gamma prior for $\gamma$

$$p(\gamma) = \mathcal{G}(\mu_\gamma, \nu_\gamma) \tag{A.12}$$

where $\mu_\gamma$ and $\nu_\gamma$ are the mean and degrees of freedom, respectively. Hence, $p(\gamma^*|\cdot)$ and $p(\gamma^{k-1}|\cdot)$ are also Gamma distributions. Finally, the proposal density $q(\gamma^*|\gamma^{k-1})$ and $q(\gamma^{k-1}|\gamma^*)$ are chosen to be Gamma distributions as well. The sequence of $\{\gamma^k\}_{k=\text{burn.in}}^K$ (draws of $\gamma$ after a sufficient burn-in phase) can be viewed as draws from the (target) marginal posterior distribution of $\gamma$.

Sample $c$

Similarly to the marginal posterior distribution of $\gamma$, the posterior distribution of $c$ does not have a closed-form. We apply the Griddy Gibbs sampler of Ritter and Tanner [1992]. We assume a uniform prior for $c$.
\[ p(c) = \mathcal{U}(\underline{u}, \overline{u}) \] (A.13)

where \( \underline{u} \) and \( \overline{u} \) represent the lower and upper limit of the 80 percent middle of \( z_t \). First, we select grid points over the space of \( c \) we want to draw from, here this space is \( \underline{u}, \overline{u} \). Then, we evaluate the conditional posterior distribution at the grid values and compute the corresponding (inverse) CDF. Finally, we can evaluate the marginal posterior distribution of \( c \) by drawing a random variable from \( \mathcal{U}(0, 1) \) and transforming the draw via the inverse CDF to obtain a draw of \( c \) from its (approximated) marginal posterior distribution.

**Algorithm**

The previous steps correspond to algorithm A.1. We repeat this algorithm 20,000 times while the first 10,000 iterations are disregarded as burn-in. The remaining 10,000 draws of \( \{\Pi_E, \Pi_R, \Omega_E, \Omega_R, \gamma, c\} \) are used to approximate the joint posterior distribution of the model and used to compute the orthogonalized and generalized impulse response functions and spending multipliers.

**Algorithm A.1 Multi-Move Gibbs Sampler**

1. Initialize: Choose \( \pi^0_E, \pi^0_R, \Omega^0_E, \Omega^0_R, \gamma^0, c^0 \);
2. Draw \( \Omega_E|\pi_E, \pi_R, \gamma, c \sim IW(S_E, T) \) and \( \Omega_R|\pi_E, \pi_R, \gamma, c \sim IW(S_R, T) \);
3. Draw \( \pi_E, \pi_R|\Omega_E, \Omega_R, \gamma, c \sim N(\overline{\pi}, \overline{V}) \);
4. Draw \( \gamma|\pi_E, \pi_R, \Omega_E, \Omega_R, c \) using a Metropolis-Hastings step;
5. Draw \( c|\pi_E, \pi_R, \Omega_E, \Omega_R, \gamma \) from a Griddy Gibbs sampler;
6. Repeat steps 2 through 5 and keep the desired number of draws after a burn-in phase.

**Alternative Algorithm**

Algorithm A.1 draws \( \Omega_E, \Omega_R \) independently. If dependence between these parameters is of interest, an alternative replaces that step (step 2) with a Metropolis-Hastings step. This alternative is somewhat slower than Algorithm A.1 but produces the same results.

Here we explain the Metropolis-Hastings step to obtain draws of \( \Omega_E, \Omega_R \) conditional on
the other parameters. The MH algorithm operates by drawing parameters from a proposal density where a sampler is available, and then comparing the proposal density with the posterior. The posterior in this case is

$$|\Sigma|^{-1/2}|\Omega_E\Omega_R|^{-(v_0+p+1)/2}\exp\left\{-\frac{1}{2}\text{tr}\left(\Sigma\hat{u}\hat{u}'\right)\right\}\exp\left\{-\frac{1}{2}\text{tr}\left(S_0\Omega_E^{-1} + S_0\Omega_R^{-1}\right)\right\} \quad (A.14)$$

where the elements may be alternately expressed as

$$|\Sigma|^{-1/2} = \prod_{t=1}^{T} |\Omega_t|^{-1/2}$$

$$|\Omega_E\Omega_R|^{-(v_0+p+1)/2} = |\Omega_E|^{-(v_0+p+1)/2}|\Omega_R|^{-(v_0+p+1)/2}$$

$$\exp\left\{-\frac{1}{2}\text{tr}\left(\Sigma^{-1}\hat{u}\hat{u}'\right)\right\} = \exp\left\{-\frac{1}{2}\sum_{t=1}^{T}\text{tr}\left(\hat{u}_t\Omega_t^{-1}\hat{u}_t\right)\right\}.$$

For a given draw, draw two candidate matrices $\Omega^*_E, \Omega^*_R$ from a proposal density. We choose the following Inverse-Wishart proposal density

$$\Omega^*_E \sim IW(\tilde{S}_E, T), \quad \Omega^*_R \sim IW(\tilde{S}_R, T)$$

where the shape parameters $\tilde{S}_E, \tilde{S}_R, T$ are (preferably) close to the true posterior density\(^\text{12}\). Then, to make the proposal density arbitrary, we compare the posterior density to the proposal density for both the candidate draws $\Omega^*_E, \Omega^*_R$ and the previous set of accepted draws, $\Omega^{k-1}_E, \Omega^{k-1}_R$. Denote the proposal density as $q(\Omega^*_E, \Omega^*_R | \cdot)$ and the conditional posterior density $p(\Omega^*_E, \Omega^*_R | \cdot)$ where the dot represents the given parameters and the data. Clearly $p(\Omega^*_E, \Omega^*_R | \cdot)$ is proportional to expression A.14 and the proposal density is proportional to

$$q(\Omega^*_E, \Omega^*_R | \tilde{S}_E, \tilde{S}_R, T) \propto |\Omega_E\Omega_R|^{-(T+p+1)/2}\exp\left\{-\frac{1}{2}\text{tr}\left(\tilde{S}_E[\Omega_E]^{-1} + \tilde{S}_R[\Omega_R]^{-1}\right)\right\}.$$  

As in Chib and Greenberg [1995], the MH algorithm requires keeping $\Omega^*_E, \Omega^*_R$ with

\(^{12}\)For $\tilde{S}_E, \tilde{S}_R$, we suggest the weighted residual covariance matrices $S_{E,T}, S_{R,T}$, defined in equation (A.8).
probability

\[
\min \left[ \frac{p(\Omega_E^*, \Omega_R^*)/q(\Omega_E^*, \Omega_R^*)}{p(\Omega_E^{k-1}, \Omega_R^{k-1})/q(\Omega_E^{k-1}, \Omega_R^{k-1})} \right]
\]

and, if not, keeping \(\Omega_E^{k-1}, \Omega_R^{k-1}\) instead. Note that the ratios in the numerator and denominator simplify slightly:

\[
p(\Omega_E^*, \Omega_R^*)/q(\Omega_E^*, \Omega_R^*) \propto |\Sigma^*|^{-1/2} |\Omega_E^*|^{-(v_0 + p + 1)/2} \exp \left\{ -\frac{1}{2} \text{tr} \left( \Sigma^* \hat{u} \hat{u}' \right) \right\}
\]

\[
\times \exp \left\{ -\frac{1}{2} \text{tr} \left( S_0[\Omega_E^*]^{-1} + S_0[\Omega_R^*]^{-1} \right) \right\}
\]

\[
\times |\Omega_E^*|^{(T+p+1)/2} \exp \left\{ \frac{1}{2} \text{tr} \left( \hat{S}_E[\Omega_E^*]^{-1} + \hat{S}_R[\Omega_R^*]^{-1} \right) \right\}
\]

\[
\propto |\Sigma^*|^{-1/2} \exp \left\{ -\frac{1}{2} \text{tr} \left( \Sigma^* \hat{u}' \hat{u} \right) \right\} |\Omega_E^*|^{(T-v_0)/2}
\]

\[
\times \exp \left\{ \frac{1}{2} \text{tr} \left( (\hat{S}_E - S_0)[\Omega_E^*]^{-1} + (\hat{S}_R - S_0)[\Omega_R^*]^{-1} \right) \right\}
\]

and the same for \(\Omega_E^{k-1}, \Omega_R^{k-1}\), proportional to the same constant.

**A.1.3 Derivation of The Conditional MLE**

Here, we derive the MLE, \(\hat{\Pi}\), used to calculate the posterior mode used in the sampling algorithm (step 3). This result is fairly standard. Take \(\gamma, c, \Omega_E, \Omega_R\) as given. The minimizer of \(S(\Pi) = \sum_{t=1}^T u_t^t \Omega_t^{-1} u_t = u^t \Sigma^{-1} u\) can be found using the following expression

\[
S(\Pi) = \text{vec}(U')\Sigma^{-1}\text{vec}(U')
\]

\[
= \text{vec}(Y' - \Pi'W')\Sigma^{-1}\text{vec}(Y' - \Pi'W')
\]

\[
= (\text{vec}(Y') - (W \otimes I)\text{vec}(\Pi'))' \Sigma^{-1} (\text{vec}(Y') - (W \otimes I)\text{vec}(\Pi'))
\]

\[
= \text{vec}(Y')'\Sigma^{-1}\text{vec}(Y') + \text{vec}(\Pi')'(W' \otimes I)\Sigma^{-1}(W \otimes I)\text{vec}(\Pi')
\]

\[
- 2\text{vec}(\Pi')'(W' \otimes I)\Sigma^{-1}\text{vec}(Y')
\]

which leads to the first-order condition (the second-order condition is obviously satisfied).
\[
\frac{\partial S(\Pi)}{\partial \text{vec}(\Pi')} = 2(W' \otimes I)\Sigma^{-1}(W \otimes I)\text{vec}(\Pi') - 2(W' \otimes I)\Sigma^{-1}\text{vec}(Y') = 0
\]

Or

\[
\text{vec}(\Pi') = (\left(W \otimes I\right)'\Sigma^{-1}(W \otimes I))^{-1}\left((W \otimes I)'\Sigma^{-1}\text{vec}(Y')\right) \quad \text{(A.15)}
\]

Then \(\text{vec}(\Pi)\) can be found using the commutation matrix \(K\) and the identity

\[
\text{vec}(\Pi) = K\text{vec}(\Pi')
\]

which connects the expression in (A.15) to that in Auerbach and Gorodnichenko [2012].

The weighted annihilator matrix \(M_V\) is given by

\[
I_{kT} - (W \otimes I) \left((W \otimes I)'\Sigma^{-1}(W \otimes I)\right)^{-1}(W \otimes I)'\Sigma^{-1}.
\]

### A.1.4 Estimation of the GIRF

While the computation of orthogonalized impulse response functions is standard in the literature, generalized impulse response functions need further explanation. The generalized impulse response functions of Koop et al. [1996] can be computed after we have obtained a set of draws from the posterior. To do this, we follow the steps given in Algorithm A.2 on the next page.
Algorithm A.2 Sampler for Generalized Impulse Responses

1. Using the full sample of data $X_1,...,X_T$, denote the set of all possible histories $\Lambda$ as the set of blocks $\lambda_i = [X_{i-p+1},...,X_i]$ of the minimum length necessary to evaluate the model (determined either by lag length of $\Pi(l)$ or the number of MA terms in $z_{t-1}$. In our case, histories $\lambda_i \in \Lambda$ require $p = 6$ observations.

2. Create a set of recessionary histories $\Lambda^R$ and a set of expansionary histories $\Lambda^E$. For each history $\lambda_i$, $\lambda_i \in \Lambda^R$ if $z_{\lambda_i}$ is in the lowest decile of $z$ and $\lambda_i \in \Lambda^E$ otherwise.

3. Draw a set of parameters $\pi^E, \pi^R, \Omega^E, \Omega^R, \gamma, c$ from the posterior set.

4. Draw one history $\lambda_i \in \Lambda^R$ at random. Construct $\Omega_{\lambda_i}$ as

$$
\Omega_{\lambda_i} = (1 - F(z_{\lambda_i}))(\Omega_E + F(z_{\lambda_i})\Omega_R).
$$

5. Draw one orthogonal matrix $Q$ and use the lower Cholesky decomposition of $\Omega_{\lambda_i}$ to produce the structural shocks $e_{\lambda_i} = Q^{-1} Chol(\Omega_{\lambda_i})^{-1} \hat{U}$. 

6. Draw with replacement $h$ shocks from $e_{\lambda_i}$ to produce a bootstrapped shock series $e^*_{\lambda_i} = [e^*_{\lambda_i,t},...,e^*_{\lambda_i,t+h}]$.

7. Duplicate the shock series $e^*_{\lambda_i}$ but add the impulse vector $v_t$ at time $t$.

$$
e^*_\lambda = [e^*_{\lambda,t} + v_t, ..., e^*_{\lambda,t+h}]$$

In our case, $v_t$ is a one-standard-deviation increase in government spending.

8. Return $e^*_{\lambda_i}$ and $e^*_\lambda$ to the reduced-form model representation by

$$
u^*_{\lambda_i} = Chol(\Omega_{\lambda_i})Q e^*_{\lambda_i}
$$

$$
u^*_\lambda = Chol(\Omega_{\lambda_i})Q e^*_\lambda
$$

9. Use $u^*_{\lambda_i}$ and $u^*_\lambda$ to roll the model forwards and obtain a candidate for GIRF($h, \lambda_i$).

10. If GIRF($h, \lambda_i$) satisfies the correct sign restrictions, keep it. Otherwise, repeat steps 5 - 9 with a different orthogonal matrix $Q$.

11. Repeat steps 4 - 10 $R$ times and take the average. This is the expected response, conditional on the parameter draws from the posterior.

12. Repeat steps 4 - 11 for expansionary histories $\lambda_i \in \Lambda^E$.

13. Repeat steps 3 - 12 with a different draw from the posterior until the desired number of draws has been obtained.
A.2 Impulse Response Functions

Figure A.1: Orthogonalized Impulse Response Functions - Recursive Identification

Note: The solid red and blue lines are the median IRFs for expansions and recessions, respectively. The red and blue shaded areas are the 68 percent credible bands. This Figure shows the IRFs using orthogonalized IRFs with recursive identification as in Auerbach and Gorodnichenko [2012].

Figure A.2: Orthogonalized Impulse Response Functions - Sign Restrictions

Note: The solid red and blue lines are the median IRFs for expansions and recessions, respectively. The red and blue shaded areas are the 68 percent credible bands. This Figure shows the IRFs using our sign restriction approach.
Figure A.3: Generalized Impulse Response Functions - Sign Restrictions

Note: The solid red and blue lines are the median IRFs for expansions and recessions, respectively. The red and blue shaded areas are the 68 percent credible bands. This Figure shows the GIRFs using our sign restriction approach.
A.3 Additional Figures

Figure A.4: GIRF Multipliers - Government Investment

![Figure A.4: GIRF Multipliers - Government Investment](image1)

Note: Figure shows the distributions of estimated multipliers for expansions and recessions across different time horizons for government investment. Using GIRFs to estimate the multipliers, there is no evidence for state-dependence in the medium- and long-run.

Figure A.5: GIRF Multipliers - Government Consumption

![Figure A.5: GIRF Multipliers - Government Consumption](image2)

Note: Figure shows the distributions of estimated multipliers for expansions and recessions across different time horizons for government consumption. Using GIRFs to estimate the multipliers, there is no evidence for state-dependence in the medium- and long-run.
Figure A.6: GIRF Multipliers - Government Defense Spending

Note: Figure shows the distributions of estimated multipliers for expansions and recessions across different time horizons for defense spending. Using GIRFs to estimate the multipliers, there is no evidence for state-dependence in the medium- and long-run.

Figure A.7: GIRF Multipliers - Government Nondefense Spending

Note: Figure shows the distributions of estimated multipliers for expansions and recessions across different time horizons nondefense spending. Using GIRFs to estimate the multipliers, there is no evidence for state-dependence in the medium- and long-run.
Figure A.8: GIRF Multipliers - Federal Spending

Note: Figure shows the distributions of estimated multipliers for expansions and recessions across different time horizons federal spending. Using GIRFs to estimate the multipliers, there is no evidence for state-dependence in the medium- and long-run.

Figure A.9: GIRF Multiplier - State & Local Spending

Note: Figure shows the distributions of estimated multipliers for expansions and recessions across different time horizons using state and local spending spending. Using GIRFs to estimate the multipliers, there is no evidence for state-dependence in the medium- and long-run.
Figure A.10: GIRF Multiplier - Ramey’s News Defense Shocks

Note: Figure shows the distributions of estimated multipliers for expansions and recessions across different time horizons. Using GIRFs to estimate the multipliers, there is no evidence for state-dependence in the medium- and long-run. In this specification, we consider Ramey’s News Defense Shocks instead of actual government spending as the shock variable.

Figure A.11: GIRF Multiplier - Large Positive Shocks

Note: Figure shows the distributions of estimated multipliers for expansions and recessions across different time horizons using government consumption expenditures & gross investment. Using GIRFs to estimate the multipliers, there is no evidence for state-dependence in the medium- and long-run. In this specification, we add two units instead of one unit to government consumption expenditures & gross investment.
Figure A.12: GIRF Multiplier - Negative Shocks

Note: Figure shows the distributions of estimated multipliers for expansions and recessions across different time horizons using government consumption expenditures & gross investment. Using GIRFs to estimate the multipliers, there is no evidence for state-dependence in the medium- and long-run. In this specification, we add minus one unit instead of one unit to government consumption expenditures & gross investment.
A.4 Model Parameter Posteriors

The state indicator function $F$, as given elsewhere,

$$F(z_t) = \frac{exp(-\gamma(z_t - c))}{1 + exp(-\gamma(z_t - c))}$$  \hspace{1cm} (A.16)

employs two tuning parameters: $c$ and $\gamma$. The posterior density of these parameters is shown below.

Figure A.13: Posterior Density of $c$

Note: Figure shows the marginal posterior distribution of the parameter $c$. The flatness of the posterior suggests that the tuning parameter $c$ has little effect on the explanatory power of the model. If we take the mode around $c = .4$ seriously, it would suggest that the dynamics (regime) of the economy transition most rapidly when $z_t$ is approximately .4.
Figure A.14: Posterior Density of $\gamma$

Note: Figure shows the marginal posterior distribution of the parameter $\gamma$ from equation (A.16). There is little density near $\gamma = 0$, suggesting that a linear model is inappropriate. Similarly, the right tail gives negligible weight to the possibility that $\gamma > 8$, which does not support the use of a threshold type model implied by $\gamma = \infty$.

Figure A.15: Posterior Density of the Median of $F(z_t)$

Note: Figure shows the histogram of the posterior medians of $F(z_t)$ evaluated at each $z_t$ in the sample. The distribution is bimodal with peaks at each end, supporting the idea that the two-regime model is reasonable.
Appendix B: Appendix to “Government Spending Between Active and Passive Monetary Policy?”

B.1 Mathematical Appendix

This section provides a more detailed overview of the TVP-ST-VAR model used in the main text. The model is a hybrid between the smooth-transition VAR model, popularized by Auerbach and Gorodnichenko [2012], and a univariate regression with time-varying parameters. The latter part is used to estimate the state determining variable $z_t$, which is then employed to distinguish between different monetary policy regimes in the main model. As argued in the main text, this extension is necessary because monetary policy is not well explained by a single observable variable. In the following, we explain the model in greater detail and lay out each step of the estimation procedure. First, the model is characterized by the following equations

$$X_t = G(z_{t-1})\Pi_{AM}X_{t-1} + (1 - G(z_{t-1}))\Pi_{PM}X_{t-1} + u_t$$ (B.1)

$$u_t \sim N(0, \Omega_t)$$ (B.2)

$$\Omega_t = G(z_{t-1})\Omega_{AM} + (1 - G(z_{t-1}))\Omega_{PM}$$ (B.3)
\[ G(z_t) = \frac{1}{1 + \exp(-\gamma(z_t - c))} \]  

(B.4)

\( X_t \) is a vector of endogenous variables and represents the underlying economy. The state of the economy, \( G(z) \) evolves continuously over time. The model B.1 - B.4 approximates the true evolution of this state via an evolving mix of two extreme states. Here, monetary policy is purely active in the AM regime and purely passive in the PM regime. The relative weight assigned to the purely active regime is given by the transition function \( G \), which constitutes our measure of monetary policy activism. \( G \) is parameterized as a logistic sigmoid curve—continuous, monotonically increasing, and bounded between 0 and 1. If \( z_t \) increases, the value of \( G \) increases and the model assigns relatively more weight towards the AM regime. Here, \( \gamma \) governs the speed with which this transition occurs. If \( \gamma \approx 0, G \approx 0.5 \forall t \), the model becomes a linear model and no transition occurs. If \( \gamma \approx \infty \), the model jumps from one extreme regime to another as soon as \( z_t \) the threshold parameter \( c \). In this case, the model becomes a threshold VAR model, thus nesting the Markov-type models Krolzig [1998].

To characterize monetary policy as the evolving, potential nonlinear, state of the economy, we follow the DSGE literature and estimate an interest-rate rule, i.e., a version of the Taylor rule as introduced by Taylor [1993]. Because there is strong evidence that the coefficients of the Taylor rule and the variance of the corresponding residuals vary over time, we estimate a Taylor rule with time-varying parameters and stochastic volatility. More formally, we estimate

\[ y_t = z_t' \phi_t + e_t, \quad e_t \sim N(0, \sigma_t^2) \]  

(B.5)

\[ \phi_t = \phi_{t-1} + v_t, \quad v_t \sim N(0, q) \]  

(B.6)

\[ \log(\sigma_t) = \log(\sigma_{t-1}) + \eta_t, \quad \eta_t \sim N(0, w) \]  

(B.7)

\[ ^{13} \text{We exploit this feature in our replication exercise. See the main text.} \]
where \( y_t \) is the federal funds rate, \( z_t \) consists of a constant, inflation and output. \( \phi_t \) is a vector of time-varying coefficients. \( \phi_t \) follows a random walk process which allows for both temporary and permanent shifts in the parameters. We use the time-varying inflation parameter, \( \phi_{\pi,t} \) as \( z_t \) in the model B.1 - B.4. The variance of the residuals \( e_t \) can also vary over time. We assume that \( \log(\sigma_t) \) follows a random walk.

We estimate B.1-B.4 and B.5-B.7 separately in two steps using Bayesian estimation techniques. Bayesian methods allow the data to become informative about the model structure. This feature is particularly important because the data are also informative about the evolution of monetary policy which is key to understanding how the effect of government spending depends on monetary policy. We next lay out each step of the estimation procedure.

B.1.1 Estimation of Taylor Rule with Time-Varying Coefficients and Stochastic Volatility

B.5 and B.6 form a linear-Gaussian state space model. This structure allows us to sample the latent states \( \phi_t \) via Kalman filtering techniques. In contrast, B.5 and B.7 form a Gaussian but nonlinear state space model. We follow Kim et al. [1998] and Primiceri [2005], Del Negro and Primiceri [2015] and transform this structure into a linear-Gaussian state space model and then use Kalman filtering techniques to sample \( \sigma_t \). To sample the time-varying coefficients \( \phi^T \), the stochastic volatility \( \sigma^T \) and the hyperparameter \( q \) and \( w \), we employ the following algorithm

**Algorithm B.1 Multi-Move Gibbs Sampler for TVP Taylor rule**

1. Initialize: \( \phi_0, q_0, \sigma_0 \) and \( w_0 \);
2. Draw \( \phi^T | X^T, q, \sigma^T, w \) using Kalman filtering methods;
3. Draw \( q|\phi^T \);
4. Draw \( \sigma^T | X^T, \phi^T, q, w \) using Kalman filtering methods;
5. Draw \( w|\sigma^T \)
6. Repeat steps 2 through 5 and keep the desired number of draws after a burn-in phase.
Time-Varying Coefficients

Conditional on $\sigma^T$ and $q$, B.5 is linear and Gaussian with a known variance. Following Frühwirth-Schnatter [1994] and Carter and Kohn [1994], the density of $\phi^T$, $p(\phi^T|X^T, \sigma^T, q)$, can be factored as

$$p(\phi^T|X^T, \sigma^T, q) = p(\phi_T|X_T, \sigma_T, q) \prod_{t=1}^{T-1} p(\phi_t|\phi_{t+1}, X^t, \sigma^T, q) \quad (B.8)$$

where

$$p(\phi_t|\phi_{t+1}, X^t, \sigma^T, q) \sim N(\phi_{t|t+1}, P_{t|t+1}) \quad (B.9)$$

$$\phi_{t|t+1} = E(\phi_t|\phi_{t+1}, X^t, \sigma^T, q) \quad (B.10)$$

$$P_{t|t+1} = Var(\phi_t|\phi_{t+1}, X^t, \sigma^T, q). \quad (B.11)$$

We use the (forward) Kalman filter to estimate $\phi_T|T$ and $P_T|T$, the mean and variance of the distribution of $\phi_T$. A draw of this distribution is then used in the (backwards) Kalman smoother to estimate $\phi_{t|t+1}$ and $P_{t|t+1}$. Draws from the corresponding distributions yield the whole sequence of $\phi_t$ for $t = \{1, ..., T - 1\}$. A general overview about Kalman filtering techniques is given in Appendix B.1.1.

Stochastic Volatility

Consider the following model

$$y_t - Z_t^\prime \phi_t = y_t^* = \sigma_t \xi_t, \xi_t \sim N(0, 1) \quad (B.12)$$

$$log(\sigma_t) = log(\sigma_{t-1}) + \eta_t, \eta_t \sim N(0, w) \quad (B.13)$$

The model B.12 and B.13 forms a Gaussian, but nonlinear state space model. However, the model can be transformed into a linear model by squaring and taking the log of B.12

---

14Here, the superscript $T$ indicates the entire history of a time-varying parameter up to time $T$; i.e. $\sigma^T = \{\sigma_1, ..., \sigma_T\}$. 

---

179
\[ \log(y_t^2) = 2 \log(\sigma_t) + \log(\xi_t^2) \quad (B.14) \]

or

\[ y_t^{**} = 2h_t + \nu_t. \quad (B.15) \]

(B.15) and (B.13) form a linear model, but \( \nu_t \sim \log \chi^2(1) \). Following Kim et al. [1998] and Primiceri [2005], we approximate the \( \log \chi^2(1) \) distribution by a mixture of seven normal distributions with weights \( q_i \), means \( m_i - 1.2704 \) and variance \( v_i^2 \). The constants \( \{q_i, m_i, v_i^2\} \) are known and provided in Table B.1 Following Kim et al. [1998] and Primiceri [2005], we define \( s^T = [s_1, ..., s_T] \) where \( s_t \) represents the indicator variable, \( i \), that belongs to the normal distribution that is used to approximate the distribution of \( \nu \) in period \( t \), i.e., \( \nu_t | s_t \sim N(m_i - 1.2704, v_i^2) \) and \( \text{Prob}(s_t = i) = q_i \). Then, conditional on \( h^T, w, s^T \) and the data (B.15) and (B.13) is approximately linear and Gaussian. This structure implies that \( h_t \) can be sampled using Kalman filtering. Similar to the step that samples \( \phi^T \), the distribution of \( h^T \) can be factored as

\[ p(h^T | X^T, \phi^T, w, s^T) = p(h_T | X_T, \phi_T, w, s^T) \prod_{t=1}^{T-1} p(h_t | h_{t+1}, X^t, \phi^T, w, s^T) \quad (B.16) \]

with

\[ p(h_t | h_{t+1}, X^t, \phi^T, w, s^T) \sim N(h_t | h_{t+1}, B_{t|t+1}) \quad (B.17) \]

\[ h_{t|t+1} = E(h_t | h_{t+1}, X^t, \phi^T, w, s^T) \quad (B.18) \]

\[ B_{t|t+1} = \text{Var}(h_t | h_{t+1}, X^t, \phi^T, w, s^T). \quad (B.19) \]
Table B.1: Mixture of Normal Distribution

<table>
<thead>
<tr>
<th>$w$</th>
<th>$q_i$</th>
<th>$m_i$</th>
<th>$v_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00730</td>
<td>-10.12999</td>
<td>5.79596</td>
</tr>
<tr>
<td>2</td>
<td>0.10556</td>
<td>-3.97281</td>
<td>2.61369</td>
</tr>
<tr>
<td>3</td>
<td>0.00002</td>
<td>-8.566866</td>
<td>5.17950</td>
</tr>
<tr>
<td>4</td>
<td>0.044395</td>
<td>2.77786</td>
<td>0.16735</td>
</tr>
<tr>
<td>5</td>
<td>0.34001</td>
<td>0.61942</td>
<td>0.64009</td>
</tr>
<tr>
<td>6</td>
<td>0.24566</td>
<td>1.79518</td>
<td>0.34023</td>
</tr>
<tr>
<td>7</td>
<td>0.25750</td>
<td>-1.08819</td>
<td>1.26261</td>
</tr>
</tbody>
</table>

Note: See Kim et al. [1998] for details.

**Hyperparameters for $q$ and $w$**

Finally, we discuss the hyperparameters for the variances $q$ and $w$. Conditional on $\phi^T$, $\sigma^T$ and $X^T$, the residuals $v_t$ and $\eta_t$ are observable, and $q$ and $w$ both have an inverse Gamma distribution, i.e.

\[
q \sim IG(\Psi_T, \psi_T) \tag{B.20}
\]

\[
w \sim IG(K_T, \kappa_T) \tag{B.21}
\]

where $\Psi_T = \Psi_0 + \sum_{j=1}^{T-1}(\phi_j - \phi_{j-1})(\phi_j - \phi_{j-1})'$, $\psi_0 = \psi_T + T$, $K_T = K_0 + \sum_{j=1}^{T-1}(h_j - h_{j-1})^2$ and $\kappa_T = \kappa_0 + T$ with $\Psi_0 = 0.04 \times I$, $\psi_0 = 40$, $K_0 = 1$ and $\kappa_0 = 2$.

Figure B.1 shows the estimated evolution of the inflation parameter, $\phi_{\pi,t}$ of the baseline specification from the main text. We then use the pointwise posterior median estimate of the inflation parameter as $z_t$ in the VAR part of the model.

**Kalman Filter and Smoother**

In Section B.1.1 we used Kalman filtering methods to filter $\alpha^T$ and $\sigma^T$. This section provides a more detailed overview about these methods. Consider the following general state space model:
Figure B.1: Estimated Inflation Parameter Series

Note: The Figure shows the estimate posterior distribution of the inflation parameter, $\phi_{\pi,t}$ of our baseline specification from the main text. The black line represents the pointwise posterior median estimate along with the 95 percent credible bands in red.

\[ y_t = Z\alpha_t + e_t, \quad e_t \sim N(0, H) \]  \hspace{1cm} (B.22)

\[ \alpha_t = T\alpha_{t-1} + R\eta_t, \quad \eta_t \sim N(0, Q). \]  \hspace{1cm} (B.23)

Kalman filtering techniques take $Z, H, T, R, Q$ as given and provide an estimate of $\alpha_t$ via the following forecasting and updating steps

\[ a_{t|t-1} = Ta_{t-1} \]  \hspace{1cm} (B.24)

\[ P_{t|t-1} = TP_{t-1}T' + RQR \]  \hspace{1cm} (B.25)

\[ a_t = a_{t|t-1} + P_{t|t-1}Z'F_t^{-1}(y_t - Za_{t|t-1}) \]  \hspace{1cm} (B.26)

\[ P_t = P_{t|t-1} - P_{t|t-1}Z'F_t^{-1}ZP_{t|t-1} \]  \hspace{1cm} (B.27)

where $F_t = ZP_{t|t-1}Z' + H$. The Kalman filter runs forward for $t = 1, ..., T$. Then,
\[ \alpha_T \sim N(a_T, P_T). \]  
(B.28)

Conditional on \( \alpha_T \), the Kalman smoother runs backwards and samples \( \alpha_t \) for \( t = T - 1, T - 2, \ldots, 0 \) using the following steps

\[
\alpha_{t|t+1} = \alpha_t + P_t T_{t+1} P_{t+1}^{-1} (\alpha_{t+1} - Z_o_t) \tag{B.29}
\]
\[
P_{t|t+1} = P_t - P_t T_{t+1} P_{t+1}^{-1} T P_t \tag{B.30}
\]

Then, \( \alpha_t \sim N(\alpha_{t+1|t}, P_{t+1|t}) \) for \( t = T - 1, T - 2, \ldots, 0 \).

### B.1.2 Estimation of the ST-VAR Model

In this section, we describe the Bayesian estimation of the ST-VAR model. To sample the model parameters \( \{\pi_{AM}, \pi_{PM}, \Omega_{AM}, \Omega_{PM}, c, \gamma\} \), we employ the following algorithm:

**Algorithm B.2** Multi-Move Gibbs Sampler for ST-VAR model

1. Initialize: Choose \( \pi_{AM}^0, \pi_{PM}^0, \Omega_{AM}^0, \Omega_{PM}^0, \gamma^0, c^0 \);
2. Draw \( \pi_{AM}|\Omega_{AM}, \Omega_{PM}, \gamma, c \sim N(m_{AM}, V_{AM}) \) and \( \pi_{PM}|\Omega_{AM}, \Omega_{PM}, \gamma, c \sim N(m_{PM}, V_{PM}) \);
3. Draw \( \gamma|\pi_{AM}, \pi_{PM}, \Omega_{AM}, \Omega_{PM}, c \) using a Metropolis-Hastings step;
4. Draw \( c|\pi_{AM}, \pi_{PM}, \Omega_{AM}, \Omega_{PM}, \gamma \) from a Griddy Gibbs sampler;
5. Draw \( \Omega_{AM}, \Omega_{PM}|\pi_{AM}, \pi_{PM}, \gamma, c \) using a Metropolis-Hastings step;
6. Repeat steps 2 through 5 and keep the desired number of draws after a burn-in phase.

**Sample \( \Pi_{AM} \) and \( \Pi_{PM} \)**

Define \( \pi_r = vec(\Pi_r) \) for \( r \in \{AM, PM\} \). For \( \pi_{AM} \) and \( \pi_{PM} \), we choose the Minnesota prior, i.e.,

\[ p(\pi_r) \sim N(m_0, V_0). \]  
(B.31)

\(^{15}\)See, for instance, Primiceri [2005] for further details.
Table B.2: Correlation of Variables, 1954Q3-2008Q3

<table>
<thead>
<tr>
<th></th>
<th>Real Government Spending</th>
<th>Real Tax Receipts</th>
<th>Real GDP</th>
<th>Ramey’s News Shocks</th>
<th>GDP Inflation</th>
<th>Federal Funds Rate</th>
<th>Real Government Debt</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Correlation With:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real Government Spending</td>
<td>1.00</td>
<td>0.10</td>
<td>0.25</td>
<td>0.07</td>
<td>-0.13</td>
<td>-0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>Real Tax Receipts</td>
<td>0.10</td>
<td>1.00</td>
<td>0.51</td>
<td>-0.09</td>
<td>-0.10</td>
<td>-0.05</td>
<td>-0.05</td>
</tr>
<tr>
<td>Real GDP</td>
<td>0.25</td>
<td>0.51</td>
<td>1.00</td>
<td>-0.01</td>
<td>-0.24</td>
<td>-0.20</td>
<td>-0.01</td>
</tr>
<tr>
<td>Ramey’s News Shocks</td>
<td>0.07</td>
<td>-0.09</td>
<td>-0.01</td>
<td>1.00</td>
<td>-0.02</td>
<td>-0.08</td>
<td>-0.05</td>
</tr>
<tr>
<td>GDP Inflation</td>
<td>-0.13</td>
<td>-0.10</td>
<td>-0.24</td>
<td>-0.02</td>
<td>1.00</td>
<td>0.72</td>
<td>0.36</td>
</tr>
<tr>
<td>Federal Funds Rate</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.20</td>
<td>-0.08</td>
<td>0.72</td>
<td>1.00</td>
<td>0.37</td>
</tr>
<tr>
<td>Real Government Debt</td>
<td>0.02</td>
<td>-0.05</td>
<td>-0.01</td>
<td>-0.05</td>
<td>0.36</td>
<td>0.37</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>With Lag Of:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real Government Spending</td>
<td>0.06</td>
<td>-0.08</td>
<td>-0.06</td>
<td>0.02</td>
<td>-0.13</td>
<td>-0.06</td>
<td>-0.07</td>
</tr>
<tr>
<td>Real Tax Receipts</td>
<td>0.03</td>
<td>-0.04</td>
<td>0.15</td>
<td>0.03</td>
<td>-0.07</td>
<td>0.02</td>
<td>-0.10</td>
</tr>
<tr>
<td>Real GDP</td>
<td>0.03</td>
<td>0.25</td>
<td>0.27</td>
<td>-0.04</td>
<td>-0.22</td>
<td>-0.11</td>
<td>-0.04</td>
</tr>
<tr>
<td>Ramey’s News Shocks</td>
<td>-0.03</td>
<td>-0.00</td>
<td>-0.15</td>
<td>0.01</td>
<td>-0.01</td>
<td>-0.10</td>
<td>-0.06</td>
</tr>
<tr>
<td>GDP Inflation</td>
<td>-0.12</td>
<td>-0.11</td>
<td>-0.24</td>
<td>-0.00</td>
<td>0.99</td>
<td>0.71</td>
<td>0.41</td>
</tr>
<tr>
<td>Federal Funds Rate</td>
<td>-0.03</td>
<td>-0.14</td>
<td>-0.29</td>
<td>-0.10</td>
<td>0.71</td>
<td>0.96</td>
<td>0.41</td>
</tr>
<tr>
<td>Real Government Debt</td>
<td>-0.03</td>
<td>0.07</td>
<td>0.07</td>
<td>-0.06</td>
<td>0.31</td>
<td>0.36</td>
<td>0.69</td>
</tr>
</tbody>
</table>

Note: We model the real economy using the seven variables shown as our vector of endogenous variables, $X_t$. Shown above are $X_t$’s correlation matrix and the correlation between $X_{t-1}$ and $X_t$. Even for the covariance stationary data (with some variables in log difference) the correlation is somewhat high. Autocorrelation, shown on the diagonal of the lower matrix, is also high in several cases. As such, the degree of collinearity present in the ST-VAR necessitates Bayesian estimation methods.
Table B.3: Correlation of Variables, 2008Q4-2015Q4

<table>
<thead>
<tr>
<th></th>
<th>Real Government Spending</th>
<th>Real Tax Receipts</th>
<th>Real GDP</th>
<th>Ramey’s News Shocks</th>
<th>GDP Inflation</th>
<th>Federal Funds Rate</th>
<th>Real Government Debt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation With:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real Government Spending</td>
<td>1.00</td>
<td>-0.38</td>
<td>-0.21</td>
<td>0.34</td>
<td>-0.36</td>
<td>0.40</td>
<td>0.26</td>
</tr>
<tr>
<td>Real Tax Receipts</td>
<td>-0.38</td>
<td>1.00</td>
<td>0.47</td>
<td>-0.14</td>
<td>-0.14</td>
<td>-0.35</td>
<td>-0.30</td>
</tr>
<tr>
<td>Real GDP</td>
<td>-0.21</td>
<td>0.47</td>
<td>1.00</td>
<td>-0.14</td>
<td>-0.15</td>
<td>-0.59</td>
<td>-0.57</td>
</tr>
<tr>
<td>Ramey’s News Shocks</td>
<td>0.34</td>
<td>-0.14</td>
<td>0.16</td>
<td>1.00</td>
<td>-0.30</td>
<td>0.00</td>
<td>-0.22</td>
</tr>
<tr>
<td>GDP Inflation</td>
<td>-0.36</td>
<td>-0.14</td>
<td>-0.15</td>
<td>-0.30</td>
<td>1.00</td>
<td>-0.08</td>
<td>-0.10</td>
</tr>
<tr>
<td>Federal Funds Rate</td>
<td>0.40</td>
<td>-0.35</td>
<td>-0.59</td>
<td>0.00</td>
<td>-0.08</td>
<td>1.00</td>
<td>0.56</td>
</tr>
<tr>
<td>Real Government Debt</td>
<td>0.26</td>
<td>-0.30</td>
<td>-0.57</td>
<td>-0.22</td>
<td>-0.10</td>
<td>0.56</td>
<td>1.00</td>
</tr>
<tr>
<td>With Lag Of:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real Government Spending</td>
<td>0.65</td>
<td>-0.36</td>
<td>-0.34</td>
<td>0.25</td>
<td>-0.54</td>
<td>0.46</td>
<td>0.35</td>
</tr>
<tr>
<td>Real Tax Receipts</td>
<td>-0.41</td>
<td>0.17</td>
<td>0.26</td>
<td>-0.00</td>
<td>0.06</td>
<td>-0.34</td>
<td>-0.39</td>
</tr>
<tr>
<td>Real GDP</td>
<td>-0.43</td>
<td>0.57</td>
<td>0.42</td>
<td>0.06</td>
<td>0.02</td>
<td>-0.51</td>
<td>-0.33</td>
</tr>
<tr>
<td>Ramey’s News Shocks</td>
<td>-0.03</td>
<td>-0.08</td>
<td>0.08</td>
<td>-0.08</td>
<td>-0.17</td>
<td>-0.00</td>
<td>0.10</td>
</tr>
<tr>
<td>GDP Inflation</td>
<td>-0.12</td>
<td>-0.22</td>
<td>-0.19</td>
<td>-0.29</td>
<td>0.84</td>
<td>-0.14</td>
<td>-0.14</td>
</tr>
<tr>
<td>Federal Funds Rate</td>
<td>0.38</td>
<td>-0.40</td>
<td>-0.57</td>
<td>-0.06</td>
<td>-0.09</td>
<td>0.97</td>
<td>0.68</td>
</tr>
<tr>
<td>Real Government Debt</td>
<td>0.23</td>
<td>-0.37</td>
<td>-0.44</td>
<td>-0.02</td>
<td>-0.17</td>
<td>0.47</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Note: Similar to the previous table, except this table covers the period from 2008Q4 to 2015Q4. The correlation structure between some variables is different during this period. For example, the high level of correlation between real government spending and real tax receipts, GDP inflation, and the Federal Funds rate was not previously present. We would remark that the correlation between the Federal Funds Rate and Real Government Debt is even higher than in the previous table.
The Minnesota prior assumes that $X_t$ follows a multivariate random walk process. $\pi_0$ is a vector equal to zero, except for the elements that correspond to a variable’s own lags. In this way the Minnesota prior shrinks many parameters of $\pi_r$ towards zero. In addition, $V_0$ has the following form:

$$V(i,j)_0 = \begin{cases} 
\frac{\theta_1}{\ell^2} & \text{for parameter on own lags, } j = i \\
\frac{\theta_2 \sigma_{s,i}^2}{\ell^2 \sigma_{s,j}^2} & \text{for parameter on foreign lags, } j \neq i
\end{cases}$$ (B.32)

where $\ell$ is the lag length and $\theta_1$ and $\theta_2$ are two hyperparameters. The Minnesota prior is also a conjugate prior, which means that the posterior distribution has the same form as the likelihood model for the data. In that case, the posterior of $\pi_r$ is also normal, i.e.,

$$p(\pi_r|X^T, \Omega_r, \gamma, c) \sim N(m_r, V_r)$$ (B.33)

where

$$V_r = \left(\left(X_{r,t-1}'X_{r,t-1} \otimes \Omega_r^{-1}\right) + V_0^{-1}\right)^{-1}$$ (B.34)

$$m_r = V_r \times \left(\left(X_{r,t-1}'X_{r,t-1} \otimes \Omega_r^{-1}\right)\hat{\pi}_r + V_0^{-1}\pi_0\right).$$ (B.35)

Minnesota priors for nonstationary data traditionally choose $\pi_0$ such that the first-own-lag of each variable has a coefficient of one; the rest are zero. Our first-differenced data are covariance stationary, and so we choose a covariance stationary prior. However, we believe a priori that our seven variables, even those that are first differenced, are characterized by a high degree of persistence.

Tables B.2 and B.3 contain correlation matrices for our seven variables, together with the matrix of those seven variables’ correlation with their own first lags. The first of these tables shows correlation for the era before the zero lower bound, while the latter table shows correlations during the zero lower bound. Four of the seven variables have first autocorrelations exceeding .6 in one or both samples. Because sample autocorrelations are well-known to be biased downwards when the true population correlation is near unity
(Diebold and Kilian [2000] call this phenomenon *Dickey-Fuller Bias*) we choose $\pi_0$ such that each variable is highly persistent AR(1), and set the first-own-lag coefficients to be .9.

**Sample $\gamma$**

The posterior distribution of $\gamma$ does not have a closed-form expression. Hence, we also rely on MCMC methods to approximate the marginal posterior distribution of $\gamma$. We follow Gefang and Strachan [2010] and Galvão and Owyang [2018] and use the Metropolis-Hasting algorithm to draw $\gamma$. Denote $\gamma^*$ as a candidate draw from the proposal density $q(\gamma^*|\gamma^{k-1})$ where $\gamma^{k-1}$ is the draw from the previous iteration. Then, we set $\gamma^k = \gamma^*$ with probability

$$
\alpha(\gamma^*|\gamma^{k-1}) = \min\left\{\frac{p(\gamma^*|\cdot)}{q(\gamma^*|\gamma^{k-1}|\gamma^*)}/\frac{p(\gamma^{k-1}|\cdot)}{q(\gamma^{k-1}|\gamma^*)}, 1\right\} \tag{B.36}
$$

$p(\gamma^*|\cdot)$ is the posterior distribution of $\gamma^*$ while $p(\gamma^{k-1}|\cdot)$ is the posterior distribution of $\gamma^{k-1}$. Moreover, $q(\gamma^*|\gamma^{k-1})$ and $q(\gamma^{k-1}|\gamma^*)$ are the corresponding proposal densities. The proposal density is a gamma distribution. Under mild regularity conditions, the sequence of $\gamma^k$, after a sufficient burn-in phase, will approximate the true marginal posterior distribution of $\gamma$ (see Chernozhukov and Hong [2003]).

**Sample $c$**

The threshold parameter $c$ does not have a closed form for its posterior distribution. Again, we follow Gefang and Strachan [2010] and use the Griddy-Gibbs sampler proposed by Ritter and Tanner [1992] to draw $c$, where $c$ is bounded between the minimum and maximum value of $z_t$. First, we select grid points to evaluate the density of $z_t$ at these grid points and approximate the corresponding CDF. Second, we draw a random variable from a standard uniform distribution and plug it into the inverse CDF of $z_t$. This yields a draw from the marginal posterior distribution of $c$.\textsuperscript{1617}

---

\textsuperscript{16}This property is known as the inversion method. Suppose $X$ is a random variable with an unknown distribution and $u$ is standard uniformly distributed. Then $F^{-1}(u)$ can be used to generated draws of $X$ with the specified CDF $F$.

\textsuperscript{17}Galvão and Owyang [2018] draw $\gamma$ and $c$ jointly using a single Metropolis-Hastings step, as previously described. Both methods (theirs and ours) are equally reasonable. After experimenting with $\gamma$ and $c$ drawn jointly, we did not observe a noticeable change in the speed of the sampler.
Sample $\Omega_E$ and $\Omega_R$

We assume that the purely AM and PM regimes are characterized not only by their own coefficient matrix $\Pi_{AM}$ and $\Pi_{PM},$ but also by their own covariance matrix $\Omega_{AM}$ and $\Omega_{PM}.$ Under the assumption of heteroskedasticity, $\Omega_{AM}$ and $\Omega_{PM}$ are not conjugate. In addition, the joint posterior distribution of $\Omega_{AM}$ and $\Omega_{PM}$ cannot be written as the product of their marginal posterior distributions. Hence, they are also not independent and must be drawn jointly. We follow Galvão and Owyang [2018] and use the Metropolis-Hastings algorithm proposed by Chib and Greenberg [1995] to sample $\Omega_{AM}$ and $\Omega_{PM}.$

Denote $\Omega^{*}_{AM}$ and $\Omega^{*}_{PM}$ as a pair of candidate draws from the proposal densities $q(\Omega^{*}_{AM}|\Omega_{k-1}^{AM},\Omega_{k-1}^{PM})$ and $q(\Omega^{*}_{PM}|\Omega_{k-1}^{AM},\Omega_{k-1}^{PM}),$ respectively. $\Omega^{k-1}_{AM}$ and $\Omega^{k-1}_{PM}$ are draws from the previous iteration. Then, we set $\{\Omega_{AM}^{k}, \Omega_{PM}^{k}\} = \{\Omega^{*}_{AM}, \Omega^{*}_{PM}\}$ with probability

$$\alpha(\Omega^{*}_{AM},\Omega^{*}_{PM}|\Omega_{AM}^{k-1},\Omega_{PM}^{k-1}) = \min\left\{ \frac{p(\Omega^{*}_{E},\Omega^{*}_{R}|\cdot)/q(\Omega^{*}_{E},\Omega^{*}_{R}|\cdot)}{p(\Omega_{E}^{k-1},\Omega_{R}^{k-1}|\cdot)/q(\Omega_{E}^{k-1},\Omega_{R}^{k-1}|\cdot)}, 1 \right\} (B.37)$$

$p(\Omega^{*}_{AM},\Omega^{*}_{PM}|\cdot)$ and $p(\Omega^{k-1}_{AM},\Omega^{k-1}_{PM}|\cdot)$ are the corresponding posterior distributions of $\Omega^{*}_{AM},\Omega^{*}_{PM}$ and $\Omega^{k-1}_{AM},\Omega^{k-1}_{PM}.$ Moreover, $q(\Omega^{*}_{AM},\Omega^{*}_{PM}|\cdot)$ and $q(\Omega^{k-1}_{AM},\Omega^{k-1}_{PM}|\cdot)$ are their respective proposal densities. The proposal density for both $\Omega_{AM}$ and $\Omega_{PM}$ is an inverse-Wishart distribution. Under relative mild regularity conditions, the sequences of $\{\Omega_{AM}^{k}, \Omega_{PM}^{k}\},$ after a sufficient burn-in phase, will approximate the true posterior distribution of $\Omega_{AM}$ and $\Omega_{PM}$ (see Chernozhukov and Hong [2003]).

### B.1.3 Generalized Impulse Response Functions

In nonlinear models, the impulse response parameters depend on the shock sign, the shock size and the timing of the shock. Furthermore, the state of the economy can change over time, and may change differently after a shock. To incorporate these features, we follow Koop et al. [1996] and use generalized impulse response functions to estimate the dynamic effects of the shock. The generalized impulse responses defined as the difference of two simulated paths of the economies, i.e.,
\[
GIRF(h) = (1 - G(z_{t+h-1})) \Pi_{AM} X_{t+h-1} + G(z_{t+h-1}) \Pi_{PM} X_{t+h-1} + u'_{t+h}
- (1 - G(z_{t+h-1})) \Pi_{AM} X_{t+h-1} + G(z_{t+h-1}) \Pi_{PM} X_{t+h-1} + u_{t+h}
\]  
(B.38)

First, we draw a set of reduced-form parameters \(\Omega_{AM}, \Omega_{PM}, \Pi_{AM}\) and \(\Pi_{PM}\). Second, the generalized impulse response functions require an initial condition. Hence, we draw an initial condition that comes with the lags for the VAR \(X_{t-1}\), the value of the state variable \(z_{t-1}\) and a sequence of the reduced-form residuals of \(H\) periods \(u^H_t\). Third, we transform the sequence of reduced-form residuals into structural shocks and “back”

\[
\epsilon_{t+h} = (Chol(\Omega_t)Q)^{-1} u_{t+h} \tag{B.39}
\]

\[
\hat{\epsilon}_{t+h} = \epsilon_{t+h} + \delta \tag{B.40}
\]

\[
u^*_{t+h} = Chol(\Omega_t)Q\hat{\epsilon}_{t+h} \tag{B.41}
\]

where \(\delta\) represents the shock size, \(Chol(\Omega_t)\) is the Cholesky decomposition of \(\Omega_t\) and \(Q\) is an orthogonal matrix, called rotation matrix. We follow Rubio-Ramirez et al. [2010] and draw \(Q\) from the Haar measure via the QR decomposition of a matrix of standard normal random variables. Fourth, conditional on the starting period, we roll the model forward using Equation (B.1) of our main model and add the value of \(u_{t+h}\) to the path of the economy without the government spending shock and \(u^*_{t+h}\) to the path of the economy with the government spending shock. Fifth, we take the difference between the simulated economies to obtain one candidate draw of the generalized impulse response functions. If the candidate satisfies the sign restrictions, we store the draw. If not, we discard the candidate, draw a new rotation matrix and check again if the imposed sign restrictions are satisfied. We repeat this procedure until a sufficient number of candidates are accepted and move to the next initial condition. Then, we average over the accepted candidates and initial conditions to obtain one realization of the generalized impulse response functions. Finally, we repeat
the whole procedure for the next draw of the reduced form parameters, \( \Omega_{AM}, \Omega_{PM}, \Pi_{AM} \) and \( \Pi_{PM} \). This procedure yields a distribution of generalized impulse response functions that are consistent with our sign restrictions.

**B.2 Taylor Rules**

This section discusses some of the alternative methods available to estimate the Taylor rule. The inflation parameter of the Taylor Rule is used in our main model to distinguish between different monetary policy regimes, and serves as a foundation of the rest of the analysis. Importantly, our measure of monetary policy activism \( G(z_{t-1}) \) is robust with regard to possible Taylor rule specifications.

There is a healthy literature regarding the estimation of Taylor rules. Taylor [1999] estimates a simple Taylor rule by regressing the federal funds rate as the monetary policy instrument on real output and the inflation rate using least squares. Clarida et al. [2000] also estimates a simple Taylor rule. However, they exploit a large set of instruments and employ the general method of moments. Both studies find that monetary policy changed considerably before and after 1980. Orphanides [2001, 2002, 2004] and Boivin [2006] choose a different strategy, and suggest real-time data should be used to analyze policymakers’ decisions because revised data may contain information that was unavailable to policy makers during the time they faced policy decisions. These studies also find that monetary policy may also have been active in the 1970s, which contradicts the conclusion drawn by Taylor [1999] and Clarida et al. [2000].

Moreover, Cogley and Sargent [2001] find that monetary policy has varied substantially over time as well. Sims [2001] argues that the change in monetary policy is due to time-varying heteroskedasticity in the Taylor-rule residuals. However, Cogley and Sargent [2005] demonstrate that their 2001-results are robust after accounting for stochastic volatility in their model.

de Vries and Li [2013] illustrate that the Taylor rule residuals are serially correlated. In this case, instruments such as lagged regressors, real-time data or the Clarida et al. [2000]’s
instruments are not necessarily valid instruments. Fortunately, the potential endogeneity problem does not cause a substantial bias. Finally, Carvalho et al. [2019] provide substantial evidence that the bias in a Taylor rule estimated with least squares is small.

One key question addressed by the literature is whether or not monetary policy has changed over time. According to the literature, the choice of data (revised vs. real-time data) and assumptions regarding the residuals (constant vs. stochastic volatility) might be important when estimating the Taylor rule. Recent evidence, however, suggests that differences in the estimators are small. In this section, we estimate the Taylor rule in four different specifications: (i) revised data with current inflation and real GDP growth rates (main specification used in the paper), (ii) revised data with lagged inflation and real GDP growth rates, (iii) real-time data with current inflation and real GDP growth rates, and (iv) real-time data with lagged inflation and real GDP growth rates. Figure B.2 presents the corresponding evolution of $G(z_{t-1})$ in each of these models.

Figure B.2: Evolution of Monetary Policy: Revised vs. Real-Time Data

Note: The Figure compares the estimated evolutions of $G(z_{t-1})$ using current and lagged regressors from revised and real-time data, respectively. The results indicate that if differences exist, they are small and appear in the pre-Volcker period.

Figure B.2 presents the evolution of $G(z_{t-1})$ rather than the evolution of the inflation
parameter because $G(z_{t-1})$ determines the dynamics of our model in the main text while the inflation parameter is just one ingredient in $G(z_{t-1})$. Figure B.2 shows that there exist differences between revised vs real-time data and between current and lagged regressors. However, these differences are small and mostly arise in the pre-Volcker era. For example, not all estimations “recognize” the Martin disinflation in 1959/60. However, all estimated evolutions capture the change in monetary policy around 1980, the responses to the 1957/58, 1990/91 and 2000/01 recessions, and the substantial deviation from the Taylor principle between 2002 and 2005 as pointed out by Taylor [2007].

Figure B.3: Evolution of Monetary Policy: Different Price Indices

![Figure B.3: Evolution of Monetary Policy: Different Price Indices](image)

Note: The Figure compares the estimated evolutions of $G(z_{t-1})$ using different price indices to compute inflation. The results indicate that if differences exist, they are small and appear mostly in the pre-Volcker period and in the second-half of the 1990s.

Another way to test the robustness of our results is to use different price indices to compute inflation. In Figure B.3, we show that $G(z_{t-1})$ displays a similar path regardless of whether we compute inflation based on (i) the GDP deflator, (ii) CPI or (iii) PCE. As in Figure B.2, if differences exist, they appear in the pre-Volcker era. For example, the Martin disinflation is not as pronounced for CPI or PCE as it is for the GDP deflator. In contrast, all three specifications display the changes in monetary policy in 1957/58, around
Figure B.4: Evolution of Monetary Policy: Comparison with the Clarida et al. [2000] Instruments

Note: The Figure compares the estimated evolution of $G(z_{t-1})$ from our main specification and that estimated with the Clarida et al. [2000] Instruments.


Because of concerns about reverse-causality, Clarida et al. [2000] estimate the Taylor rule coefficients using an instrumental variables procedure. Their instruments (for inflation and output) include four lags of inflation, the output gap, the federal funds rate, the short-long spread, and commodity price inflation. We consider a two-stage regression, where we first regress inflation and output on these instruments and use their fitted values in the Taylor rule. While this method could be advantageous in the presence of significant endogeneity, we find the variation in coefficients (and estimated regimes) is similar to the estimates from our primary model. Figure B.4 illustrates that differences only occur between 1995 and 2000.

A possible explanation for the similar estimates of $G(z_{t-1})$, regardless of whether we use (i) current or lagged regressors, (ii) revised or real-time data or (iii) different price indices to compute inflation, is serial correlation in the residuals of the different Taylor rule specifications. To investigate this feature of the residuals, we conduct the Ljung-Box Q-Test.
We use Bayesian methods to estimate the different Taylor rules. Hence, we can also compute the posterior distribution of the residuals. We then conduct the Ljung-Box Q-Test on each draw of the residuals. The null of the Ljung-Box Q-Test is that the residuals are not serially correlated. In Figure B.5, we show the distributions of p-values of this test with one, two and four lags for our baseline specification. Figure B.5 provides strong evidence for serial correlation in the residuals. For example, the share of p-values that is below 0.05 for the test with one lag is above 70 percent. This implies that more than 70 percent of the p-values are below 0.05. For the tests with two and four lags, the share of small p-values increases to above 90 percent. This test for our baseline specification provides strong evidence that the residuals are indeed serially correlated. According to de Vries and Li [2013], if the residuals of the Taylor rule are serially correlated, the estimation using standard instruments such as lagged regressors or the Clarida et al. [2000]’s instruments face endogeneity problems, but the resulting bias is small.

Figure B.5: Ljung-Box Q-Test

Note: The figure shows the distribution of p-values for a Ljung-Box Q-Test. The Figure provides evidence that the residuals of the estimated Taylor rule exhibit serial correlation. For example over 70 percent of the p-values for the Test with one lag is below 0.05. For the tests with two and four lags, the percentage is larger.
Finally, another way to check the robustness of our estimate of $G(z_{t-1})$ is the comparison with the real interest rates. Theorists sometimes consider the real interest rate when making their predictions about whether the government spending multiplier depends on monetary policy. In Figure B.6, we compare the estimate of $G(z_{t-1})$ from the main text with the real interest rate defined as the difference between the federal funds rate and annualized inflation based on the GDP deflator.

Figure B.6 documents a close relationship between our estimated $G(z_{t-1})$ and the real interest rate. When monetary policy was very passive e.g., during the late 1950s, second half of 1970s or the first half of the 2000s, real interest rates were relatively low. In contrast, when monetary policy was very active e.g., during the 1980s or the second half of the 1990s, real interest rates were relatively high. Figure B.6 displays further that changes in $G(z_{t-1})$ and changes in the real interest rates occur almost at the same time. For example, when the monetary policy regime changes and becomes more active, real interest rates increase and vice versa if the monetary policy regime becomes more passive.

Figure B.6: The Sample Paths of $G(z_{t-1})$ and the Real Interest Rate

Note: The red line shows the median estimate of $G(z_{t-1})$ along with the 68 percent credible bands in pink. The black line corresponds to the real interest rate. The Figures demonstrates that during times of passive monetary policy, the real interest rate is relatively low, and relatively high during times of active monetary policy.
B.3 Additional Exercises

B.3.1 Robustness Checks

In this section, we conduct robustness checks with respect to several aspects of our main specification. First, we redo our main exercise but replace our sign restriction approach with the Cholesky method to identify government spending shocks. This exercise is also shown by Figure 4 in the main text. Second, in our baseline exercise, we use a five-quarter moving average of the median estimate of $\phi_{\pi,t}$ as the state variable $z_t$. In this section, we (i) consider the median estimate of $\phi_{\pi,t}$ as $z_t$ without any smoothing, (ii) change the length of the moving average to three and sixth quarters, and (iii), use the five-quarter moving average of the pointwise posterior 5th and 95th percentiles of $\phi_{\pi,t}$, respectively, as $z_t$. Lastly, in the baseline exercise, we update $\phi_{\pi,t+h}$ using the Kalman filter based on forecasted values of the federal funds rate, inflation rate and the real GDP growth rate from the generalized impulse response functions. This choice is consistent with the initial estimation of $\phi_{\pi,t}$. In this section, we consider alternative updating rules. For example, we conduct the updating of $\phi_{\pi,t+h}$ via (i) an updating regression, (ii) include $\phi_{\pi,t}$ in $X_t$ and treat it as an observable variable, and (iii) update $\phi_{\pi,t}$ via a AR(1) process with an exogenous persistent parameter.

Table B.4 presents the pointwise posterior median estimates for different forecast horizons along with the 68-percent credible bands. In addition, Figure B.7 shows the distribution of the difference between multipliers from the initial very active and the initial very passive regimes after five years.

For our baseline specification, the pointwise posterior median estimate of the government spending multiplier decreases from 5.7 to 0.6 for the very active regime, and from 4.5 to 1.2 for the very passive regime. Even though the pointwise posterior median estimates suggest that there exists a difference in the estimated multipliers between the initial regimes, the credible bands highly overlap. Panel A of Figure B.7 shows that the distribution of the differences between the multipliers after 5 years is centered around zero. This suggests that the two multiplier distributions do not differ in any meaningful way. When we replace our sign restriction approach with the Cholesky method, the pointwise posterior
median estimates decrease in magnitude at all horizons and differ across regimes. However, the distribution of the differences between the multipliers does not suggest a systematic difference (Panel B of Figure B.7).

We now alter the length of the moving average. First, we use the pointwise median estimate of $\phi_{\pi,t}$ as $z_t$, setting the moving average to zero. In this specification, the pointwise posterior median estimates decrease over time and are quite similar after five years. Second, we change the length of the moving average to three and sixth quarters, respectively. Similarly to the main exercise, the pointwise posterior median estimates decrease over time. The difference after five years shrinks for the three-quarter moving average compared to our main exercise, and widens for the six-quarter moving average specification. However, the distribution of the difference between multipliers does not detect a systematic difference (Panel D and E). Next, we consider different moments of the posterior distribution of $\phi_{\pi,t}$. Instead of the pointwise posterior median, we use the five-quarter moving average of the pointwise posterior $5^{th}$ and $95^{th}$ percentile of $\phi_{\pi,t}$. Table B.4 and Panel F and G of Figure B.7 show results similar to our main specification.

Finally, we consider different updating rules for $\phi_{\pi,t}$ after the government spending shock. In the baseline exercise, we update $\phi_{\pi,t+h}$ as $z_{t+h}$ using the Kalman filter based on forecasted observations of the federal funds rate, inflation rate and the real GDP growth rate from the generalized impulse response functions. First, we do not update $\phi_{\pi,t+h}$ at all and keep the monetary policy regimes purely active and purely passive for the entire time after the shock. In this specification, the median estimates are similar in the short run, decrease over time but diverge in the long run. Panel H of Figure B.4 shows that the distribution of the difference between the multipliers from the initial very active and very passive regimes after five years is skewed to the left. This result suggests that the government spending multiplier is larger if monetary policy is and remains purely passive after the shock. Second, we update $\phi_{\pi,t+h}$ after the shock using a regression based on the forecasted values of the federal funds rate, inflation rate and output growth rate, i.e.

$$
\hat{i}_{t+h} = \hat{c}_{t+h} + \hat{\phi}_{\pi,t+h} \hat{\pi}_{t+h} + \hat{\phi}_{\gamma,t+h} \hat{\gamma}_{t+h} + \hat{\epsilon}_{mp,t}.
$$

(B.42)
Then, we use the point estimate of $\hat{\phi}_{t+h}$ as $z_{t+h}$ to update the weights $G(z_{t+h})$ and $1 - G(z_{t+h})$. For this specification, the government spending multiplier decreases in magnitude over time but is similar in magnitude at each horizon across initial regimes. Next, we include $\phi_{t}$ in $X_t$, and forecast $\phi_{t+h}$ directly via the generalized impulse response functions. This exercise shows diverging pointwise posterior median estimates of the multiplier across initial regimes, but the distribution of the differences after five years is centered around zero (Panel J). Lastly, we update $\phi_{t}$ via an AR(1) process

$$\phi_{t+h} = \rho \phi_{t+h-1}. \quad (B.43)$$

where we set the persistence parameter $\rho$ exogenously to 0.95. In this case, the central bank updates its monetary policy regime after the government spending shock but in a very sluggish way. The pointwise posterior median estimates decrease and diverge over time. Similarly to the exercise with the constant regimes, the distribution of the differences after five years is skewed to the left.

Our robustness checks underline our main conclusion: the government spending multiplier does not depend on the initial monetary policy regime because the monetary policy regime itself responds quickly to the economic conditions after the government spending shock. This result holds regardless of whether we consider (i) a different identification strategy, (ii) use different lengths of the moving average of $\phi_{t}$ as $z_{t}$, (iii) consider different moments of the posterior distribution of $\phi_{t}$ as $z_{t}$, or (iv) different updating rules for $\phi_{t+h}$ after the government spending shock. The only times we find differences in the estimated multipliers across initial regimes is if we force the monetary policy regimes to remain artificially different for a long period of time, such as in Panels H and K of Figure B.4. However, these counterfactuals are difficult to reconcile with the data that suggest that the central bank responds quickly after a government spending shock.
B.3.2 Consequences of the Binary Interpretation of Monetary Policy

Monetary policy activism is not a binary phenomenon. Here, we examine the consequences of considering only the two halves of the policy spectrum. We repeat our main exercise but only distinguish between “active” and “passive” monetary policy. Using the generalized impulse response functions, we employ a threshold value: If \( G(\Delta t_{t-1}) \geq 0.5 \), then the monetary policy regime in \( t \) is declared as “active” and as “passive” otherwise. We then estimate the generalized impulse response function for each regime, drawing random initial conditions from the two subsets.

Figure B.8 shows the response of the monetary policy regime to the government spending shock. If monetary policy is initially “active”, the central bank changes its policy regime only
Figure B.8: Response of Monetary Policy Regime when Monetary Policy is Binary

Note: The figure shows the pointwise-posterior median evolution of $G(z)$, along with the 68 percent credible bands, when monetary policy is interpreted as binary: just active (red) and just passive (blue). The figure suggests that shortly after the shock and regardless of its initial condition, the central bank responds actively to inflation. Slightly. In contrast, if the initial regime is “passive”, the central bank responds quickly and transitions fast to the “active” regime. Shortly after the shock, the central bank responds aggressively to inflation regardless of the initial regime. Figure B.9 illustrates that the government spending multiplier does not depend on the initial monetary policy regime – either in the short run or in the long run. These results mirror our main findings.

In the main text, we also find that monetary policy is not just active or passive but differences exist even within these categories in terms of how active or passive the monetary policy regime may be. In Figure B.10, we demonstrate the consequences of the binary interpretation of monetary policy when the regime is in fact continuous. Figure B.10 plots the distribution of $G(z_{t-1})$ to display the distribution of initial monetary policy regimes. There are two spikes close to zero and one. However, there are also many observations in the interior of the unit interval, and many of those are close to the threshold value. We seriously doubt that these observations are characteristic of the ends of the spectrum—or vice-versa. In
Figure B.9: Multiplier Estimates when Monetary Policy is Responsive but Binary

Note: The figure displays the multiplier estimates for the scenario when the monetary policy regime is responsive but binary: just active (red) or just passive (blue). Regardless of the forecast horizon after the shock, there exists little evidence that the government spending multiplier depends on monetary policy.

this thresholded GIRF comparison, on average, we compare initial monetary policy regimes of 0:16 and 0:86 to each other. These values are far from the monetary policy regimes of zero and one, and both regimes incorporate many observations in the neighborhood of .5—possibly obfuscating the difference in multipliers at the ends of the spectrum. For example, Caggiano et al. [2015] and Ramey and Zubairy [2018] compare spending multipliers relying on the binary interpretation of the business cycle. Both studies conclude that there are no significant differences in the estimated multipliers. However, results thus obtained are difficult to interpret because they could come from two sources: (i) multipliers that are indeed regime-independent; or (ii) initial “regimes” that are not sufficiently different.

B.3.3 A Narrative History of Monetary Policy

In this Section, we provide a narrative overview of the history of U.S. monetary policy since 1954. The key goal is to develop a better understanding of why monetary policy has
Figure B.10: Distribution of Initial Monetary Policy Regimes

Note: Figure shows the distribution of the initial “passive” monetary policy regimes (blue) and the initial “active” monetary policy regimes (red). Bisecting the sample into “active” and “passive” regimes based on the initial value of $G(z)$ lumps many of the central observations into one regime or the other. However, many initial monetary policy regime are close to the threshold value of 0.5 and the “distinct” regimes are relatively close to each other.

changed over time and to support our empirical estimation of monetary policy in Figure 3 of the main text. The section follows closely Romer and Romer [2004] and Romer and Romer [2013]. Those authors analyze government reports such as Minutes and Transcripts of Federal Open Market Committee, Annual Reports of the Board of Governors of the Federal Reserve System and the Congressional testimony of the Fed chairman. They conclude that changes in monetary policy mostly occurred because the chairman of the Fed and/or other FOMC members changed their beliefs about whether inflation has negative long-run consequences, their estimate of the natural rate of unemployment and/or the responsiveness of inflation to economic slack. In Figure B.11, the narrative points align with our estimated history of monetary policy.
• William McChesney Martin Jr. was the chairman of the Fed between April 1951 and January 1970. Romer and Romer [2004] describe Martin as having similar beliefs about monetary policy as Paul Volcker and Alan Greenspan. Martin believed that high inflation was harmful in the long run and that monetary policy should respond to bad economic conditions.

• The Fed lowered nominal interest rates substantially in response to the 1953/54 and to the 1957/58 recessions. In Figure B.11, our measure of monetary policy activism takes a value close to zero at the beginning of our sample, indicating that monetary policy has been very passive during the late 1950s.

• Following the 1957/58 recession, inflation started to rise. The Fed responded via extreme monetary tightening. Romer and Romer [2004] suggest calling this period the Martin Disinflation. \( G(z_{t-1}) \) in Figure B.11 increases substantially in 1959.

• In the 1960s, the Fed changed its views and adopted “New Economics”. In particular, the Fed believed that there was a long-run trade-off between inflation and unemployment and that inflation would be low even for low levels of unemployment. Consequently, the Fed loosened its policy and did not tighten monetary policy during the second half of the 1960s even when output growth was high and inflation was increasing.

• At the end of Martin's tenure, the Fed returned to the natural rate framework but kept its optimistic view about the economy. In late 1968, the Fed tightened monetary policy substantially to combat increasing inflation. In 1968, \( G(z_{t-1}) \) increases significantly.

• Arthur Burns followed William McChesney Martin Jr. as the chairman of the Fed in February 1970. Initially, he continued Martin’s beliefs about monetary policy but he later became pessimistic about the responsiveness of inflation to economic slack. Towards the end of his term, Burns would return to his initial stance.

• During Burns’ first year, the Fed lowered nominal interest rates substantially. Their optimistic estimate of the natural rate of unemployment led the Fed to assume that
there was significant slack in the economy. $G(z_{t-1})$ in Figure B.11, displays this change as well: the monetary policy regime became more passive.

- When inflation rates did not decrease as expected, the Fed changed its views and became overly pessimistic about the responsiveness of inflation to economic slack. According to Fed beliefs at that time, raising interest rates to cause economic slack would have been inefficient. The Nixon administration imposed price controls in August 1971, which temporarily lowered inflation.

- Price controls were removed in January 1973 and inflation started to rise again. In addition, the Fed pursued expansionary monetary policy.

- In the mid-1970s, the Fed was pessimistic about the responsiveness of inflation to economic slack. In 1974, the Fed conducted contractionary monetary policy even though output was already decreasing. The Fed wanted to reduce inflation and was willing to accept a recession. Our estimate of monetary policy displays this change as well: $G(z_{t-1})$ increases in the beginning of 1974.

- Due to increasing unemployment rates, the Fed lowered nominal interest rates in the winter of 1974/75. Monetary policy remained expansionary until the end of Burns’ tenure in February 1978. Our model replicates this change: the value of $G(z_{t-1})$ decreases after the winter of 1974/75 and remains low until 1979.

- William Miller would follow Arthur Burns to become the Fed’s chairman between March 1978 and September 1979. The Fed became more optimistic about the natural rate of unemployment and believed monetary policy is ineffective to combat inflation: even though the Fed worried about high inflation during the late-1970s, the Fed’s views on the responsiveness of inflation to slack and their high estimate of the natural rate precluded the Fed from tightening monetary policy.

- Paul Volcker became the chairman of the Fed in August 1979. The views about monetary policy changed fundamentally and remained the same during his chairmanship and under his successor Alan Greenspan: (i) high inflation has strong
negative long-run consequences and little benefits, (ii) inflation responds to economic slack and (iii) the natural rate was even higher than previously assumed.

• In response to high inflation in the late 1970s, the Volcker Fed tightened monetary policy substantially which is believed to have caused the 1981/82 recession. Because of their high estimate of the natural rate, the Fed allowed a very high rate of unemployment in order to reduce inflation. According to our model, this change is the largest during our sample period. $G(z_{t-1})$ increases from almost zero to almost one during a period of three years (1979 - 1982).

• Alan Greenspan replaced Paul Volcker as chairman of the Fed in August 1987 and served until January 2006.

• In response to the 1990/91 recession, the Fed lowered nominal interest rates. Our estimate of $G(z_{t-1})$ changes from one to almost zero.

• A puzzle in the second half of the 1990s: Romer and Romer [2004] report that during the second half of the 1990s, inflation did not rise despite a long-lasting expansion. Greenspan argued that the economy became more competitive. Firms would rather cut costs than increase prices. In addition, Romer and Romer [2004] say that the Fed left real interest rates unchanged. According to our estimates, however, the Fed changed its responsiveness to inflation substantially. $G(z_{t-1})$ moves from almost zero to one and remains very active until 2001.

• The Fed loosened monetary policy in response to the 2000/01 recession. We estimate that the policy regime changed from being very active to be very passive.

• Taylor [2007] argued that the Fed deviated substantially from the Taylor Principle between 2002 and 2005. He concludes that this deviation contributed enormously to the housing bubble, which led to the Financial Crisis in 2008. Our model replicates this behavior as well. Between 2002 and 2005, our estimate of $G(z)$ reaches its lowest level for the sample before the 2008 Financial crisis.
• In June 2004, the Fed raised the target rate for the federal funds rate for the first time since the 2000/01 recession. The Fed believed that monetary policy remained accommodative. The Fed kept raising the target rate for the federal funds rate to 5.25 in June 2006. However, the Fed stopped announcing that monetary policy is accommodative in August 2005. Our estimate of $G(z)$ leaves its high level at one at roughly the same time.

• In May 2005, the Fed first expressed concerns about a slowdown in economic growth, in part, because of a “gradual cooling” of the housing market.

• Ben Bernanke followed Alan Greenspan in February 2006.

• In September 2007, the Fed announced to reduce its target rate for the federal funds rate for the first time since its response to the 2000/01 recession.

• Zero Lower Bound: The Fed kept lowering nominal interest rates until the Fed finally cut them to zero in December 2008. Nominal interest rates remained at zero until December 2015.

• Janet Yellen replaced Ben Bernanke as the Fed’s chairman in February 2014.

• Liftoff: In December 2015, the Fed raised the target range for the federal funds rate to 1/4 to 1/2 percent.
Figure B.11: Evolution of Monetary Policy between 1954 and 2016

Note: Figure shows the pointwise-posterior median estimate of $G(z)$, along with the 68 percent credible bands. $G(z)$ can be interpreted as a measure for monetary policy activism. Grey bars represent recessions as defined by the National Bureau of Economic Research (NBER). The Figure combines our estimated evolution of monetary policy with narrative evidence given by Romer and Romer [2004], and the author’s reading of the FOMC press releases.
Table B.4: Robustness Checks

<table>
<thead>
<tr>
<th></th>
<th>1 Quarter</th>
<th>1 Year</th>
<th>2 Years</th>
<th>3 Years</th>
<th>4 Years</th>
<th>5 Years</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline Specification: Very Active</strong></td>
<td>[5.7]</td>
<td>[4.2]</td>
<td>[2.1]</td>
<td>[1.2]</td>
<td>[0.8]</td>
<td>[0.6]</td>
</tr>
<tr>
<td></td>
<td>[4.1, 8.2]</td>
<td>[3.0, 5.7]</td>
<td>[0.8, 3.8]</td>
<td>[0-6.3,0]</td>
<td>[-1.2,3.0]</td>
<td>[-1.6,2.6]</td>
</tr>
<tr>
<td><strong>Baseline Specification: Very Passive</strong></td>
<td>[4.5]</td>
<td>[3.6]</td>
<td>[2.4]</td>
<td>[1.6]</td>
<td>[1.4]</td>
<td>[1.2]</td>
</tr>
<tr>
<td></td>
<td>[3.0, 7.1]</td>
<td>[2.3, 5.5]</td>
<td>[0.7, 5.6]</td>
<td>[-0.7,5.3]</td>
<td>[-1.3,4.9]</td>
<td>[-1.7,4.8]</td>
</tr>
<tr>
<td><strong>Recursive Identification: Very Active</strong></td>
<td>[0.6]</td>
<td>[-0.4]</td>
<td>[-1.2]</td>
<td>[-0.9]</td>
<td>[-0.5]</td>
<td>[0.2]</td>
</tr>
<tr>
<td></td>
<td>[-0.3, 1.6]</td>
<td>[-2.7, 1.6]</td>
<td>[-5.7,2.4]</td>
<td>[-6.6,5.3]</td>
<td>[-5.9,5.8]</td>
<td>[-5.9,8.1]</td>
</tr>
<tr>
<td><strong>Recursive Identification: Very Passive</strong></td>
<td>[1.1]</td>
<td>[0.7]</td>
<td>[0.4]</td>
<td>[0.3]</td>
<td>[0.3]</td>
<td>[0.3]</td>
</tr>
<tr>
<td></td>
<td>[0.1, 1.9]</td>
<td>[-0.8, 1.8]</td>
<td>[-1.9,1.6]</td>
<td>[-2.3,2.3]</td>
<td>[-2.5,2.6]</td>
<td>[-2.6,2.7]</td>
</tr>
<tr>
<td>$\phi_{e,t}$ as $z_t$: Very Active</td>
<td>[5.2]</td>
<td>[4.5]</td>
<td>[3.4]</td>
<td>[2.4]</td>
<td>[1.8]</td>
<td>[1.8]</td>
</tr>
<tr>
<td></td>
<td>[3.6, 7.8]</td>
<td>[3.2, 6.5]</td>
<td>[1.3, 7.1]</td>
<td>[-0.8,6.6]</td>
<td>[-1.9,6.6]</td>
<td>[-2.8,6.3]</td>
</tr>
<tr>
<td>$\phi_{e,t}$ as $z_t$: Very Passive</td>
<td>[4.3]</td>
<td>[3.8]</td>
<td>[3.1]</td>
<td>[2.5]</td>
<td>[2.0]</td>
<td>[1.9]</td>
</tr>
<tr>
<td></td>
<td>[3.0, 6.2]</td>
<td>[2.7, 5.5]</td>
<td>[1.4, 5.9]</td>
<td>[-0.1,6.1]</td>
<td>[-1.1,5.9]</td>
<td>[-1.7,6.2]</td>
</tr>
<tr>
<td><strong>3 Quarter Moving Average as $z_t$: Very Active</strong></td>
<td>[5.7]</td>
<td>[4.1]</td>
<td>[2.3]</td>
<td>[1.6]</td>
<td>[1.3]</td>
<td>[1.2]</td>
</tr>
<tr>
<td></td>
<td>[3.8, 8.6]</td>
<td>[2.9, 5.9]</td>
<td>[0.8, 4.8]</td>
<td>[-0.8,4.3]</td>
<td>[-1.5,3.9]</td>
<td>[-1.6,7.4]</td>
</tr>
<tr>
<td><strong>3 Quarter Moving Average as $z_t$: Very Passive</strong></td>
<td>[4.0]</td>
<td>[3.3]</td>
<td>[2.4]</td>
<td>[1.8]</td>
<td>[1.6]</td>
<td>[1.4]</td>
</tr>
<tr>
<td></td>
<td>[2.7, 6.4]</td>
<td>[2.3, 5.6]</td>
<td>[0.6, 5.6]</td>
<td>[-1.1,5.7]</td>
<td>[-1.8,5.4]</td>
<td>[-2.1,5.4]</td>
</tr>
<tr>
<td><strong>6 Quarter Moving Average as $z_t$: Very Active</strong></td>
<td>[5.9]</td>
<td>[4.9]</td>
<td>[2.6]</td>
<td>[1.2]</td>
<td>[0.8]</td>
<td>[0.6]</td>
</tr>
<tr>
<td></td>
<td>[4.8, 7.3]</td>
<td>[4.6, 6.4]</td>
<td>[0.6, 5.3]</td>
<td>[-1.0,3.5]</td>
<td>[-1.5,3.3]</td>
<td>[-1.6,3.0]</td>
</tr>
<tr>
<td><strong>6 Quarter Moving Average as $z_t$: Very Passive</strong></td>
<td>[5.1]</td>
<td>[5.1]</td>
<td>[3.9]</td>
<td>[2.6]</td>
<td>[1.9]</td>
<td>[1.6]</td>
</tr>
<tr>
<td></td>
<td>[3.9, 6.8]</td>
<td>[3.0, 6.6]</td>
<td>[1.8, 7.1]</td>
<td>[-0.3,6.6]</td>
<td>[-1.5,5.3]</td>
<td>[-1.9,5.3]</td>
</tr>
<tr>
<td>$5^{th}$ percentile as $z_t$: Very Active</td>
<td>[5.4]</td>
<td>[4.2]</td>
<td>[2.2]</td>
<td>[1.3]</td>
<td>[1.0]</td>
<td>[0.8]</td>
</tr>
<tr>
<td></td>
<td>[3.9, 7.0]</td>
<td>[3.0, 5.7]</td>
<td>[0.7,4.0]</td>
<td>[-0.6,3.2]</td>
<td>[-1.1,3.1]</td>
<td>[-1.3,2.8]</td>
</tr>
<tr>
<td>$5^{th}$ percentile as $z_t$: Very Passive</td>
<td>[4.2]</td>
<td>[3.3]</td>
<td>[2.1]</td>
<td>[1.5]</td>
<td>[1.2]</td>
<td>[1.1]</td>
</tr>
<tr>
<td></td>
<td>[2.7, 6.7]</td>
<td>[2.2, 5.6]</td>
<td>[0.6, 4.7]</td>
<td>[-0.7,4.2]</td>
<td>[-1.1,3.9]</td>
<td>[-1.5,3.9]</td>
</tr>
<tr>
<td>$95^{th}$ percentile as $z_t$: Very Active</td>
<td>[5.9]</td>
<td>[4.5]</td>
<td>[2.6]</td>
<td>[1.6]</td>
<td>[1.2]</td>
<td>[1.0]</td>
</tr>
<tr>
<td></td>
<td>[4.1, 8.8]</td>
<td>[3.3, 6.3]</td>
<td>[1.3, 4.7]</td>
<td>[-0.6,4.2]</td>
<td>[-1.5,3.7]</td>
<td>[-1.9,3.7]</td>
</tr>
<tr>
<td>$95^{th}$ percentile as $z_t$: Very Passive</td>
<td>[4.8]</td>
<td>[4.0]</td>
<td>[3.1]</td>
<td>[2.3]</td>
<td>[1.8]</td>
<td>[1.5]</td>
</tr>
<tr>
<td></td>
<td>[3.8, 7.0]</td>
<td>[2.6, 6.1]</td>
<td>[0.7, 7.2]</td>
<td>[-1.1,7.3]</td>
<td>[-2.4,7.1]</td>
<td>[-3.0,7.2]</td>
</tr>
<tr>
<td>Constant Regimes: Purely Active</td>
<td>[5.2]</td>
<td>[4.8]</td>
<td>[2.9]</td>
<td>[1.8]</td>
<td>[1.3]</td>
<td>[1.1]</td>
</tr>
<tr>
<td></td>
<td>[2.7, 10.5]</td>
<td>[2.3, 9.9]</td>
<td>[0.9, 6.4]</td>
<td>[-0.1,4.3]</td>
<td>[-0.5,4.5]</td>
<td>[-0.8,3.1]</td>
</tr>
<tr>
<td>Constant Regimes: Purely Passive</td>
<td>[5.5]</td>
<td>[4.0]</td>
<td>[3.5]</td>
<td>[3.5]</td>
<td>[3.5]</td>
<td>[3.5]</td>
</tr>
<tr>
<td></td>
<td>[2.6, 11.7]</td>
<td>[2.1, 7.6]</td>
<td>[1.8, 7.2]</td>
<td>[1.5,7.4]</td>
<td>[1.5,7.7]</td>
<td>[4.8,8.8]</td>
</tr>
<tr>
<td><strong>Updating Regression: Very Active</strong></td>
<td>[5.1]</td>
<td>[4.5]</td>
<td>[2.8]</td>
<td>[2.0]</td>
<td>[1.6]</td>
<td>[1.4]</td>
</tr>
<tr>
<td></td>
<td>[3.6, 7.0]</td>
<td>[3.1, 6.3]</td>
<td>[1.2, 6.6]</td>
<td>[-1.3,6.2]</td>
<td>[-2.0,5.7]</td>
<td>[-2.5,5.6]</td>
</tr>
<tr>
<td><strong>Updating Regression: Very Passive</strong></td>
<td>[4.9]</td>
<td>[4.2]</td>
<td>[2.8]</td>
<td>[1.9]</td>
<td>[1.7]</td>
<td>[1.5]</td>
</tr>
<tr>
<td></td>
<td>[3.5, 7.1]</td>
<td>[2.9, 6.1]</td>
<td>[0.5, 6.4]</td>
<td>[-1.5,6.1]</td>
<td>[-1.8,6.4]</td>
<td>[-2.2,5.9]</td>
</tr>
<tr>
<td>$z_t$ in $X_t$: Very Active</td>
<td>[4.8]</td>
<td>[4.4]</td>
<td>[2.9]</td>
<td>[1.5]</td>
<td>[0.9]</td>
<td>[0.6]</td>
</tr>
<tr>
<td></td>
<td>[4.1, 5.9]</td>
<td>[3.6, 5.5]</td>
<td>[1.3, 5.7]</td>
<td>[-0.7,4.3]</td>
<td>[-1.7,3.7]</td>
<td>[-1.9,4.4]</td>
</tr>
<tr>
<td>$z_t$ in $X_t$: Very Passive</td>
<td>[4.5]</td>
<td>[4.7]</td>
<td>[3.5]</td>
<td>[2.0]</td>
<td>[1.5]</td>
<td>[1.3]</td>
</tr>
<tr>
<td></td>
<td>[3.6, 5.5]</td>
<td>[3.7, 5.9]</td>
<td>[-0.1,8.8]</td>
<td>[-2.5,6.7]</td>
<td>[-3.4,5.9]</td>
<td>[-2.2,5.5]</td>
</tr>
<tr>
<td>AR(1) process: Very Active</td>
<td>[5.7]</td>
<td>[4.2]</td>
<td>[2.1]</td>
<td>[1.2]</td>
<td>[0.8]</td>
<td>[0.7]</td>
</tr>
<tr>
<td></td>
<td>[2.7, 8.3]</td>
<td>[3.6, 7.5]</td>
<td>[1.0, 3.6]</td>
<td>[-0.3,2.8]</td>
<td>[-1.0,2.5]</td>
<td>[-1.2,2.4]</td>
</tr>
<tr>
<td>AR(1) process: Very Passive</td>
<td>[4.2]</td>
<td>[3.7]</td>
<td>[3.5]</td>
<td>[3.2]</td>
<td>[3.0]</td>
<td>[2.9]</td>
</tr>
<tr>
<td></td>
<td>[2.7, 6.4]</td>
<td>[2.5, 5.6]</td>
<td>[1.8, 6.3]</td>
<td>[-1.2,6.3]</td>
<td>[1.1,6.3]</td>
<td>[0.8,6.3]</td>
</tr>
</tbody>
</table>

Note: Table shows the pointwise posterior median estimates along with the 68-percent credible bands for our main specification and several robustness checks. Our robustness checks support our main conclusion that the multiplier does not depend on the initial monetary policy regime because the monetary policy regime itself changes after the government spending shock.
Appendix C: Appendix to “When is an Expectile the Best Linear Unbiased Estimator?”

Appendix C.1

Identification: Existence of the solution to the expectile regression problem in section 3.4 follows from Assumption 3. Assumption 3 requires that the matrix $X$ has full column rank. In turn, this implies that the first-order condition to the least squares problem

$$
\hat{\beta}_\tau = \arg \min_{\beta} \int \zeta_\tau(y - x'\beta)dP_n
$$

$$
\implies \hat{\beta}_\tau = (X'WX)^{-1}X'Wy
$$

exists and is unique. If Assumption 3 is violated, then the matrix $X'WX$ will not be invertible and the expectile regression coefficients cannot be evaluated. Because

$$
\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B)),
$$

invertibility of $X'WX$ requires that the data matrix $X$ has full rank. This is standard for ordinary least squares, but in the expectile case full rank is also sufficient to ensure that $X'WX$ is invertible. To show that this is true, observe that a unique Moore-Penrose pseudoinverse $(X'X)^{-1}X'$ exists for any matrix $X$ with full column rank. Then certainly $X'X$ is invertible. Likewise, for $\tau \in (0, 1)$, the matrix of expectile weights $W$ is diagonal (symmetric, positive definite) and has a positive definite square root $W^{1/2}$ with full rank. Then $W^{1/2}X$ is the $n \times k$ matrix $[W^{1/2}X]_{i,j} = W_{ii}^{1/2}[X]_{i,j}$. If this matrix is not of full rank, then there is some vector $v \in \mathbb{R}^k$ such that $\sum_{j=1}^k W_{ii}^{1/2}[X]_{i,j}r_j = 0$ for all $i$, which implies $\sum_{j=1}^k [X]_{i,j}r_j = 0$ for all $i$, which is impossible because $X$ has full rank. Then $W^{1/2}X$ has
a unique Moore-Penrose pseudoinverse \((X'WX)^{-1}X'W^{1/2}\) whenever \(X\) is of full rank, and \(X'WX\) is invertible.

In addition, Assumption 3 requires that there are at least \(k\) observations; \(n \geq k\). If \(n = k\), then the equation

\[ y = X\beta \]

has an unique, “exact” solution for \(\beta = X^{-1}y\) because \(X\) is square matrix with full rank (and thus invertible). If \(n > k\), then we might consider using the same Moore-Penrose pseudoinverse of \(X\), which is \((X'X)^{-1}X'\) to obtain a solution: \(\beta = (X'X)^{-1}X'y\). The pseudoinverse of a full-rank matrix is unique, but there are other matrices which can be used to “solve” a linear equation of this type. They include \((X'AX)^{-1}X'A\), with an infinite number of possibilities for the choice of \(A\), including the expectile weights \(A = W\). In that case, the mean regression is also the expectile regression for every \(\tau \in (0,1)\). Because the case where \(n = k\) is uninteresting, we will restrict our attention to the case where \(n > k\) and sample expectiles do not co-locate. Any generalized inverse \((X'AX)^{-1}X'A\) can be used to produce an “estimate” of \(\beta\). We will produce a unique most-efficient choice set from that set of possibilities in Appendix C.2: The Linear Unbiased Estimator and section 3.4.2.

Appendix C.2: The Linear Unbiased Estimator

Here, we prove that the Expectile coefficients are unbiased with known variance. This follows closely the “standard” result found in Greene [2003].

Start by taking Assumption 1 and Assumption 2. Then the model is linear and we have weighted strict exogeneity. If we write Assumption 2, replacing \(\epsilon_i\) with \(y_i - x_i'\beta\), we have an assumption on the population moment

\[ E(w_i(y_i - x_i'\beta)|X) = 0. \]

This also implies orthogonality given in equation 3.16:

\[ E(x_iw_i(y_i - x_i'\beta)|X) = 0_k. \]
The sample counterpart\textsuperscript{18} to this is as below, which leads to the estimator. We choose our estimator $\hat{\beta}_\tau$ in order to ensure that equation 3.16 holds in-sample.

$$0_k = \frac{1}{n} \sum_{i=1}^{n} x_i w_i (y_i - x_i' \hat{\beta}_\tau)$$

$$= \sum_{i=1}^{n} x_i w_i (y_i - x_i' \hat{\beta}_\tau)$$

$$= \sum_{i=1}^{n} x_i w_i y_i - \sum_{i=1}^{n} x_i w_i x_i' \hat{\beta}_\tau$$

$$= X'W y - X'W \hat{\beta}_\tau$$

Clearly, this implies $\hat{\beta}_\tau = (X'W X)^{-1} X'W y$ as would be the case with any weighted least squares problem. This is a \textit{linear} estimator in the usual sense: not only is the model linear in parameters (Assumption 1) but we also have $\hat{\beta}_\tau$ as a linear function of $y$; we merely left-multiply by the matrix $(X'W X)^{-1} X'W$ and obtain our estimate. That is quite convenient algebraically and for numeric computation.

We also wish to know whether the estimator $\hat{\beta}_\tau$ is unbiased. We say that the estimator is unbiased if its expected value is equal to its true value or if the expected sampling error is zero. The sampling error for $\hat{\beta}_\tau$ in this case is the difference between the estimator and its true value, say $\beta_\tau$. First decompose the estimator as follows

$$\hat{\beta}_\tau = (X'W X)^{-1} X'W y$$

$$= (X'W X)^{-1} X'W (X\beta_\tau + \epsilon)$$

$$= \beta_\tau + (X'W X)^{-1} X'W \epsilon. \quad \text{(C.2)}$$

The question is whether $E(\hat{\beta}_\tau - \beta_\tau) = 0$. We will take $W$ as known and use the tower

\textsuperscript{18} As the theoretical expectile can be defined as the expectation of the variable under a modified distribution, $\tilde{F}$, and the sample moment conditions are the weighted orthogonality conditions under the empirical distribution $\tilde{F}_n(y) = n^{-1} \sum_{i=1}^{n} I(y_i \leq y)$, so too can the sample moment conditions be expressed as standard moment conditions with respect to $\tilde{F}_n$, a weighted empirical distribution conforming to the definition in 3.7.


\[ E(\hat{\beta}_r - \beta_r) = E(\beta_r + (X'WX)^{-1}X'W\epsilon - \beta_r) \]
\[ = E \left( (X'WX)^{-1}X'W\epsilon \right) \]
\[ = E \left( E \left( (X'WX)^{-1}X'W\epsilon | X, W \right) \right) \]
\[ = E \left( (X'WX)^{-1}E \left( \sum_{i=1}^{n} x_i w_i \epsilon_i | X, W \right) \right) = 0 \quad \text{(C.3)} \]

Then \( \hat{\beta}_r \) is an unbiased estimator. As is the case with GLS-type estimators \( W \) is not necessarily known and this estimator is not necessarily feasible. However, we have devoted section B.3.3 of this paper to the feasibility of the weights \( w_i \). There, we provide some examples of data and model structures where \( W \) is known (exact) with probability one; then the estimator is perfectly feasible. In those cases, \( \hat{\beta}_r \) is a sort of “oracle” estimator because it has perfect foreknowledge of the weights \( w_i \). Otherwise, the unbiased version of \( \hat{\beta}_r \) is not feasible to estimate and its feasible counterpart with estimated weights may be inferior.

Next we state the variance of our estimator \( \hat{\beta}_r \). From equations C.2 and C.3 we can see that

\[ \text{Var}(\hat{\beta}_r | X, W) = E \left( \hat{\beta}_r - \beta_r \right) \left( \hat{\beta}_r - \beta_r \right)' | X, W \]
\[ = E \left( (X'WX)^{-1}X'W\epsilon \epsilon' X'W \left( (X'WX)^{-1} | X, W \right) \right) \]
\[ = (X'WX)^{-1}X'WE \left( \epsilon \epsilon' | X, W \right) W X \left( (X'WX)^{-1} \right) \quad \text{(C.4)} \]

Supposing that the un-weighted non-central variance\(^{19}\) of \( \epsilon | X \) is \( E(\epsilon \epsilon' | X, W) = \Sigma \), we have our solution. This sandwich-type formula\(^{20}\) would simplify under conditions such as, for

---

\(^{19}\)We have \( E(\epsilon_i | X) \neq 0 \), so \( \text{Var}(\epsilon | X) \neq E(\epsilon \epsilon' | X) \). The identity in equation 3.18 is not applicable because we are interested in the variance around \( X'\beta_r \) under the usual distribution of errors, not the weighted variance, and the variance or weighted variance around the usual measure of central tendency \( E(y | X) \).

\(^{20}\)Replacing \( E(\epsilon \epsilon' | X, W) \) with the estimator \( \hat{\Sigma} = \text{diag}(\hat{\epsilon}_i^2) \) produces the usual sandwich type heteroscedasticity-robust estimator of Eicker [1967] or White et al. [1980]. That estimator, with the
instance, \( E(W \epsilon' | X) = \nu^2 I_n \) which we assumed in Assumption 4. Then, if that assumption holds\(^{21}\),

\[
    Var(\hat{\beta}_\tau | X, W) = \nu^2 (X'WX)^{-1}. \tag{C.5}
\]

As is the case with ordinary least squares, the last assumption is difficult to take seriously. Nevertheless, this is the result that nests the classical OLS covariance, \( \sigma^2 (X'X)^{-1} \), when \( \tau = .5 \). In practice, we suggest deference to the heteroscedasticity-robust covariance matrix in equation C.4.

**Appendix C.3**

Additional proofs of optimality conditions are found here.

**Proof of Proposition 4:** The \( \tau^{th} \) expectile of \( Ay \) given \( X \) is

\[
    E_\tau(Ay | X, W) = E_\tau(AX \beta_\tau + A\epsilon | X)
    = E_\tau(AX \beta_\tau | X) + E_\tau(A\epsilon | X)
    = AX \beta_\tau.
\]

Clearly, we can construct a predictor of \( Ay \) using any estimate we may have for \( \beta_\tau \). If we use the unbiased estimator \( \hat{\beta}_\tau = (X'WX)^{-1} X'Wy \), we have

\[
    E(A\hat{y} | X, W) = E(AX \hat{\beta}_\tau | X, W)
    = AX \hat{\beta}_\tau
\]

\(^{21}\)A standard estimator for equation C.5 is

\[
    \hat{\nu}^2 (X'\hat{W}X)^{-1} = \frac{1}{n-k} \sum_{i=1}^{n} \hat{w}_i^2 \left( X'\hat{W}X \right)^{-1}.
\]

See the discussion in Section 3.6.
Thus, $A\hat{y}$ is an unbiased predictor of the $\tau^{th}$ expectile of $Ay$. Its variance is given by

$$Var(A\hat{y}|X, W) = E ((A\hat{y} - E(A\hat{y}|X, W))(A\hat{y} - E(A\hat{y}|X, W))'|X, W)$$

$$= E ((A\hat{y} - AX\beta_\tau)(A\hat{y} - AX\beta_\tau)'|X, W)$$

$$= AX E \left( (\hat{\beta}_\tau - \beta_\tau)(\hat{\beta}_\tau - \beta_\tau)'|X, W \right) X'A'$$

$$= \nu^2 AX (X'WX)^{-1} X'A'. $$

For any other unbiased estimator $\tilde{\beta}_\tau$ as given before, we have the predictor $A\tilde{y} = AX\tilde{\beta}_\tau$ and $E(A\tilde{y}|X, W) = AX\beta_\tau$. But the variance of $A\tilde{y}$ is

$$Var(A\tilde{y}|X, W) = E ((A\tilde{y} - E(A\tilde{y}|X, W))(A\tilde{y} - E(A\tilde{y}|X, W))'|X, W)$$

$$= AX E \left( (\tilde{\beta}_\tau - \beta_\tau)(\tilde{\beta}_\tau - \beta_\tau)'|X, W \right) X'A'$$

$$= AX \left( D + (X'WX)^{-1} X'W \right) \nu^2 W^{-1} \left( D + (X'WX)^{-1} X'W \right)' X'A'$$

$$= \nu^2 \left( AXDW^{-1}D'X'X + AX (X'WX)^{-1} X'A' \right)$$

$$\geq \nu^2 AX (X'WX)^{-1} X'A' = Var(A\hat{y}|X, W)$$

Again, the positive-definite matrix $AXDW^{-1}D'X'A'$ shows that our variance is at least as great as before. By extension, we see that $AX\hat{\beta}_\tau$ is the best linear predictor of $AX\beta_\tau$ for a linear function $B = (AX)$ of $\beta_\tau$. This proves proposition 4.

**Proof of Proposition 6:** We have

$$\hat{\beta}_{\tau, GLS} = (X'X)^{-1} X'\tilde{y}$$

$$= (X'V'VX)^{-1} X'V'V\tilde{y}$$

$$= (X'\Sigma^{-1}X)^{-1} X'\Sigma^{-1}\tilde{y}$$

$$= \beta_\tau + (X'\Sigma^{-1}X)^{-1} X'\Sigma^{-1}\epsilon$$

214
where, because the sequence \( \{Y, X\} \) is independent,

\[
E \left( (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} \epsilon | X, W \right) \\
= (X' \Sigma^{-1} X)^{-1} X' E \left( W^{1/2} \Omega^{-1} W^{1/2} \epsilon | X, W \right) \\
= (X' \Sigma^{-1} X)^{-1} X' \Omega^{-1} \underbrace{E (W \epsilon | X, W)}_0 .
\]

Then the expectile GLS estimator is unbiased. It remains to be shown whether this estimator has the lowest possible variance for the class of linear estimators. Obviously the variance of the estimator itself is \( \nu^2 (X' \Sigma^{-1} X)^{-1} \).

The proof is similar to the example in the previous section. Because any unbiased linear estimator can be written \( \hat{\beta}_\tau = Cy \) where \( CX = I_n \) and \( C = D + (X' \Sigma^{-1} X) X' \Sigma^{-1} \), \( DX = 0 \), we can write

\[
Var(Cy | X, W) = Var((D + (X' \Sigma^{-1} X) X' \Sigma^{-1}) y | X, W) \\
= \nu^2 \left( D + (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} \right) \Sigma \left( D + (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} \right) \\
= \nu^2 D \Sigma D' + \nu^2 (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} \Sigma \Sigma^{-1} X (X' \Sigma^{-1} X)^{-1} \\
\geq \nu^2 (X' \Sigma^{-1} X)^{-1} = Var(\hat{\beta}_{\tau, GLS})
\]

Thus we have shown that the GLS estimator is the “best” linear unbiased estimator for \( \beta_{\tau} \) in the sense of minimum variance. This completes the proof of proposition 6.

**Proof of Proposition 8 on page 96:** The proposition states that a particular individual with covariates \( z \) has atypical odds of a positive versus negative residual \( \epsilon^* \):

\[
\frac{Pr(\epsilon^* \geq 0)}{Pr(\epsilon^* < 0)} = \alpha \frac{Pr(\epsilon_i \geq 0)}{Pr(\epsilon_i < 0)}
\]
where $\epsilon_i$ is a residual from the general population. To improve our estimator and our predictor, we can incorporate that new information in the following way.

$$E(\epsilon^*|z, X) = E(\epsilon^*|z, X, \epsilon^* \geq 0) \Pr(\epsilon^* \geq 0) + E(\epsilon^*|z, X, \epsilon^* < 0) \Pr(\epsilon^* < 0)$$

$$= E(\epsilon^*|z, X, \epsilon^* \geq 0) \Pr(\epsilon_i \geq 0) \times \alpha \Pr(\epsilon^* < 0) \Pr(\epsilon_i < 0)$$

$$+ E(\epsilon^*|z, X, \epsilon_i < 0) \Pr(\epsilon^* < 0) \times \frac{1}{\alpha} \Pr(\epsilon_i \geq 0) \Pr(\epsilon_i < 0)$$

$$= E(\epsilon^*|z, X, \epsilon^* \geq 0) \times \tau + E(\epsilon^*|z, X, \epsilon_i < 0) \times (1 - \tau) \quad (C.7)$$

for some $\tau \in (0, 1)$. But clearly this is $E_{\tau}(\epsilon^*|z, X)$ or $E(w^* \epsilon^*|z, X)$ with proper expectile weights. We may produce an unbiased estimator based on this information by using the sample moment, for instance. This is the same moment condition in equation C.1 and leads to the usual expectile regression estimator. We know that $E(w^* \epsilon^*|z, X) = 0$ must hold for all possible $z$ and, thus, it holds for all $x_i$, implying $E(w_i \epsilon_i|X, W)$ for all $i$ in $1, ..., n$. The same condition must be true for any other unbiased estimator.

Because any unbiased estimator must elicit the $\tau^{th}$ expectile of the observable distribution, we may find the predictor with the lowest variance by minimize the expression for variance directly. Following the fact that $E_{\tau}(y|X) = X\beta_{\tau}$, any unbiased linear predictor $z'\tilde{\beta}$ has $E_{\tau}(z'\tilde{\beta}|X) = z'CX\beta_{\tau} = z'\beta_{\tau}$. The variance of the unbiased predictor is

$$Var(z'\tilde{\beta}|X, W) = E \left( (z'\tilde{\beta} - E_{\tau}(z'\tilde{\beta}|X, W))^2 | X, W \right)$$

$$= z'CE \left( (y - X\beta_{\tau})(y - X\beta_{\tau})' | X, W \right) C'z$$

$$= \nu^2 z'C\Sigma C'z$$

where $\nu^2 \Sigma = \nu^2 W^{-1/2}\Omega W^{-1/2}$ denotes $E(\epsilon^*|X, W)$ as before. This is constant with respect to the choice of unbiased estimator. Then we can minimize the variance, subject to

---

22 The weights in equation C.7 add to one because, as seen previously, they are equal to $\Pr(\epsilon^* \geq 0)$ and $\Pr(\epsilon^* < 0)$, respectively.

23 The variance of the predictor in equation C.8 is obtained using the real distribution of data, so the first expectation operator is not weighted. But we want the predicto to be unbiased under the alternative distribution of errors, which makes the second expectations operator weighted.

---
unbiasedness of the estimator, by constrained optimization. Writing the Lagrangian below,

\[ L(z'C) = \frac{1}{2} z' C \Sigma C' z + \lambda' (X'C' z - z) \]  

(C.9)

minimize with respect to \( z'C \):

\[ \frac{\partial L(z'C)}{L(z'C)} = z'C \Sigma + \lambda' X' = 0 \]  

(C.10)

And the first-order condition with respect to \( \lambda \) leaves \( CX - I = 0 \). Together, these imply

\[ X' \Sigma^{-1} X \lambda = X' C' z \]

\[ \implies \lambda = (X' \Sigma^{-1} X)^{-1} z \]  

(C.11)

and

\[ \Sigma C' z = X \lambda = X (X' \Sigma^{-1} X)^{-1} z \]

\[ \implies C' = \Sigma^{-1} X (X' \Sigma^{-1} X)^{-1} \]  

(C.12)

Thus, the best linear unbiased predictor \( z' \hat{\beta}_\tau \) makes use of the GLS-type estimator

\[ \hat{\beta}_{\tau, \text{GLS}} = C y = (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} y \]  

(C.13)

where \( \nu^2 \Sigma = E((y - X \beta_\tau)(y - X \beta_\tau)'|X,W) \). Under the ideal conditions such as in Expectile Assumption 4, \( E((y - X \beta_\tau)(y - X \beta_\tau)'|X,W) = \nu^2 W^{-1} \), we have the usual expectile estimator

\[ \hat{\beta}_\tau = C y = (X' W X)^{-1} X' W y. \]  

(C.14)

But this need not be the case. We have proven proposition 8.

**Proof of Proposition 11 on page 99:** Identical to those given elsewhere. We have an unbiased linear predictor given by \( z' C y \) with \( E(z' C y|X,W) = z' \beta_\tau \). The variance of any
such predictor is

\[
\text{Var}(z'Cy|X, W) = z'\text{Var}(Cy - C\beta_t|X, W) z \\
= z'\left(D + (X'S^{-1}X)^{-1}X'S^{-1}\right)\nu^2\Sigma \left(D + (X'S^{-1}X)^{-1}X'S^{-1}\right) z \\
= \nu^2z'D\Sigma D'z + \nu^2z'(X'S^{-1}X)^{-1}z \\
\geq \nu^2z'(X'S^{-1}X)^{-1}z = \text{Var}(z'\hat{\beta}_{\tau,GLS}|X, W). \tag{C.15}
\]

This completes the proof of proposition 11 on page 99.

**Proof of Proposition 12 on page 100:**

*Proof.* We know that

\[
E(\epsilon_i|X) = E(\epsilon_i|X, \epsilon_i \geq 0) \Pr(\epsilon_i \geq 0) + E(\epsilon_i|X, \epsilon_i < 0) \Pr(\epsilon_i < 0) \tag{C.16}
\]

\[
= c
\]

So,

\[
E(w_i\epsilon_i|X) = \tau E(\epsilon_i|X, \epsilon_i \geq 0) \Pr(\epsilon_i \geq 0) + (1 - \tau)E(\epsilon_i|X, \epsilon_i < 0) \Pr(\epsilon_i < 0) \\
= E(\epsilon_i|X) - (1 - \tau)E(\epsilon_i|X, \epsilon_i \geq 0) \Pr(\epsilon_i \geq 0) - \tau E(\epsilon_i|X, \epsilon_i < 0) \Pr(\epsilon_i < 0) \\
= c - (1 - \tau)E(\epsilon_i|X, \epsilon_i \geq 0) \Pr(\epsilon_i \geq 0) - \tau E(\epsilon_i|X, \epsilon_i < 0) \Pr(\epsilon_i < 0). \tag{C.17}
\]

The last expression above will be zero if

\[
c = (1 - \tau)E(\epsilon_i|X, \epsilon_i \geq 0) \Pr(\epsilon_i \geq 0) + \tau E(\epsilon_i|X, \epsilon_i < 0) \Pr(\epsilon_i < 0) \\
= E(\epsilon_i|X, \epsilon_i \geq 0) \Pr(\epsilon_i \geq 0) + \tau (E(\epsilon_i|X, \epsilon_i < 0) \Pr(\epsilon_i < 0) - E(\epsilon_i|X, \epsilon_i \geq 0) \Pr(\epsilon_i \geq 0))
\]

so the unique \(\tau\) that satisfies this condition must be

\[
\tau = \frac{E(\epsilon_i|X, \epsilon_i \geq 0) \Pr(\epsilon_i \geq 0) - c}{E(\epsilon_i|X, \epsilon_i \geq 0) \Pr(\epsilon_i \geq 0) - E(\epsilon_i|X, \epsilon_i < 0) \Pr(\epsilon_i < 0)}. \tag{C.18}
\]
But, because of equation C.16, we can replace $c$ and write $\tau$ as

$$\tau = \frac{-E(\epsilon_i|X, \epsilon_i < 0) \Pr(\epsilon_i < 0)}{E(\epsilon_i|X, \epsilon_i \geq 0) \Pr(\epsilon_i \geq 0) - E(\epsilon_i|X, \epsilon_i < 0) \Pr(\epsilon_i < 0)}.$$  \hspace{1cm} (C.19)

Clearly this has the form $\frac{a}{a+b}$ for positive $a, b$; therefore $\tau$ must be in $(0, 1)$. This proves proposition 12 on page 100.

\[ \square \]

**Appendix C.4**

Throughout our discussion of the expectile estimator and its corresponding predictor, we have employed the assumption that the true weight matrix $W$ is known perfectly. In practice, this may fail to be the case, but it holds in at least one example as in Figure 3.2 on page 77.

The estimator for expectile coefficients

$$\hat{\beta}_\tau = (X'WX)^{-1}X'Wy$$  \hspace{1cm} (C.20)

has a typical generalized least squares form with diagonal weight matrix given by

$$[W]_{ii} = \begin{cases} 
\tau & \text{if } y_i - x'_i\hat{\beta}_\tau \geq 0 \\
1 - \tau & \text{if } y_i - x'_i\hat{\beta}_\tau < 0.
\end{cases}$$  \hspace{1cm} (C.21)

GLS estimators of this form are usually considered to be infeasible because their weights (above) are not known a priori. The “feasible” GLS-type estimator uses estimated weights obtained jointly with the linear coefficients $\hat{\beta}_{\tau,FGLS}$. In our case, that gives

$$\hat{\beta}_{\tau,FGLS} = (X'\hat{W}X)^{-1}X'\hat{W}y$$

$$[\hat{W}]_{ii} = \hat{w}_i = \begin{cases} 
\tau & \text{if } y_i - x'_i\hat{\beta}_{\tau,FGLS} \geq 0 \\
1 - \tau & \text{if } y_i - x'_i\hat{\beta}_{\tau,FGLS} < 0.
\end{cases}$$  \hspace{1cm} (C.22)

Estimation of $\hat{\beta}_{\tau,FGLS}, \hat{W}$ is usually achieved by iteratively reweighted least squares, which is a simple algorithmic procedure. Given some initial condition for $\hat{\beta}_\tau$, such as OLS
estimates, weights $\hat{W}$ are obtained and a new $\hat{\beta}_{\text{FGLS}}$ can be evaluated. Repeating this procedure with the new value $\hat{\beta}_{\text{FGLS}}$, the estimated coefficients will converge relatively quickly as the sub-Hessian for this problem is globally negative semidefinite (see Philipps [2021a]). Then the converged estimator is the correct (exact) sample linear expectile coefficient vector. This is similar overall to the procedure used by Zellner [1962] for the mean regression.

It is clear from the definition in equation C.22 on the preceding page that many of the estimated weights will be exactly correct and the entire weight matrix is estimated consistently so long as $\hat{\beta}_{\text{FGLS}}$ is consistent, which was proven by Newey and Powell [1987]. In the heteroscedastic case, we have $\Sigma_{ii} = w_i \omega_{ii}$. Then consistency of $\hat{\beta}_{\text{FGLS}}$ requires a consistent estimator of $w_i$ and one of $\omega_{ii}$. See Barry et al. [2018] or Philipps [2021a] for asymptotic conditions for that estimator.

However, the expectile regression model differs from the usual GLS example because the optimal estimator is not always infeasible. For any nondegenerate distribution $F$, continuous or discrete, the expectile function $E_Y(\tau) : [0, 1] \mapsto \text{Support}(Y)$ is surjective, causing the true expectile to fall between points with positive probability density (or mass) of $Y$ with probability one for $\tau \in (0, 1)$. In any sample, the empirical CDF $F_n$ will itself be discrete, causing the sample expectile (or linear predictor) to fall between observations almost surely. If the true expectile and the sample expectile fall between the same set of observations,

$$\{ i : y_i - x_i' \hat{\beta}_{\text{FGLS}} \geq 0 \} \equiv \{ i : y_i - x_i' \beta \geq 0 \} \quad \text{(C.23)}$$

then the estimated weights are exact; $\hat{w}_i = w_i$. In that case, the optimal expectile estimator $\hat{\beta}_\tau = (X'WX)^{-1}X'Wy$ is feasible. Using the location model as an example, we can see that this happens with positive probability.

**Lemma 21.** Let $Y$ be distributed according to $F$ with Lebesgue measure. Let $\{y_i\}_{i=1}^n$ be an i.i.d. sample from $F$. The distribution of the random variable $\hat{\mu}_\tau$ has positive density on the interval $[\inf(Y), \sup(Y)]$.

**Proof.** We estimate the $\tau^{th}$ sample expectile $\hat{\mu}_\tau$ by minimizing $R_n(\theta; \tau) = n^{-1} \sum_{i=1}^n s_\tau(y_i -
\(\theta\), where the “swoosh” function \(\varsigma\) is as in equation C.24 or equation 3.5 on page 73,

\[
\varsigma(u) = u^2|\tau - I(u < 0)|. \tag{C.24}
\]

Because the solution of this strictly convex minimization problem (and the first-order condition) is unique, we have

\[
\hat{\mu}_\tau = \arg \min_{\theta} n^{-1} \sum_{i=1}^{n} \varsigma(y_i - \theta).
\]

\[
\implies n^{-1} \sum_{i=1}^{n} (y_i - \hat{\mu}_\tau)|\tau - I(y_i - \hat{\mu}_\tau < 0)| = 0 \tag{C.25}
\]

or

\[
n^{-1}(1 - \tau) \sum_{i=1}^{n} (y_i - \hat{\mu}_\tau)I(y_i - \hat{\mu}_\tau < 0) = n^{-1} \tau \sum_{i=1}^{n} (y_i - \hat{\mu}_\tau)I(y_i - \hat{\mu}_\tau \geq 0) \tag{C.26}
\]

The probability of the sample expectile \(\hat{\mu}_\tau\) being less than or equal to a particular value \(x\), is then the same as

\[
\Pr\left(\tau \int_{x}^{\infty} (y - x)dF_n \right) \leq \left(1 - \tau\right) \int_{-\infty}^{x} (x - y)dF_n \right), \tag{C.27}
\]

where the term on the left is clearly monotone decreasing with \(x\) and the term on the right is monotone increasing. This means that, as \(x\) increases, the probability of the event \(\mu_{\tau} \leq x\) is nondecreasing (the set of possible values for \(D := \{y_i\}_{i=1}^{n} : \mu_{\tau} \leq x\} \) is increasing). In particular, in means that the probability of the event \(\Pr(\mu_{\tau} \leq x) = \int_{D} f(y)dy\) is monotone increasing if \(F\) has Lebesgue measure, implying that the distribution of \(\mu_{\tau}\) has positive density on \([\inf(Y), \sup(Y)]\).

\[\square\]

**Proposition 22.** Let \(Y\) be distributed according to \(F\) with Lebesgue measure. Let \(\{y_i\}_{i=1}^{n}\) be an i.i.d. sample from \(F\). The sample expectile weights \(\hat{w}_i\) for \(\hat{\mu}_\tau : \tau \in (0,1)\) are exact with positive probability.

**Proof.** This follows directly from the previous lemma, as the finite-sample distribution of \(\hat{\mu}_\tau\)
has positive density on the interval \([\inf(Y), \sup(Y)]\). For \(\tau \in (0, 1)\), \(\mu_\tau\) falls into the interior of this interval with probability one. Then, for any sample, the probability that there is at least one observation above \(\mu_\tau\) and at least one observation below \(\mu_\tau\) is positive. Then the interval \([\max(y_i|y_i < \mu_\tau), \min(y_i|y_i \geq \mu_\tau)]\) exists with positive probability. Then, because the distribution of \(\hat{\mu}_\tau\) has positive density, the probability that \(\max(y_i|y_i < \mu_\tau) < \hat{\mu}_\tau < \min(y_i|y_i \geq \mu_\tau)\) is also positive.

The same result is applicable asymptotically to distributions that do not have Lebesgue measure. Because the support of the distribution of \(\hat{\mu}_\tau\) is asymptotically dense, there is a nonempty set of possible samples producing \(\max(y_i|y_i < \mu_\tau) < \mu_\tau < \min(y_i|y_i \geq \mu_\tau)\) as \(n \to \infty\).

Suppose that the distribution \(F\) does not have Lebesgue measure, but let \(F\) be continuous at \(\mu_\tau\). Let there be some sequence \(a_n \to \infty\) such that \(a_n(\mu_{\tau,n} - \mu_\tau) \rightsquigarrow Z\) as \(n \to \infty\). Very broad conditions for this limiting behavior in expectiles were published recently Holzmann and Klar [2016]. So long as \(Z\) has density, clearly,

\[
\int_{a_n(\max(y_i|y_i < \mu_\tau) - \mu_\tau)}^{a_n(\min(y_i|y_i \geq \mu_\tau) - \mu_\tau)} dZ > 0 \quad \text{(C.28)}
\]

for any \(n < \infty\) and asymptotically so long as \(a_n(\min(y_i|y_i \geq \mu_\tau) - \max(y_i|y_i < \mu_\tau)) \not\to 0\). In one interesting example, where the distribution \(F\) is fully discrete or has no density in an open ball around \(\mu_\tau\), the length of the interval \(\min(y_i|y_i \geq \mu_\tau) - \max(y_i|y_i < \mu_\tau)\) diverges as \(n \to \infty\), so

\[
a_n(\min(y_i|y_i \geq \mu_\tau) - \mu_\tau) \to \infty \quad \text{(C.29)}
\]

\[
a_n(\mu_\tau - \max(y_i|y_i < \mu_\tau)) \to \infty. \quad \text{(C.30)}
\]

which implies

\[
\lim_{n \to \infty} \int_{a_n(\max(y_i|y_i < \mu_\tau) - \mu_\tau)}^{a_n(\min(y_i|y_i \geq \mu_\tau) - \mu_\tau)} dZ = \int_{-\infty}^{\infty} dZ = 1 \quad \text{(C.31)}
\]

That is, the sample expectile weights \(\hat{w}_i\) are correct with probability one in the limit.

Furthermore, it is possible to construct a realistic example where sample weights are
exact with probability one in a finite sample. This will be the case in well-specified binary response models, in particular.

Appendix C.5

Among regression models where the dependent variable is not assumed to have a continuous distribution, the binary response case is quite common. These models occur when $Y$ takes one of two values, usually labeled 0 and 1, with some probability dependent on covariates $X$. See Heckman and Snyder Jr [1996], Heckman and MaCurdy [1985] for examples.

Because the dependent variable is binary and a Bernoulli distribution is indexed by a single parameter (the probability of a nonzero outcome) it is popular to finish the specification of a stochastic binary response model by assuming that

$$E(Y|X = x) = \Pr(Y = 1|x) = G(x'\beta)$$

for some convenient class of function $G$. The linear probability model $G(x'\beta) = x'\beta$ is the obvious choice if we treat the model as being no different from any other linear regression. But because $X'\beta$ is unbounded for unbounded $X$ and the least squares criterion is indifferent towards whether or not the predictor falls within the unit interval for all observations, other designs have been advocated. These include the famous Logit and Probit models; see Aldrich et al. [1984] for a thorough comparison or see Greene [2003] for an overview. These two models and others are designed to remain bounded within the unit interval for all $X$.

Without respect to the individual function $G$ chosen in any particular application, we say that the binary response model is “well specified” if it maps the inner product of two $k$-vectors, $X'_i\beta$, to the interior of the unit interval$^{24}$

$$G : \mathbb{R}^k \mapsto (0, 1)$$

and we note the obvious result.

---

$^{24}$There is a question regarding whether $G$ should map its inputs to the open interval $(0, 1)$ or the closed interval $[0, 1]$. In general, it would be undesirable to predict that $\Pr(y_i = 1|x_i) = 1$ or 0 and to observe an error term, as this would make the model logically incoherent.
Proposition 23. Let \( y, X \) belong to a binary response regression problem with \( y_i \in \{0,1\} \forall i \). Let the predictor \( G(x' \beta) \) be well-specified as in equation C.33. Then the estimated expectile weights are exact (the optimal weights matrix \( W \) is feasible) with probability one.

The proof of this proposition is obvious: the response variable takes only two values and the predictor falls strictly between them. If the \( i^{th} \) residual from the true binary data generating process is positive, then \( y_i \) equals one. Then \( y_i > G(x' \beta) \) regardless of \( x, \beta \), and \( \hat{w}_i = w_i = \tau \). The same logic applies to negative residuals.

The function \( G(x' \beta) \) is now the predictor, which removes the obvious interpretation from the coefficient vector \( \beta \) when \( G \) is nonlinear. However, most of the results we might be interested in obtaining can be reproduced for the predictor \( G(x' \beta) \). For instance, the variance of a Bernoulli variable is obvious as is the conditional distribution of errors. For simple parametric sigmoid functions or probability distributions that are common choices for \( G \), further algebraic results can be obtained. See Angrist and Pischke [2008] for applications to the mean regression. Expectile logit and probit models of this form have not been developed but are a promising area for future research. We invite other authors to consider this topic.

Appendix C.6

Expectile Projection and Annihilator Matrices: Here we discuss briefly the projection and annihilator matrices from this problem, which is helpful for the result in appendix A3. It is reasonably obvious that these matrices are symmetric but not idempotent. That fact is standard for GLS-type estimators, but reduces to a particular meaningful form for expectile weights. First, notice that \( P_{5.5} = X(X'X)^{-1}X' \) and \( M_{.5} = I - X(X'X)^{-1}X' \) are symmetric and idempotent. In the asymmetrically weighted case, we have

\[
P_{\tau} = X(X'W'X)^{-1}X'W \neq WX(X'W'X)^{-1}X' = P'_{\tau}
\]

and

\[
M_{\tau} = I - P_{\tau} \neq I - P'_{\tau} = M'_{\tau}
\]
so neither matrix is symmetric. However,

\[ P_\tau P_\tau = X(X'WX)^{-1}X'WX(X'WX)^{-1}X'W = X(X'WX)^{-1}X'W = P_\tau \]

\[ M_\tau M_\tau = (I - P_\tau)(I - P_\tau) = II - IP_\tau - P_\tau I + P_\tau P_\tau = I - P_\tau = M_\tau \]

both matrices are idempotent. These are simple in the location model: consider the special case where \( X \) is merely a vector of ones; \( 1_n \). Then

\[ P_\tau = 1_n(1'_nW1_n)^{-1}1'_nW \]

\[ = \left( \sum_{i=1}^{n} w_i \right)^{-1}1_n1'_nW \]

\[ = \left( \sum_{i=1}^{n} w_i \right)^{-1} \begin{pmatrix} w_1 & w_2 & \cdots & w_n \\ w_1 & w_2 & \cdots & w_n \\ \vdots & \vdots & \ddots & \vdots \\ w_1 & w_2 & \cdots & w_n \end{pmatrix} \]

note that left multiplication by \( P_\tau \) maps the vector \( y \) as follows

\[ P_\tau y = 1_n \frac{(1'_nW1_n)^{-1}1'_nWy}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \]

\[ = 1_n \bar{y}_\tau \]
However, if we write \( P'_r y \), we have

\[
P'_r y = W_1 (1'_n W_1 n)^{-1} 1'_n y
\]

\[
= \left( \sum_{i=1}^{n} w_i \right)^{-1} \begin{pmatrix} w_1 & w_1 & \cdots & w_1 \\ w_2 & w_2 & \cdots & w_2 \\ \vdots & \vdots & \ddots & \vdots \\ w_n & w_n & \cdots & w_n \end{pmatrix} y
\]

\[
= \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} \left( \sum_{i=1}^{n} y_i \right) \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}.
\]

Similarly,

\[
P'^r_1 P'^r_1 = W_1 (1'_n W_1 n)^{-1} 1'_n (1'_n W_1 n)^{-1} 1'_n W
\]

\[
= \left( \sum_{i=1}^{n} w_i \right)^{-2} \begin{pmatrix} w_1 & w_1 & \cdots & w_1 \\ w_2 & w_2 & \cdots & w_2 \\ \vdots & \vdots & \ddots & \vdots \\ w_n & w_n & \cdots & w_n \end{pmatrix} \begin{pmatrix} w_1 & w_2 & \cdots & w_n \\ w_1 & w_2 & \cdots & w_n \\ \vdots & \vdots & \ddots & \vdots \\ w_1 & w_2 & \cdots & w_n \end{pmatrix}
\]

\[
= n \left( \sum_{i=1}^{n} w_i \right)^{-2} \begin{pmatrix} w_1^2 & w_1 w_2 & \cdots & w_1 w_n \\ w_2 w_1 & w_2^2 & \cdots & w_2 w_n \\ \vdots & \vdots & \ddots & \vdots \\ w_n w_1 & w_n w_2 & \cdots & w_n^2 \end{pmatrix}.
\]
whereas,

\[ P_1^{-1}P_1' = 1_n(1_n'W1_n)^{-1}1_n'WW1_n(1_n'W1_n)^{-1}1_n' \]

\[ = \left( \sum_{i=1}^{n} w_i \right)^{-2} \begin{pmatrix} w_1 & w_2 & \cdots & w_n \\ w_1 & w_2 & \cdots & w_n \\ \vdots & \vdots & \ddots & \vdots \\ w_1 & w_2 & \cdots & w_n \end{pmatrix} \begin{pmatrix} w_1 & w_1 & \cdots & w_1 \\ w_2 & w_2 & \cdots & w_2 \\ \vdots & \vdots & \ddots & \vdots \\ w_n & w_n & \cdots & w_n \end{pmatrix} \]

\[ = \left( \sum_{i=1}^{n} w_i \right)^{-2} \sum_{i=1}^{n} w_i^2 \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \]

And the difference is clear.

**Appendix C.7**

**Trace of \( M_\tau'WM_\tau \):** Here, we show that \( \text{trace}(M_\tau'WM_\tau) = \text{trace}(W) - \text{trace}(WP_\tau) \).

The expected value of the weighted sum of squared errors is

\[ E(WSSE|X, W) = E(\hat{\epsilon}'W\hat{\epsilon}|X, W) \]

\[ = E(\hat{\epsilon}'M_\tau'WM_\tau\hat{\epsilon}|X, W) \]

\[ = \sum_{i=1}^{n} \sum_{j=1}^{n} [M_\tau'WM_\tau]_{ij} E(\epsilon_i\epsilon_j|X, W) \]

\[ = \sum_{i=1}^{n} [M_\tau'WM_\tau]_{ii} \sigma_\tau^2 \]

which relies on the trace of \( M_\tau'WM_\tau \).

\[ M_\tau'WM_\tau = (I - P_\tau)'W(I - P_\tau) \]

\[ = W - WP_\tau - P_\tau'W + P_\tau'WP_\tau. \]
This is straightforward. The two negative matrices have the same trace:

\[
\text{trace}(WP_{\tau}) = \text{trace}((X^TWX)^{-1}X'WX)
\]

\[
= \text{trace} \left( \left( \sum_{i=1}^{n} w_i x_i x_i' \right)^{-1} \left( \sum_{i=1}^{n} w_i^2 x_i x_i' \right) \right)
\]

\[
\leq k
\]

and

\[
\text{trace}(P_{\tau}'WP_{\tau}) = \text{trace}(WX(X^TWX)^{-1}X'WX(X^TWX)^{-1}X'W)
\]

\[
= \text{trace}((X^TWX)^{-1}X'WX(X^TWX)^{-1}X'WWX)
\]

\[
= \text{trace}(WP_{\tau}).
\]

So trace\((M_{\tau}'WM_{\tau}) = \text{trace}(W) - \text{trace}(WP_{\tau})\). Then

\[
E(WSSE|X, W) = (\text{trace}(W) - \text{trace}(WP_{\tau})) \sigma_\tau^2
\]

In the special case where \(X = 1_n\), we have

\[
E(WSST|X, W) = E(\epsilon'\hat{\epsilon}|X, W)
\]

\[
= E((y - \bar{y}_\tau)'W(y - \bar{y}_\tau)|X, W)
\]

\[
= (\text{trace}(W) - \text{trace}(WP_{\tau}^{1})) \sigma_\tau^2
\]

where

\[
\text{trace}(WP_{\tau}^{1}) = \text{trace}((1_n'W1_n)^{-1}1_n'WW1_n)
\]

\[
= \frac{\sum_{i=1}^{n} w_i^2}{\sum_{i=1}^{n} w_i} \leq 1.
\]

Note that this value is strictly contained in the unit interval because \(w_i^2 < w_i\) for \(\tau\in(0,1)\).
Table C.1: Quantile regression coefficients for the simple regression of *Deny* on *Black* are degenerate; compare to the regression lines in figure C.1 or the expectile regression coefficients in table 3.1. In addition to the coefficients being either 0 or 1, the analytical standard errors reported by STATA (or other software) for this problem are not valid.

### Appendix C.8

**Figures:** Supplementary figures for the two data applications are given here.

![Figure C.1](image)

Figure C.1: Simple regression of *Deny* on *Black* produces the linear probability model shown in black in the figure above. The fitted values are (0,.09) and (1,.28). These reveal little more about the data than the summary statistics would.
Figure C.2: A mean regression. Repatriation intensity is defined as the percentage decrease in population of Mexican nationals from 1930-40. The covariate is the proportion of the population that were Mexican nationals in 1930. On average, municipalities with a higher proportion of Mexican nationals in 1930 saw a larger repatriation intensity.

Figure C.3: Estimated expectile coefficients from a multivariate expectile GLS model. The effect of Black on Deny achieves its maximum of .174 at $\tau = .83$. Similar to the results in figures 3.3 and 3.6, individuals who are relatively likely to be denied have a larger racial disparity than those who are relatively unlikely to be denied.
Figure C.4: Estimated expectile coefficients from the multivariate expectile GLS model are very similar to those in Figure 3.7; the loss of smoothness evident from this figure occurs because observations are dropped if their GLS weights cannot be calculated. This occurs primarily at high values of $\tau$. 

![Graph of estimated expectile coefficients](image.png)