When Expectiles are BLUE

Collin S. Philipps*

November 7, 2019

Abstract

We generalize the classical Gauss-Markov framework to incorporate expectile regression. Expectile regression produces the best linear unbiased estimator for regression lines other than the mean in model designs with asymmetric conditional variance of the error term. In some cases where OLS assumptions are violated, an expectile regression estimator is the BLUE for the mean regression. The weighted estimator is also feasible in some cases; we provide an example. The usual unbiased estimator for residual mean squared error may be biased in this application: we suggest an alternative and discuss.

Keywords: Expectile Regression, Generalized Quantile Regression, Best Linear Unbiased Estimation, Generalized Least Squares

JEL Codes: C0, C1, C4

1 Introduction

The Gauss-Markov Theorem is a foundational result in econometrics. Under simple assumptions, that theorem states that the ordinarly least squares (OLS) estimator of regression coefficients is the **Best Linear Unbiased Estimator** (BLUE) in the sense that it has the minimum variance of any such estimator. It may be possible to find a linear estimator with the same variance (see [3]), but none better. The result extends to generalized least squares (GLS) under heteroscedasticity. In this paper, we show that there are scenarios where *an expectile regression* estimator is the BLUE.

The expectile regression of Newey and Powell [37] is not known universally, but has recently become more popular in the theoretical literature. This is primarily because expectile regression has the potential to elicit results not found using mean regression: see [45], [25], or [44] for discussions. Formally, expectiles are a type of *generalized quantile* [16], *m-quantile* [13], or L_p*-quantile* [14]. These families extend the original quantile regression of Koenker and Bassett [28] to estimation by alternative loss functions. For the location-scale model, different L_p-quantiles produce the same sets of regression lines, making a wide class of estimators available for each such line [16]. Expectiles are produced by using an asymmetrically weighted least squares estimator, which nests ordinary least squares (OLS) when the weights are constant. The virtues of the asymmetric class of estimators

*Department of Economics, University of Illinois Urbana-Champaign.

are widely extolled: see [26, 29]. In that class, expectiles are the unique example with a (piecewise) quadratic loss function and an interpretation as an expected value. Expectiles have other unique properties; see Ziegel [52] or Bellini et al. [11]. This document presents other fundamental properties of those estimators that should not be overlooked.

The motivating research question for this article is whether the Gauss-Markov theorem is extensible to generalized quantiles. It is. Because generalized quantiles produced by different scoring functions produce the same sets of possible regression lines [16], an obvious question is whether any one of these scoring functions might produce the "BLUE". Others have suggested that expectiles may be the efficient estimator for the m-quantile class [40] without exploring the concept in detail. Following that suspicion, we find that a simple generalization of the Gauss-Markov assumptions–focused on asymmetric variance of residuals–*does* extend the major result to expectiles.

To obtain this result, we discuss the "usual" regression assumptions in the context of the linear expectile regression proposed by Newey and Powell [37]. Because the expectile coefficients are produced by minimizing a quadratic (weighted least squares) function, the solution produces a linear estimator in the classical sense[1]. The linear form of the estimator conforms to the standard structure common to generalized least squares estimators. However, expectile regression coefficients are not estimators of the mean regression coefficients except in one special case. We show that the estimator that is the BLUE for a given expectile can be obtained trivially. We will also show that the estimated expectile coefficients and the corresponding predictor are the BLUE in the traditional sense (minimum variance linear unbiased) for the mean regression coefficients when the usual Gauss-Markov assumptions are violated in a particular way.

A critical impediment to the adoption of expectiles has been the lack of any *elementary* treatment of the subject. We seek to remedy this fault in some small way by re-casting the expectile regression in the classical framework. In that framework, a substantial number of additional results become obvious. Among these, we obtain the expectile GLS (or generalized expectile regression) estimator and show that it is the BLUE for Newey and Powell's regression expectiles under heteroscedasticity. We also provide two examples where one of Newey and Powell's expectile estimators (or our GLS variant) is the BLUE in a mean regression. Specifically, we adopt four assumptions from the classical linear model: linearity in parameters, strict exogeneity, 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. The fifth assumption typically presented in the classical linear model is 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. Quasi-likelihood models that elicit expectiles do exist: see [38].

Asymptotic properties of the expectile regression estimators are well-studied: see [37], [23], [8]. Thus, this article is primarily concerned with finite sample properties in the classical framework. However, some components of our project have not been studied in the asymptotic framework. We discuss the variance of the residuals in the finite sample as well as asymptotically. Interestingly, neither of the "usual" estimators for residual variance are unbiased when adapted to expectile regression, except in special cases. We propose a third option and show that it is unbiased (at least when the estimator is feasible) and consistent.

The rest of the document is arranged as follows. Section 2 introduces expectiles formally an

[1] Generalized quantiles are defined as minimizers of asymmetrically weighted generalized loss functions, see Daouia et al. [16]. 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 [29].

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 estimator (or predictor) in very simple model misspecification problems. Section 5 discusses the feasibility of the expectile weights and explains how they may be feasible. Section 6 is devoted to the estimator of the variance parameter for a regression of this type and section 7 concludes.

2 Preliminaries

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

$$\mathbf{y} = \mathbf{X}\beta + \epsilon. \tag{2.1}$$

The data $\mathbf{y}, \mathbf{X}$ 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 $\mathbf{x}_i$, the linear coefficients β and the vector of error terms ϵ. Namely,

$$\begin{aligned}
\mathbf{y} &= [y_1, ..., y_n]', \\
\mathbf{X} &= [\mathbf{x}_1, ..., \mathbf{x}_n]', \\
\mathbf{x}_i &= [1, x_{i,2}, ..., x_{i,k}]', \\
\beta &= [\beta_1, ..., \beta_k]', \\
\epsilon &= [\epsilon_1, ..., \epsilon_n]'.
\end{aligned}$$

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 $\{y_i, \mathbf{x}_i\}$ are jointly i.i.d. but $F_{Y|X}$ is not necessarily 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. [8] or the closely related result in the working paper by Philipps [38]. The model conforms to the following assumption:

Assumption 0: The process $(Y, X) \coloneqq \{y_i, \mathbf{x}_i : \Xi \to \mathbb{R}^{k+1}, i = 1, ..., n\}$ is defined on a (joint) probability space $(\Xi, \mathcal{F}, P)$ where Ξ 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 more than two finite moments.

The *generalized quantiles* of a distribution F are a set of summary statistics indexed by $\tau \in (0, 1)$, similar in many ways to quantiles [26]. They are the class of minimizers

$$\theta_\tau(Y) = \arg\min_\theta \int |\tau - I(y < \theta)| \rho(y - \theta) dF(y) \tag{2.2}$$

obtained by minimizing the expected value of some objective function ρ 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 ρ, these produce the standard m-statistic when $\tau = .5$ and the τ^{th} m-quantile more generally [13]. When ρ is an L_p loss function,

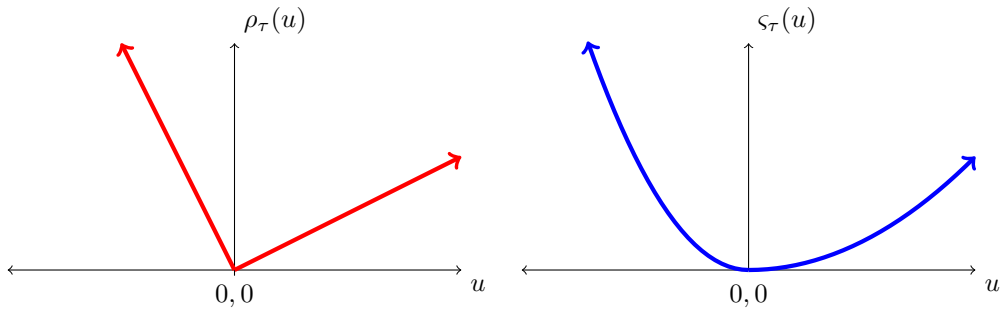

Figure 1: On the left, the L_1 quantile loss function. On the right, the L_2 quantile (expectile) loss function. The coefficient of asymmetry is $\tau = .2$ in both cases. In the symmetric case, the loss function at left produces the Least Absolute Deviation (LAD) estimator while the loss function at right produces Ordinary Least Squares (OLS).

$\rho(y-\theta) = \|y-\theta\|_p^p$, these are the L_p-quantiles of Chen [14]. The standard L_1 and L_2 quantile objective functions are shown in Figure 1.

2.1 Expectiles

The L_2-quantile loss function is an asymmetrically weighted least squares criterion and produces Expectiles. Because these are the only linear estimators[2] in the class of L_p-quantiles, they will be the focus of this document. Expectiles are weighted averages that range from the minimum of F to its maximum and nest the usual arithmetic mean at $\tau = .5$, exactly in the same way that quantiles nest the median. One definition of the τ^{th} expectile of a distribution, which is a special case of 2.2 on the preceding page, is

$$\mu_\tau(Y) := \arg\min_\theta \int \varsigma_\tau(y-\theta)dF(y). \tag{2.3}$$

Thus, Newey and Powell [37] show that expectiles can be obtained by minimizing a particular "swoosh" function

$$\varsigma_\tau(u) = \begin{cases} \tau u^2 & \text{if } u \geq 0 \\ (1-\tau)u^2 & \text{if } u < 0 \end{cases} \tag{2.4}$$

which is a piecewise quadratic loss function similar to the piecewise linear "check" function of [28]. As such, expectiles have been characterized primarily by their relationship to quantiles. However, the asymmetric least squares form of equation 2.4 makes evaluation of the τ^{th} linear sample quantiles into a nearly-standard weighted least squares problem: replace θ with $x'\beta$ and

[2] A "linear" estimator can be written as $\tilde{\beta} = Cy$ for some matrix C, i.e. it is linear in each y_i. The estimators corresponding to 2.2 on the previous page have first-order conditions

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

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 ρ is quadratic.

replace F with $\mathbb{F}(z) = n^{-1}\sum_{i=1}^{n} I(y_i - \mathbf{x}_i'\beta \geq z)$

$$\begin{aligned}\hat{\beta_\tau} &:= \arg\min_b \int \varsigma_\tau(y_i - \mathbf{x}_i'\beta) d\mathbb{F} \\ &= \arg\min_b \sum_{i=1}^{n} w_i(y_i - \mathbf{x}_i'\beta)^2 \end{aligned} \tag{2.5}$$

where the weights $w_i = \tau$ if $y_i \geq \mathbf{x}_i'\beta$ and $w_i = 1-\tau$ otherwise. The coefficients for the τ^{th} linear regression expectile adopt the subscript τ, say β_τ, following the notation of Koenker [26]. As a weighted least squares problem, the sum in equation 2.5 is $\hat{\epsilon}' W \hat{\epsilon}$, where W is the diagonal matrix $[W]_{ii} = w_i$. Naturally, the least asymmetric sum of squares estimator is $\hat{\beta_\tau} = (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{y}$ which Newey and Powell show to be consistent and asymptotically normal under reasonably general conditions. See Holzmann and Klar for more general asymptotics in the location model [23] or Barry et al. [8] or Philipps [38] for asymptotics of the generalized estimator under non-i.i.d. assumptions. Recent literature has added substantial context regarding the usefulness of these statistics in economics and finance: see [44], [45], or [52].

As the minimizer of the "swoosh" function in 2.4 and 2.5, the τ^{th} expectile of the distribution F has a dual interpretation. First, it can be expressed as a weighted average:

$$\begin{aligned}\mu_\tau(F) &= \int y\, \kappa w(y)\, dF(y) \\ \text{where } w(y) &= \begin{cases} \tau & y \geq \mu_\tau \\ (1-\tau) & \text{if } y < \mu_\tau \end{cases}\end{aligned}$$

and κ 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 2.4 does not vary over affine transforms of ς_τ, κ 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 $\tilde{F}$,

$$\begin{aligned}\mu_\tau(F) &= \int y\, d\tilde{F}(y) \\ \text{where } d\tilde{F} &= \kappa w(y)\, dF(y). \end{aligned} \tag{2.6}$$

This can be attributed to Breckling and Chambers [13], who produce the same fact for a more general family of estimators. Both the interpretation with respect to F and the same with respect to $\tilde{F}$ will play a role in later sections. For the asymmetrically τ-weighted expectations operator corresponding to any distribution F, we adopt the τ subscript for our notation

$$E_\tau(Y) = \int y\, \kappa w(y)\, dF(y) = \int y\, d\tilde{F}(y) = \mu_\tau(F). \tag{2.7}$$

We will use this notation[4] extensively. The expectile operator $E_\tau(\cdot)$ is clearly a special case of the usual expectations operator $E_F(\cdot)$ (and vice versa) and inherits all of its properties. In the case

[3] We say that the weights $w(y)$ are proper if $E(w(Y)) = 1$. The weights $\kappa w(Y)$ are proper by construction.

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

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

where $\tau = .5$, we have the usual expectations under F, so we suppress the subscript and simply write $E(\cdot)$.

The sample expectile fitted values (predictor) for $\mathbf{y}$ has the usual GLS form

$$\begin{aligned}\hat{\mathbf{y}_\tau} &= \mathbf{X}\hat{\beta}_\tau \\ &= \mathbf{X}(\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{y} \\ &= P_\tau\mathbf{y} \end{aligned} \tag{2.8}$$

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

$$\begin{aligned}\hat{\epsilon} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= \left(I - \mathbf{X}(\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\right)\mathbf{y} \\ &= M_\tau\mathbf{y} \end{aligned} \tag{2.9}$$

where the matrix M_τ is the expectile annihilator matrix. The notation is standard and GLS-type projections with this form are ubiquitous, but P_τ and M_τ differ slightly from their "symmetric" counterparts. They are idempotent but not symmetric, thus they are not *orthogonal* projections in the usual sense. Instead, a projection of this type is called an *oblique projection* [9, p. 165] [50, p. 578]. Oblique projections are rarely discussed *per se* in econometrics. However, they have been studied explicitly in other applied sciences, such as signal processing [10]. Some details regarding P_τ and M_τ are given in the appendix and several of the implications of their structure are discussed in Section 6.

2.2 Example: Mexican Repatriation

In Figure 2 on the next page, the L_1 quantiles of Koenker and Bassett [28] are compared to the L_2 quantiles of Newey and Powell [37]. The data for this example are drawn from the Mexican repatriation during the great depression era of the 1930's. During that period of time, there was a substantially hostile attitude in the southwest of the United States towards Mexican nationals living north of the border. Organized labor and political groups sponsored, pushed, and sometimes forced Mexicans and Mexican Americans to leave the U.S. and return to Mexico. As a result of harassment and other targeted campaigns, approximately 1.3 million persons were repatriated (returned to Mexico) during that episode [31]. See the book by Balderrama and Rodriguez [7] for a detailed overview.

The data in the figure are the Mexican repatriation intensity, defined as the proportional decrease in the number of Mexican nationals in a given municipality from 1930 to 1940, according to the U.S. Census. Figure 2 on the following page plots the repatriation intensity relative to the proportion of the population that were Mexican nationals in 1930 per locus. On the left, regression

where $\psi_\tau(y) = \tau - I(y < 0)$, which is the derivative of Koenker and Bassett's "check" function $\rho_\tau(u)$ [28], [26, p. 36]. The absolute value of ψ_τ is sometimes called the check function, see [8]. 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 τ otherwise.

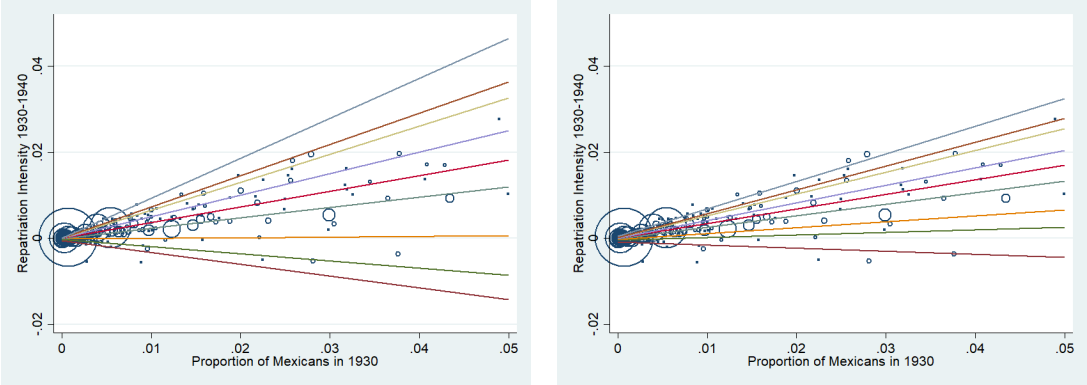

Figure 2: On the left, linear regression L_1-quantiles. On the right, linear regression L_2 quantiles. Repatriation intensity is defined as the percentage decrease in population of Mexican nationals from 1930-40. The covariate is the proportion of the population that were Mexican nationals in 1930.

quantiles are shown. On the right, regression expectiles are shown. The general positive correlation between repatriation (departures) and the proportion of Mexicans in 1930 is statistically significant, suggesting that repatriation or harassment campaigns were more intense in those localities. This result can be found using Ordinary Least Squares (the red line on the right) or median regression (the red line on the left).

The notable feature of the scatterplot in Figure 2 is the cone-shaped heteroscedasticity. The variance of the repatriation variable increases with the proportion of Mexicans in 1930. Because the location-scale is apparently quite reasonable in this example, we could take any L_1 regression line from the left and find a τ such that the L_2 regression line on the right is the same asymptotically (or vice versa). Waltrup, Sobotka, and Kneib have recently developed an effective method for doing this [45] and Daouia et al. [16] have extended it. In this application, expectiles with the same subset of τ's are slightly closer together than quantiles, which is typical [27, 11]. The econometrician's choice of estimator is dependent on what finite-sample properties are desirable.

Importantly, the example in Figure 2 contains major economic results that cannot be found using mean or median regression alone. The obvious result implied by the central regression estimates is that symptoms of racial animus are experienced more profoundly in areas where the racial minority is most present. However, higher (lower) quantiles of the conditional distribution correspond to those areas where these effects are stronger (weaker), ceteris paribus. Compare the L_1 quantile regression coefficients in Table 1 on the next page with the expectile regression coefficients in Table 2 on page 9.

Notably, the mean and median regression produce generally similar results with statistically coefficients of .34 and .36, respectively. Extending our survey to measures of non-central tendency, we see larger coefficients as τ increases and smaller coefficients as τ decreases. Given that the effect seems to be strongest at the top of the distribution, the obvious question is how severe that effect may be in that part of the sample. The quantile line of best fit at $\tau = .9$ is approximately co-located with the expectile line of best fit at $\tau = .99$ with coefficients of approximately .65 and .64, respectively. This indicates that the effect of the population demographics is approximately twice as large near the maximum of the data as it is at the mean. Because the mean effect represents

Quantile Regression Coefficients

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Intensity	Intensity	Intensity	Intensity	Intensity
$\tau =$	.005	.01	.05	.1	.3
Proportion of Mexicans	-0.273	-0.168**	0.0154	0.240***	0.364***
	(0.258)	(0.0661)	(0.0904)	(0.0327)	(0.0230)
Constant	-0.000517***	-0.000233***	-0.000153***	-3.86e-05***	0
	(7.80e-05)	(3.81e-05)	(2.09e-05)	(1.02e-05)	(4.97e-06)

	(6)	(7)	(8)	(9)	(10)
VARIABLES	Intensity	Intensity	Intensity	Intensity	Intensity
$\tau =$	.5	.7	.9	.95	.99
Proportion of Mexicans	0.364***	0.500***	0.651***	0.726***	0.926***
	(0.0230)	(0.0396)	(0.0336)	(0.0313)	(0.118)
Constant	0	0	3.18e-05***	4.68e-05***	3.54e-05
	(4.97e-06)	(6.25e-06)	(7.18e-06)	(1.54e-05)	(6.22e-05)

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

Table 1: Quantile Regression coefficients for the regression of repatriation intensity on the proportion of Mexicans in a given locale. Standard errors are the heteroscedasticity-robust sandwich estimator given in Theorem 4.1 in [26]. The negative coefficient at $\tau = .01$ is weakly significant.

Expectile Regression Coefficients

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Intensity	Intensity	Intensity	Intensity	Intensity
$\tau =$	.001	.01	.05	.1	.3
Proportion of Mexicans	-0.0673*	-0.0697**	0.0577	0.138**	0.267***
	(0.0376)	(0.0308)	(0.0604)	(0.0538)	(0.0355)
Constant	-0.00333***	-0.000889***	-0.000357**	-0.000225**	-7.51e-05
	(0.00100)	(0.000320)	(0.000145)	(9.06e-05)	(4.70e-05)
R-squared	0.077	0.064	0.031	0.171	0.524

	(6)	(7)	(8)	(9)	(10)
VARIABLES	Intensity	Intensity	Intensity	Intensity	Intensity
$\tau =$	.5	.7	.9	.95	.99
Proportion of Mexicans	0.340***	0.409***	0.507***	0.550***	0.640***
	(0.0323)	(0.0315)	(0.0260)	(0.0232)	(0.0268)
Constant	-1.43e-05	4.03e-05	0.000162**	0.000264***	0.000447***
	(4.06e-05)	(4.85e-05)	(6.45e-05)	(7.19e-05)	(0.000133)
R-squared	0.664	0.747	0.842	0.874	0.893

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

Table 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 $\frac{n}{n-k}$.

an average of different quantile effects (see the note by Subrahmanyam [42]) we can determine that the result is driven more by these locations at the top of the distribution.

In sharp contrast, we see both quantiles and expectiles produce *negative* and statistically significant coefficients near the bottom of the distribution ($\tau = .05$ for L_1 quantiles and $\tau = .01$ for expectiles). This is arguably as large a result as the OLS coefficient itself. The bottom of the distribution corresponds to those "least racist" municipalities. Not only do we fail to observe Mexican nationals fleeing those areas, we find statistically significant evidence that their populations are growing relative to other demographic groups. Thus it is possible that Mexican nationals are moving from municipalities at the top of the distribution to municipalities at the bottom.

These generalized quantile regressions have the power to identify results that would otherwise be missed. However, a major question is which estimator is "optimal" in some standard sense. For example, it is obvious that the heteroscedasticity in this application would point towards a generalized least squares estimator, rather than OLS, for the mean regression. This result can be extended to expectiles. In that way, we will develop the BLUE for the generalized quantile family in the next section. Similarly, we will show that the expectile (L_2-quantile) has an interpretation as an expected value under certain conditions and may be the optimal estimator of the mean under other conditions. Other corollary results will follow.

3 Expectile (Generalized) Least Squares

In this section, we will derive the expectile regression in a modified form of the classical Gauss-Markov framework. For any general linear model, the estimated expectile regression coefficients have the form

$$\hat{\beta_\tau} = (\mathbf{X}'A\mathbf{X})^{-1}\mathbf{X}'A\mathbf{y} \tag{3.1}$$

for some matrix A. This generalized least squares form is common in the literature. 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 3.1 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 [37] nests Ordinary Least Squares when $\tau = .5$. Otherwise, "Expectile Regression" refers to the concept generally.

In the following subsections, we treat τ as known or assume that multiple τ are of interest, as is the case when expectiles are being treated similarly to quantiles. In section 4, we consider the possibility that a particular τ is optimal for specific model designs.

3.1 Expectile Gauss-Markov Assumptions

As in the previous section, we have a sample of data observations $y_i, \mathbf{x}_i$ for $i = 1, ..., n$. Denote the column vector made from $\{y_i\}_{i=1}^n$ as $\mathbf{y}$ and the matrix made from each $\mathbf{x}_i'$ as $\mathbf{X}$. The weights w_i and the asymmetric expectations operator E_τ are defined as in the previous section.

The relationship between the following four assumptions and the classical Gauss-Markov assumptions is readily apparent: simply let $\tau = .5$ and the expectile assumptions nest the classical Gauss-Markov assumptions. We will address that relationship more thoroughly in Section 4. A

notable result is that the first and third assumptions (linearity and full rank) are the same for all expectiles, the second and fourth vary depending on which expectile is of interest. The fourth (spherical variance-covariance) is usually not taken seriously: a more general heteroscedastic version is given in Section 3.2.2.

The first assumption is obvious: the model for the τ^{th} conditional expectile of y_i given $\mathbf{x}_i$ will be linear.

Assumption 1: The model is linear.

$$\begin{aligned} y_i &= \beta_{\tau,1} x_{i,1} + ... + \beta_{\tau,k} x_{i,k} + \epsilon_i \\ &= \mathbf{x}_i' \beta_\tau + \epsilon_i \end{aligned} \tag{3.2}$$

This buys us a convenient interpretation of each coefficient as the partial derivative of the τ^{th} expectile of y_i with respect to the corresponding covariate:

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

See Stahlschmidt et al. [40] for some discussion regarding these "expectile treatment effects". Nonparametric regression models for expectiles are reasonably common in the literature, [48] but are not part of the current subject matter.

Assumption 2: Weighted Strict Exogeneity:

$$E_\tau(\epsilon_i|\mathbf{X}) = 0 \quad \text{or} \quad E(w_i \epsilon_i|\mathbf{X}) = 0 \quad \forall i \in \{1, ..., n\} \ . \tag{3.3}$$

Then the weighted error term is orthogonal to the data $\mathbf{X}$. This can be re-formulated, using $\epsilon_i = y_i - \mathbf{x}_i' \beta_\tau$ from equation 3.2, as

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

Of course, we can also simplify the latter term by requiring the $E(w_i|\mathbf{X}) = 1$, conforming to our definition of "proper" weights in footnote 4.

You can also see that equation 3.3 implies some other trivial results. For instance,

$$E_\tau(x_{ij}\epsilon_i) = 0 \quad \text{or} \quad E(w_i x_{ij} \epsilon_i) = 0 \quad \forall i, j \ . \tag{3.4}$$

This follows directly from the tower rule[5]:

$$\begin{aligned} E_\tau(x_{ij}\epsilon_i) &= E(E_\tau(x_{ij}\epsilon_i|X)) \\ &= E(x_{ij} E_\tau(\epsilon_i|X)) \\ &= 0. \end{aligned} \tag{3.5}$$

[5] 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 τ appears only once on the right side of equation 3.5 because writing $E_\tau(E_\tau(\cdot))$ would imply applying the weights twice.

The conditional moment in equation 3.3 also implies the value of the unconditional moment by the tower rule:

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

This illustrates the recurring theme that expectile regression will conform, in most cases, to the same set of mathematical properties that central L_2 estimators, including the entire GMM family, will posess. Next we present the third assumption.

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

If Assumption 3 is violated, then the matrix $\mathbf{X}'W\mathbf{X}$ will not be invertible and the expectile regression coefficients cannot be evaluated. Because $rank(AB) \leq min(rank(A), rank(B))$, invertibility of $\mathbf{X}'W\mathbf{X}$ requires that the data matrix $\mathbf{X}$ has full rank. This is standard for ordinary least squares, but in the expectile case full rank is also sufficient to ensure that $\mathbf{X}'W\mathbf{X}$ is invertible. To show that this is true, observe that a unique Moore-Penrose pseudoinverse $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ exists for any matrix $\mathbf{X}$ with full column rank. Then certainly $\mathbf{X}'\mathbf{X}$ 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 $W^{1/2}\mathbf{X}$ is the $n \times k$ matrix $[W^{1/2}\mathbf{X}]_{i,j} = W_{ii}^{1/2}[\mathbf{X}]_{i,j}$. If this matrix is not of full rank, then there is some vector $v \in \mathbb{R}^k$ such that $\sum_{j=1}^k W_{ii}^{1/2}[\mathbf{X}]_{i,j}r_j = 0$ for all i, which implies $\sum_{j=1}^k [\mathbf{X}]_{i,j}r_j = 0$ for all i, which is impossible because $\mathbf{X}$ has full rank. Then $W^{1/2}\mathbf{X}$ has a unique Moore-Penrose pseudoinverse $(\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W^{1/2}$ whenever $\mathbf{X}$ is of full rank, and $\mathbf{X}'W\mathbf{X}$ is invertible.

Assumption 3 is notable in the sense that it does not matter which expectile we are modelling: the assumption either holds for all expectiles or it holds for none. Compare this to the major result in Newey and Powell [37], where the τ^{th} expectile of the distribution exists if and only if its first moment exists. In addition, Assumption 3 requires that there are at least k observations; $n \geq k$. If $n = k$, then the equation

$$y = \mathbf{X}\beta_\tau$$

has an unique, "exact" solution for $\beta_\tau = \mathbf{X}^{-1}\mathbf{y}$ because $\mathbf{X}$ is square matrix with full rank (and thus invertible). If $n > k$, then we might consider using the same Moore-Penrose pseudoinverse of $\mathbf{X}$, which is $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ to obtain a solution: $\beta_\tau = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. 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 $(\mathbf{X}'A\mathbf{X})^{-1}\mathbf{X}'A$, 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 $(\mathbf{X}'A\mathbf{X})^{-1}\mathbf{X}'A$ can be used to produce an "estimate" of β_τ. We will produce a unique choice set from this family in sections 7 and 3.2. To do this, we make one more assumption about the variance of our model.

Assumption 4: "Spherical" error variance or Asymmetric Homoscedasticity:

$$E_\tau(\epsilon_i^2|\mathbf{X}, W) = \kappa\nu^2 \quad \text{or} \quad E(w_i\epsilon_i^2|\mathbf{X}, W) = \nu^2 \quad \forall i \ . \tag{3.6}$$

and

$$E_\tau(\epsilon_i\epsilon_j|\mathbf{X}, W) = 0 \quad \text{or} \quad E(w_i\epsilon_i\epsilon_j|\mathbf{X}, W) = 0 \quad \forall i \neq j \ . \tag{3.7}$$

These statements may also be written as $E(W\epsilon\epsilon'|\mathbf{X}, W) = \nu^2 I_n$. Assumption 4 implies that each ϵ_i has the same *weighted* variance, i.e. the conditional variance (given that the error term is positive or negative) varies by an *a priori* known ratio[6]. It also implies that each pair of distinct ϵ_i, ϵ_j are uncorrelated. This is "spherical" variance in the sense of a sphere defined under a weighted distance function. It is useful also to note that the weighted covariance in equation 3.7 uses weights which vary depending on ϵ_i; though ϵ_i and ϵ_j may be reversed and the statement remains true.

Importantly, the assumption that $E(W\epsilon\epsilon') = \nu^2 I_n$ implies that the ratio of variance for positive errors to variance for negative errors is $\frac{1-\tau}{\tau}$. Because the τ^{th} expectile is not the mean, the distribution of residuals is "skew" in precisely this way and is not zero on average. However, this is not the usual definition of skewness as we have not employed any third-moment information.

As with the previous two 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 ϵ_i given $\mathbf{X}$ as

$$\begin{aligned} WVar(\epsilon_i|\mathbf{X}) &\coloneqq E_\tau\left((\epsilon_i - E_\tau(\epsilon_i|\mathbf{X}))^2|\mathbf{X}\right) \\ &= E_\tau\left(\epsilon_i^2 - 2\epsilon_i E_\tau(\epsilon_i|\mathbf{X}) + E_\tau(\epsilon_i|\mathbf{X})^2|\mathbf{X}\right) \\ &= E_\tau(\epsilon_i^2|\mathbf{X}) - E_\tau(\epsilon_i|\mathbf{X})^2. \end{aligned} \tag{3.8}$$

This is similar to the usual expression given for variance; $Var(\epsilon_i|\mathbf{X}) = E(\epsilon_i^2|\mathbf{X}) - E(\epsilon_i|\mathbf{X})^2$, with the addition of weights i.e. with the definition of variance modified to use $\tilde{F}$ in place of the original distribution F. Also, with weighted strict exogeneity (Assumption 2) we have $E_\tau(\epsilon_i|\mathbf{X})^2 = 0$, so $WVar(\epsilon_i|\mathbf{X}) = E_\tau(\epsilon_i^2|\mathbf{X})$. For research based on the standard case where $\tau = .5$, these two small results are ubiquitous. See section 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 and proper such that $W = I_n$ uniquely. Then assumptions 1-4 are the Gauss-Markov assumptions.*

We allow this proposition to serve as our 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 $\hat{\beta}_\tau = (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{y}$ has the following properties:*

$$\begin{aligned} E(\hat{\beta}_\tau|\mathbf{X}, W) &= \beta_\tau \\ Var(\hat{\beta}_\tau|\mathbf{X}, W) &= \nu^2 \left(\mathbf{X}'W\mathbf{X}\right)^{-1} \end{aligned}$$

[6]To assist the reader, we would like to draw attention to the relationship between ν^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 $\frac{1}{n-k}\sum_{i=1}^n w_i\hat{\epsilon_i}^2$, which we will recommend as an estimator for ν^2 when the variance of $\hat{\beta_\tau}$ or the predictor is of interest. In section 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.

Thus, the estimator is unbiased with variance $\nu^2 (\mathbf{X}'W\mathbf{X})^{-1}$. A full derivation of this result is given in Appendix A1, but it is identical to the standard result for GLS estimators. See [20], for instance. Next, we will show that the estimator $\hat{\beta}_\tau$ is the BLUE.

3.2 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 [34] states that ordinary least squares is the BLUE under assumptions 1-4 with $\tau = .5$. This result is typically extended to the Generalized Least Squares estimator (GLS) of Aitken [1] under heteroscedasticity.

3.2.1 With Spherical Variance-Covariance

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 $\hat{\beta}_\tau = (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{y}$. We may write any such linear estimator as $\tilde{\beta}_\tau = Cy$ for some choice of matrix C, potentially a function of $\mathbf{X}, W$. For our estimator$\tilde{\beta}_\tau$, we have

$$\begin{aligned} E(\tilde{\beta}_\tau|\mathbf{X},W) &= E(Cy|\mathbf{X},W) \\ &= E(C\mathbf{X}\beta_\tau + C\epsilon|\mathbf{X},W) \\ &= \underbrace{E(C\mathbf{X}\beta_\tau|\mathbf{X},W)}_{C\mathbf{X}\beta_\tau} + \underbrace{E(C\epsilon|\mathbf{X},W)}_{0} \\ &= C\mathbf{X}\beta_\tau. \end{aligned}$$

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

$$\begin{aligned} I = C\mathbf{X} &= D\mathbf{X} + (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{X} \\ &= D\mathbf{X} + I. \end{aligned} \tag{3.9}$$

Then $D\mathbf{X}$ is uniquely zero. Now write the conditional variance of $\tilde{\beta}_\tau = Cy$ using this same decomposition:

$$\begin{aligned} Var(\tilde{\beta}_\tau|\mathbf{X},W) &= E\left((\tilde{\beta}_\tau - \beta_\tau)(\tilde{\beta}_\tau - \beta_\tau)'|\mathbf{X},W\right) \\ &= E\left(C\epsilon\epsilon'C'|\mathbf{X},W\right) \\ &= CE\left(\epsilon\epsilon'|\mathbf{X},W\right)C' \\ &= \left(D + (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\right)\nu^2 W^{-1}\left(D + (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\right)' \\ &= \nu^2\left(DW^{-1}D' + (\mathbf{X}'W\mathbf{X})^{-1} + DW^{-1}W\mathbf{X}(\mathbf{X}'W\mathbf{X})^{-1} + (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'WW^{-1}D\right) \\ &= \nu^2\left(DW^{-1}D' + (\mathbf{X}'W\mathbf{X})^{-1}\right). \end{aligned} \tag{3.10}$$

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

$$\begin{aligned} Var(\tilde{\beta_\tau}|\mathbf{X}, W) &= \nu^2 \left(DW^{-1}D' + (\mathbf{X}'W\mathbf{X})^{-1} \right) \\ &\geq \nu^2 (\mathbf{X}'W\mathbf{X})^{-1} = Var(\hat{\beta_\tau}|\mathbf{X}, W). \end{aligned}$$

That is, any unbiased linear estimator has at least as much variance as $\hat{\beta_\tau}$. We have just proven the proposition below.

Proposition 3. *Under assumptions 1 through 4, the expectile regression estimator $\hat{\beta_\tau} = (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{y}$ is the best linear unbiased estimator in the sense that it has the least variance.*

Next, we will ask whether the estimator $\hat{\beta_\tau}$ is useful to construct a predictor of $\mathbf{y}$ or of any (known) linear function of y; say $A\mathbf{y}$ for A possibly but not necessarily I_n. The τ^{th} expectile of $A\mathbf{y}$ given $\mathbf{X}$ is

$$\begin{aligned} E_\tau(A\mathbf{y}|\mathbf{X}, W) &= E_\tau(A\mathbf{X}\beta_\tau + A\epsilon|\mathbf{X}) \\ &= E_\tau(A\mathbf{X}\beta_\tau|\mathbf{X}) + \underbrace{E_\tau(A\epsilon|\mathbf{X})}_{0} \\ &= A\mathbf{X}\beta_\tau. \end{aligned}$$

Clearly, we can construct a predictor of $A\mathbf{y}$ using any estimate we may have for β_τ. If we use the unbiased estimator $\hat{\beta_\tau} = (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{y}$, we have

$$\begin{aligned} E(A\hat{\mathbf{y}}|\mathbf{X}, W) &= E(A\mathbf{X}\hat{\beta_\tau}|\mathbf{X}, W) \\ &= A\mathbf{X}\beta_\tau \end{aligned}$$

Thus, $A\hat{\mathbf{y}}$ is an unbiased predictor of the τ^{th} expectile of $A\mathbf{y}$. Its variance is given by

$$\begin{aligned} Var(A\hat{\mathbf{y}}|\mathbf{X}, W) &= E\left((A\hat{\mathbf{y}} - E(A\hat{\mathbf{y}}|\mathbf{X}, W))(A\hat{\mathbf{y}} - E(A\hat{\mathbf{y}}|\mathbf{X}, W))'|\mathbf{X}, W\right) \\ &= E\left((A\hat{\mathbf{y}} - A\mathbf{X}\beta_\tau)(A\hat{\mathbf{y}} - A\mathbf{X}\beta_\tau)'|\mathbf{X}, W\right) \\ &= A\mathbf{X}E\left((\hat{\beta_\tau} - \beta_\tau)(\hat{\beta_\tau} - \beta_\tau)'|\mathbf{X}, W\right)\mathbf{X}'A' \\ &= \nu^2 A\mathbf{X}(\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'A'. \end{aligned}$$

For any other unbiased estimator $\tilde{\beta_\tau}$ as given before, we have the predictor $A\tilde{\mathbf{y}} = A\mathbf{X}\tilde{\beta_\tau}$ and $E(A\tilde{\mathbf{y}}|\mathbf{X}, W) = A\mathbf{X}\beta_\tau$. But the variance of $A\tilde{\mathbf{y}}$ is

$$\begin{aligned} Var(A\tilde{\mathbf{y}}|\mathbf{X}, W) &= E\left((A\tilde{\mathbf{y}} - E(A\tilde{\mathbf{y}}|\mathbf{X}, W))(A\tilde{\mathbf{y}} - E(A\tilde{\mathbf{y}}|\mathbf{X}, W))'|\mathbf{X}, W\right) \\ &= A\mathbf{X}E\left((\tilde{\beta_\tau} - \beta_\tau)(\tilde{\beta_\tau} - \beta_\tau)'|\mathbf{X}, W\right)\mathbf{X}'A' \\ &= A\mathbf{X}\left(D + (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\right)\nu^2 W^{-1}\left(D + (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\right)'\mathbf{X}'A'. \\ &= \nu^2\left(A\mathbf{X}DW^{-1}D'\mathbf{X}'A' + A\mathbf{X}(\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'A'\right) \\ &\geq \nu^2 A\mathbf{X}(\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'A' = Var(A\hat{\mathbf{y}}|\mathbf{X}, W) \end{aligned}$$

Again, the positive-definite matrix $A\mathbf{X}DW^{-1}D'\mathbf{X}'A'$ shows that our variance is at least as great as before. By extension, we see that $A\mathbf{X}\hat{\beta_\tau}$ is the best linear predictor of $A\mathbf{X}\beta_\tau$ for a linear function $B = (A\mathbf{X})$ of β_τ. This proves the following proposition.

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

In particular, we might note that A could be any one of the elementary basis vectors such that $A\mathbf{y} = y_i$ and $A\mathbf{X}\beta = \mathbf{x}_i'\beta_\tau$. Then we have also proven the next proposition.

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

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

3.2.2 With Heteroscedasticity

The typical approach to "BLUE" regression under heteroscedasticity is to look for a transformation of the data such that the Gauss-Markov assumptions are applicable. Let assumptions 1 through 3 hold, but let $E(W^{1/2}\epsilon\epsilon'W^{1/2}|\mathbf{X}, W) \neq \nu^2 I_n$. Instead, let[7]

$$E(\epsilon\epsilon'|\mathbf{X}, W) = \nu^2\Sigma = \nu^2 W^{-1/2}\Omega W^{-1/2} \tag{3.11}$$

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

Because Σ 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 Σ^{-1}. In the usual fashion, left-multiply the entire model by V and say

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\beta_\tau + \tilde{\epsilon} \equiv V\mathbf{y} = V\mathbf{X}\beta_\tau + V\epsilon$$

It is trivial to show that assumptions 1 and 3 apply to this transformed model: clearly it is linear, $\tilde{\mathbf{X}}$ has full rank. But also,

$$\begin{aligned} E(\tilde{\epsilon}\tilde{\epsilon}'|\tilde{\mathbf{X}}, W) &= E(V\epsilon\epsilon'V'|\mathbf{X}, W) \\ &= VE(\epsilon\epsilon'|\mathbf{X}, W)V' \\ &= \nu^2 V\Sigma V' \\ &= \nu^2 I_n \end{aligned}$$

Because we have already incorporated the expectile weights in equation 3.11, this has become a perfectly standard GLS problem. We have

$$\begin{aligned} \hat{\beta}_{\tau,GLS} &= \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} \\ &= \left(\mathbf{X}'V'V\mathbf{X}\right)^{-1}\mathbf{X}'V'V\mathbf{y} \\ &= \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y} \\ &= \beta_\tau + \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\epsilon \end{aligned}$$

[7] Notice $E(w_i\epsilon_i^2|\mathbf{X}, W) = \nu^2\Omega_{ii}$; the elements of Ω are proportional to the expected squared error.

where, because the sequence $\{Y, X\}$ is independent,

$$\begin{aligned}
&E\left(\left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\epsilon|\mathbf{X}, W\right) \\
&= \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'E\left(W^{1/2}\Omega^{-1}W^{1/2}\epsilon|\mathbf{X}, W\right) \\
&= \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Omega^{-1}\underbrace{E\left(W\epsilon|\mathbf{X}, W\right)}_{0}.
\end{aligned}$$

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 $\nu^2\left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}$.

The proof is similar to the example in the previous section. Because any unbiased linear estimator can be written $\tilde{\beta_\tau} = C\mathbf{y}$ where $CX = I_n$ and $C = D + \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)\mathbf{X}'\Sigma^{-1}$, $D\mathbf{X} = 0$, we can write

$$\begin{aligned}
Var(C\mathbf{y}|\mathbf{X}, W) &= Var(((D + \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)\mathbf{X}'\Sigma^{-1})\mathbf{y}|\mathbf{X}, W) \\
&= \nu^2\left(D + \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\right)\Sigma\left(D + \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\right) \\
&= \nu^2 D\Sigma D' + \nu^2\left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\Sigma\Sigma^{-1}\mathbf{X}\left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1} \\
&= \nu^2 D\Sigma D' + \nu^2\left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1} \\
&\geq \nu^2\left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1} = Var(\hat{\beta_{\tau,GLS}})
\end{aligned} \tag{3.12}$$

Thus we have shown that the GLS estimator is the "best" linear unbiased estimator for β_τ in the sense of minimum variance. This serves as a proof of the following proposition.

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

$$\hat{\beta_{\tau,GLS}} = \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y} \tag{3.13}$$

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

We will explore this result further in the next part. Obviously, we have the corollary result that the linear predictor $A\mathbf{X}\hat{\beta}_{\tau,GLS}$ is the best unbiased linear predictor of $A\mathbf{y}$ and that $\mathbf{x}_i'\hat{\beta}_{\tau,GLS}$ is the best unbiased linear predictor of y_i. We leave that proof to the reader.

To conclude this section, we restate two 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 [1] is BLUE under the assumption of heteroscedasticity of known form with the same ratio of variance for positive and negative errors.

4 Expectiles in Misspecified OLS Regressions

As stated in Proposition 1, the four expectile assumptions nest the four Gauss-Markov assumptions when $\tau = .5$. We have shown that the expectile regression coefficients $\hat{\beta_\tau} = (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{y}$

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.

However, we can also 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 a particular nonstandard way. In the following three sections, we show three separate cases where Gauss-Markov assumptions are violated and a generalized expectile regression is the BLUE.

First, let us restate the Gauss-Markov four assumptions for the context of a typical mean regression. Because we are not estimating the τ^{th} expectile specifically, we drop the τ subscript from β.

G-M Assumption 1: The model is linear.

$$\begin{aligned} y_i &= \beta_1 x_{i,1} + ... + \beta_k x_{i,k} + \epsilon_i \\ &= \mathbf{x}_i' \beta + \epsilon_i \end{aligned} \tag{4.1}$$

G-M Assumption 2: Strict Exogeneity:

$$E(\epsilon_i | \mathbf{X}) = 0 \quad \text{or} \quad \forall i \in \{1, ..., n\} \ . \tag{4.2}$$

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

G-M Assumption 4: Spherical variance-covariance

$$E(\epsilon\epsilon' | \mathbf{X}) = \sigma^2 I_n. \tag{4.3}$$

We have shown that the first and third assumptions do not change depending on which expectile we are interested in. We have shown that the estimator is usable and has known variance when the fourth assumption is violated (see equation 7.4). 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.

4.1 Expectiles: Relaxed Exogeneity

To step away from the usual mean regression, relax assumption 2 in the smallest way possible:

$$E(\epsilon_i | \mathbf{X}) = c \ \forall i. \tag{4.4}$$

We have merely altered our assumption so that the average error term is not zero, but some other constant $c \in \mathbb{R}$. This seems odd at first, because the mean regression will produce residuals with mean zero by design. But there are reasons why nonzero residuals are sometimes desiarable. Naturally, the mean regression line is not the only line of interest. This is especially true in heteroscedastic models, where we see that the rest of the regression coefficients $\beta_2, ...\beta_k$ vary with the estimated constant. This was shown in Figure 2 on page 7. Similarly, the regression residuals may be nonzero on average after a data trimming procedure to remove data from either the right or left of the data [24]. This is a well-known procedure to increase robustness of our estimators, but it also changes the mean of ϵ_i by a potentially large amount. Models with nonzero residual means

occur also in survival analysis. Suffice it to say that there are scenarios where we wish to include a constant in our regression *and* keep the error terms different from zero on average.

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 4.4. The non-zero exogeneity condition implies the following about our model:

Proposition 7. *Let Assumptions 1,3,4 hold, let* $E(\epsilon_i|\mathbf{X}) = c$ *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 $E(w_i\epsilon_i|\mathbf{X}) = 0$, *for some* $\tau \in [0,1]$

Proof. We know that

$$\begin{aligned} E(\epsilon_i|\mathbf{X}) &= E(\epsilon_i|\mathbf{X}, \epsilon_i \geq 0)\Pr(\epsilon_i \geq 0) + E(\epsilon_i|\mathbf{X}, \epsilon_i < 0)\Pr(\epsilon_i < 0) \\ &= c \end{aligned} \tag{4.5}$$

So

$$\begin{aligned} E(w_i\epsilon_i|\mathbf{X}) &= \tau E(\epsilon_i|\mathbf{X}, \epsilon_i \geq 0)\Pr(\epsilon_i \geq 0) + (1-\tau)E(\epsilon_i|\mathbf{X}, \epsilon_i < 0)\Pr(\epsilon_i < 0) \\ &= E(\epsilon_i|\mathbf{X}) - (1-\tau)E(\epsilon_i|\mathbf{X}, \epsilon_i \geq 0)\Pr(\epsilon_i \geq 0) - \tau E(\epsilon_i|\mathbf{X}, \epsilon_i < 0)\Pr(\epsilon_i < 0) \\ &= c - (1-\tau)E(\epsilon_i|\mathbf{X}, \epsilon_i \geq 0)\Pr(\epsilon_i \geq 0) - \tau E(\epsilon_i|\mathbf{X}, \epsilon_i < 0)\Pr(\epsilon_i < 0). \end{aligned} \tag{4.6}$$

The last expression above will be zero if

$$\begin{aligned} c &= (1-\tau)E(\epsilon_i|\mathbf{X}, \epsilon_i \geq 0)\Pr(\epsilon_i \geq 0) + \tau E(\epsilon_i|\mathbf{X}, \epsilon_i < 0)\Pr(\epsilon_i < 0) \\ &= E(\epsilon_i|\mathbf{X}, \epsilon_i \geq 0)\Pr(\epsilon_i \geq 0) + \tau\left(E(\epsilon_i|\mathbf{X}, \epsilon_i < 0)\Pr(\epsilon_i < 0) - E(\epsilon_i|\mathbf{X}, \epsilon_i \geq 0)\Pr(\epsilon_i \geq 0)\right), \end{aligned}$$

so the unique τ that satisfies this condition must be

$$\tau = \frac{E(\epsilon_i|\mathbf{X}, \epsilon_i \geq 0)\Pr(\epsilon_i \geq 0) - c}{E(\epsilon_i|\mathbf{X}, \epsilon_i \geq 0)\Pr(\epsilon_i \geq 0) - E(\epsilon_i|\mathbf{X}, \epsilon_i < 0)\Pr(\epsilon_i < 0)}. \tag{4.7}$$

But, because of equation 4.5, we can replace c and write τ as

$$\tau = \frac{-E(\epsilon_i|\mathbf{X}, \epsilon_i < 0)\Pr(\epsilon_i < 0)}{E(\epsilon_i|\mathbf{X}, \epsilon_i \geq 0)\Pr(\epsilon_i \geq 0) - E(\epsilon_i|\mathbf{X}, \epsilon_i < 0)\Pr(\epsilon_i < 0)}. \tag{4.8}$$

Clearly this has the form $\frac{a}{a+b}$ for positive a, b; therefore τ must be in $(0,1)$. □

The proposition above leads to an interesting outcome. If strict exogeneity fails as in equation 4.4 on the preceding page, and $E(\epsilon_i|\mathbf{X}) = c$, then there must be a $\tau \in (0,1)$ such that expectile weighted exogeneity holds[8].

$$E(\epsilon_i|\mathbf{X}) \neq 0, \quad \text{but} \quad E_\tau(\epsilon_i|\mathbf{X}) = E(w_i\epsilon_i|\mathbf{X}) = 0. \tag{4.9}$$

[8] This, as in the proposition, excludes the degenerate case where *all* ϵ_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.

Obviously, we find that the OLS estimator is not BLUE in this case–it is biased. However, there is a $\tau \in (0,1)$ (the τ in equation 4.8) such that the τ^{th} expectile regression estimator is unbiased. These are the following two propositions.

Proposition 8. *Let Gauss-Markov assumptions 1,3,4 hold and $E(\epsilon_i|\mathbf{X}) = c$ for all $i = 1, ..., n$. The OLS estimator is biased if $c \neq 0$.*

Proof. The proof is obvious. From the definition, we have

$$\begin{aligned} E(\hat{\beta}_{OLS}|\mathbf{X}) &= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}|\mathbf{X}) \\ &= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \epsilon)|\mathbf{X}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\epsilon|\mathbf{X}) \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'1_n c \end{aligned} \tag{4.10}$$

and for $c \neq 0$, the latter term is not zero unless $\mathbf{X} = 0_{n\times k}$, which violates Assumption 3. □

However, the $\hat{\beta}_\tau = (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{y}$ with expectile weights given by τ in equation 4.7 is unbiased. This follows from the fact that $E_\tau(\epsilon_i|\mathbf{X}) = 0$ as above, and the proof is the same as that given in equation 7.3 on page 49. We have the following propostion.

Proposition 9. *Let Gauss-Markov assumptions 1,3,4 hold and $E(\epsilon_i|\mathbf{X}) = c$. Then the expectile regression estimator $\hat{\beta}_\tau = (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{y}$ with τ given in equation 4.8 is an unbiased estimator of β.*

Because $\hat{\beta}_\tau = (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{y}$ is unbiased, the variance is identical to that given in equation 7.4 on page 49 for general $E(\epsilon\epsilon'|\mathbf{X}, W) = \Sigma$. With Expectile Assumption 4 in place, $E(W\epsilon\epsilon'|\mathbf{X}, W) = \nu^2 I_n$ and the variance of $\hat{\beta}_\tau$ is

$$\begin{aligned} Var(\hat{\beta}_\tau|\mathbf{X}, W) &= E\left((\hat{\beta}_\tau - \beta_\tau)(\hat{\beta}_\tau - \beta_\tau)'|\mathbf{X}, W\right) \\ &= (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'E\left(W\epsilon\epsilon'|\mathbf{X}, W\right)W\mathbf{X}\left(\mathbf{X}'W\mathbf{X}\right)^{-1} \\ &= \nu^2\left(\mathbf{X}'W\mathbf{X}\right)^{-1}. \end{aligned} \tag{4.11}$$

Naturally, this extends to predictors of $A\mathbf{X}\beta$. The OLS predictor is biased, but the expectile predictor $A\mathbf{X}\hat{\beta}_\tau$ is unbiased. We may alter assumption 4 (in the proposition below) to develop a more robust sandwich variance or to use the GLS estimator from the previous section. Either way, we have shown that *the expectile* is the natural generalization of the mean regression to a non-central regression design implied by of equation 4.4. The fourth Gauss-Markov assumption here is a special case of the fourth expectile assumption, as in equation 3.11 on page 16, which leaves the expectile GLS estimator as the BLUE.

Proposition 10. *Let Gauss-Markov assumptions 1,3 hold. Let $E(\epsilon_i|\mathbf{X}) = c$, and let $E(\epsilon\epsilon'|\mathbf{X}, W) = \nu^2\Sigma = \nu^2 W^{-1/2}\Omega W^{-1/2}$ for some diagonal positive definite Ω. Then the expectile regression estimator $\hat{\beta}_\tau = \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y}$ with τ given in equation 4.8 is the best linear unbiased estimator of β.*

The proof in section 3.2.2 on page 16 is applicable because $E_\tau(\epsilon_i|\mathbf{X}) = 0$. Specifically, equation 3.12 shows that this estimator must be the BLUE.

4.2 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 ϵ_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 ϵ_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 ϵ_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 $\mathbf{y}$ and $\mathbf{X}$. Expectiles are especially useful for this purpose.

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

$$\begin{aligned} E(\hat{\beta}_{OLS}|\mathbf{X}) &= \beta \\ Var(\hat{\beta}_{OLS}|\mathbf{X}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \tag{4.12}$$

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

$$\begin{aligned} Var(\mathbf{x}_i'\hat{\beta}_{OLS}|\mathbf{X}) &= \mathbf{x}_i' Var(\hat{\beta}_{OLS}|\mathbf{X})\mathbf{x}_i \\ &= \sigma^2 \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'. \end{aligned} \tag{4.13}$$

This is also the best linear unbiased predictor of y_i. This is proven in the usual way by supposing that there is some other unbiased linear estimator $\tilde{\beta} = Cy$ with $C = D + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $DX =$ 0and noticing that the predictor $\mathbf{x}_i'\tilde{\beta}$ has variance

$$\begin{aligned} Var(\mathbf{x}_i'\tilde{\beta}|\mathbf{X}) &= \mathbf{x}_i' Var(\tilde{\beta}|\mathbf{X})\mathbf{x}_i \\ &= \mathbf{x}_i' C(\sigma^2 I_n)C'\mathbf{x}_i \\ &= \sigma^2 \mathbf{x}_i' \left(D + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\left(D + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)'\mathbf{x}_i \\ &= \sigma^2 \mathbf{x}_i' \left(D'D + (\mathbf{X}'\mathbf{X})^{-1}\right)\mathbf{x}_i \\ &\geq Var(\mathbf{x}_i'\hat{\beta}_{OLS}|\mathbf{X}) \end{aligned} \tag{4.14}$$

which is merely a modified case of the proof in equation 3.10. However, if we know that a particular 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[9] of ϵ_i,

$$\frac{\Pr(\epsilon^* \geq 0)}{\Pr(\epsilon^* < 0)} = \alpha \frac{\Pr(\epsilon_i \geq 0)}{\Pr(\epsilon_i < 0)} \tag{4.15}$$

[9]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.

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 $\mathbf{z}'\hat{\beta}_{OLS}$. This new information about the error term violates two Gauss-Markov assumptions: strict exogeneity and spherical variance (assumptions 2 and 4)[10], at least for the coefficients $\hat{\beta}_{OLS}$, which must therefore be biased. Is there a different set of coefficients we should consider?

To improve our estimator and our predictor, we can incorporate the new information in the following way.

$$\begin{aligned} E(\epsilon^*|\mathbf{z},\mathbf{X}) &= E(\epsilon^*|\mathbf{z},\mathbf{X},\epsilon^* \geq 0)\Pr(\epsilon^* \geq 0) + E(\epsilon^*|\mathbf{z},\mathbf{X},\epsilon_i < 0)\Pr(\epsilon^* < 0) \\ &= E(\epsilon^*|\mathbf{z},\mathbf{X},\epsilon^* \geq 0)\Pr(\epsilon_i \geq 0) \times \alpha\frac{\Pr(\epsilon^* < 0)}{\Pr(\epsilon_i < 0)} \\ &\quad + E(\epsilon^*|\mathbf{z},\mathbf{X},\epsilon_i < 0)\Pr(\epsilon_i < 0) \times \frac{1}{\alpha}\frac{\Pr(\epsilon^* \geq 0)}{\Pr(\epsilon_i \geq 0)} \\ &= E(\epsilon^*|\mathbf{z},\mathbf{X},\epsilon^* \geq 0) \times \tau + E(\epsilon^*|\mathbf{z},\mathbf{X},\epsilon_i < 0) \times 1 - \tau \end{aligned} \tag{4.16}$$

for some[11] $\tau \in (0,1)$. But clearly[12] this is $E_\tau(\epsilon^*|\mathbf{z},\mathbf{X})$ or $E(w^*\epsilon^*|\mathbf{z},\mathbf{X})$ 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 7.1 and leads to the usual expectile regression estimator. We know that $E(w^*\epsilon^*|\mathbf{z},\mathbf{X}) = 0$ must hold for all possible $\mathbf{z}$ and, thus, it holds for all $\mathbf{x}_i$, implying $E(w_i\epsilon_i|\mathbf{X},W)$ for all i in $1,...,n$. The same condition must be true for any other unbiased estimator.

Because any unbiased estimator must elicit the τ^{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 $E_\tau(y|\mathbf{X}) = \mathbf{X}\beta_\tau$, any unbiased linear predictor $\mathbf{z}'\tilde{\beta}$ has $E_\tau(\mathbf{z}'\tilde{\beta}|\mathbf{X}) =$

[10]Strict exogeneity is violated by construction. If the distribution of ϵ_i is F, then the distribution of the residual ϵ^* is the $\tilde{F}$ from equation 2.6 with $\frac{\tau}{1-\tau} = \alpha$ (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:

$$\underbrace{(.5-\tau)E(\epsilon_i I(\epsilon \geq 0)|\mathbf{X})}_{<0} + \underbrace{(.5-1+\tau)E(\epsilon_i I(\epsilon_i < 0)|\mathbf{X})}_{<0} = 0$$

which is impossible so long as ϵ_i is not uniquely zero. The variance of a vector of these atypical observations may still be spherical, but the coefficient σ^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 4.4.

[11]A straightforward calculation shows that $\frac{\tau}{1-\tau} = \alpha$. For a specific, known contingency, we may use a specific τ in equation 4.16. More generally, it is reasonable to perform the estimation using many different τ in order to survey the full spectrum of possible variation.

[12]The weights in equation 4.16 add to one because, as seen previously, they are equal to $\Pr(\epsilon^* \geq 0)$ and $\Pr(\epsilon^* < 0)$, respectively.

$E_\tau(\mathbf{z}'Cy|\mathbf{X}) = \mathbf{z}'C\mathbf{X}\beta_\tau = \mathbf{z}'\beta_\tau$. The variance[13] of the unbiased predictor is

$$\begin{aligned} Var(\mathbf{z}'\tilde{\beta}|\mathbf{X}, W) &= E\left((\mathbf{z}'\tilde{\beta} - E_\tau(\mathbf{z}'\tilde{\beta}|\mathbf{X}, W))^2|\mathbf{X}, W\right) \\ &= \mathbf{z}'CE\left((\mathbf{y} - \mathbf{X}\beta_\tau)(\mathbf{y} - \mathbf{X}\beta_\tau)'|\mathbf{X}, W\right)C'\mathbf{z} \\ &= \nu^2\mathbf{z}'C\Sigma C'\mathbf{z} \end{aligned} \tag{4.17}$$

where $\nu^2\Sigma = \nu^2 W^{-1/2}\Omega W^{-1/2}$ denotes $E(\epsilon\epsilon'|\mathbf{X}, W)$ as before. This is constant with respect to the choice of unbiased estimator. Then we can minimize the variance, subject to unbiasedness of the estimator, by constrained optimization. Writing the Lagrangian below,

$$\mathcal{L}(\mathbf{z}'C) = \frac{1}{2}\mathbf{z}'C\Sigma C'\mathbf{z} + \lambda'(X'C'\mathbf{z} - \mathbf{z}) \tag{4.18}$$

minimize with respect to $\mathbf{z}'C$:

$$\frac{\partial\mathcal{L}(\mathbf{z}'C)}{\mathcal{L}(\mathbf{z}'C)} = \mathbf{z}'C\Sigma + \lambda'\mathbf{X}' = 0 \tag{4.19}$$

And the first-order condition with respect to λ leaves $CX - I = 0$. Together, these imply

$$\begin{aligned} &\mathbf{X}'\Sigma^{-1}\mathbf{X}\lambda = \mathbf{X}'C'\mathbf{z} \\ &\quad\implies \lambda = \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{z} \end{aligned} \tag{4.20}$$

and

$$\begin{aligned} &\Sigma C'\mathbf{z} = \mathbf{X}\lambda = \mathbf{X}\left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{z} \\ &\implies C' = \Sigma^{-1}\mathbf{X}\left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1} \end{aligned} \tag{4.21}$$

Thus, the best linear unbiased predictor $\mathbf{z}'\tilde{\beta}_\tau$ makes use of the GLS-type estimator

$$\hat{\beta}_{\tau,GLS} = C\mathbf{y} = \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y} \tag{4.22}$$

where $\nu^2\Sigma = E\left((\mathbf{y} - \mathbf{X}\beta_\tau)(\mathbf{y} - \mathbf{X}\beta_\tau)'|\mathbf{X}, W\right)$. Under the ideal conditions such as in Expectile Assumption 4, $E\left((\mathbf{y} - \mathbf{X}\beta_\tau)(\mathbf{y} - \mathbf{X}\beta_\tau)'|\mathbf{X}, W\right) = \nu^2 W^{-1}$, we have the usual expectile estimator

$$\hat{\beta}_\tau = C\mathbf{y} = \left(\mathbf{X}'W\mathbf{X}\right)^{-1}\mathbf{X}'W\mathbf{y}. \tag{4.23}$$

But this need not be the case. We have proven the following proposition.

Proposition 11. *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 4.15, the optimal linear predictor is* $\mathbf{z}'\tilde{\beta}_\tau$ *where* $\tilde{\beta}_\tau$ *is the GLS-type expectile estimator,* $\hat{\beta}_{\tau,GLS} = \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y}$*,* $\nu^2\Sigma = E\left((\mathbf{y} - \mathbf{X}\beta_\tau)(\mathbf{y} - \mathbf{X}\beta_\tau)'|\mathbf{X}, W\right)$.

[13] The variance of the predictor in equation 4.17 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.

In this case, we can also show that the best linear unbiased estimator of β_τ is $\hat{\beta}_{\tau,GLS}$.

Proposition 12. *Let Gauss-Markov assumptions 1-3 hold for the data* $\mathbf{y}, \mathbf{X}$*. For any observation with a misspecified residual distribution as in equation 4.15, the best linear unbiased estimator of the coefficient vector* β_τ *is* $\hat{\beta}_{\tau,GLS} = \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y}$*,* $\nu^2\Sigma = Var\left(\mathbf{y} - \mathbf{X}\beta_\tau|\mathbf{X}, W\right)$*.*

The proof is simple. We have already shown that unbiasedness requires the predictor to use an unbiased estimator for β_τ and that the optimal linear predictor is $\mathbf{z}'\hat{\beta}_{\tau,GLS}$ for any $\mathbf{z}$, so it must hold in particular when $\mathbf{z}$ is an elementary basis vector and thus the optimal linear predictor of $[\beta_\tau]_j$ is $[\hat{\beta}_{\tau,GLS}]_j$ for any j in $1, ..., k$. A proof that $\hat{\beta}_{\tau,GLS}$ has the minimum variance of any unbiased estimator for β_τ is given in equation 3.12. Under propositions 10-11, the expectile coefficients β_τ retain the interpretation as partial derivatives of the expected value of the data with respect to $\mathbf{z}_i$. That is, for the atypical observation y^*,

$$\frac{\partial E(y^*)}{\partial \mathbf{z}_j} = \beta_{\tau,j}\ \forall j \in \{1, ...k\}. \tag{4.24}$$

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

4.3 Expectiles for Missing Data

Similar reasoning is applicable when data are missing not-at-random, but asymmetrically. Suppose that the "true" data generating process $y_i^* = \mathbf{x}_i^*\beta + \epsilon_i^*$ has an unknown distribution of errors $\tilde{F}$ as in equation 2.6, but the data is not perfectly observed and observations are missing. If we only know that positive error terms $\epsilon^* \geq 0$ are α times as likely to go missing as negative error terms, we have

$$\frac{\Pr(\epsilon_i \geq 0)}{\Pr(\epsilon_i < 0)} = \frac{1}{\alpha}\frac{\Pr(\epsilon^* \geq 0)}{\Pr(\epsilon^* < 0)} \tag{4.25}$$

where the observed data have $\frac{1}{\alpha}$ times as many positive error terms. Clearly, this is the same as equation 4.15. So we can incorporate this information as in equation 4.16 to produce a useful estimator. If $E(\epsilon^*|\mathbf{X}^*) = 0$ for the true data generating process but equation 4.25 is true for the observed data, then $E(w_i\epsilon_i|\mathbf{X}) = 0$ is true for the sample as in equation 4.16. Then we may solve for the optimal estimator by the method in the previous section.

Proposition 13. *Let Gauss-Markov assumptions 1-3 hold for the true data generating process of* $\mathbf{y}^*, \mathbf{X}^*$*. For data* $\mathbf{y}, \mathbf{X}$ *with missing observations as in 4.25, the best linear unbiased estimator for the true DGP is the GLS expectile estimator,* $\hat{\beta}_{\tau,GLS} = \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y}$*,* $\nu^2\Sigma = E(\epsilon\epsilon'|\mathbf{X}, W)$*.*

Proof. Exactly the same as before. There is some $\tau \in (0, 1)$ such that $E_\tau(\epsilon_i|\mathbf{X}) = 0$. This is the same as the proof in equation 4.16. Then any unbiased linear estimator can be written $\tilde{\beta} = C\mathbf{y}$ with $E_\tau(\tilde{\beta}|\mathbf{X}) = \mathbf{X}\beta_\tau$, where β_τ is the true τ^{th} expectile coefficient vector for the observed data. See equation 3.12 for a proof 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 4.2 on page 21, 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 $\mathbf{z}$ depends on whether we are interested in data before or after observations go missing. If we are interested in the observed

data only, then OLS is appropriate. If we are interested in prediction from the underlying data generating process, is is optimal to use the GLS expectile predictor $\mathbf{z}'\hat{\beta}_{\tau,GLS}$. The corresponding proposition and proof are below.

Proposition 14. *Let Gauss-Markov assumptions 1-3 hold for the true data generating process of* $\mathbf{y}^*, \mathbf{X}^*$*, with observations missing as in 4.25. For an observation with covariates* $\mathbf{z}$*, the best linear unbiased predictor for the true data generating process is* $\mathbf{z}'\hat{\beta}_{\tau,GLS}$ *with* $\hat{\beta}_{\tau,GLS} = \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y}$*,* $\nu^2\Sigma = E(\epsilon\epsilon'|\mathbf{X}, W)$*.*

Proof. As before, we have an unbiased linear predictor given by $\mathbf{z}'C\mathbf{y}$ with $E(\mathbf{z}'C\mathbf{y}|\mathbf{X}, W) = \mathbf{z}'\beta_\tau$. The variance of any such predictor is

$$\begin{aligned} Var(\mathbf{z}'C\mathbf{y}|\mathbf{X}, W) &= \mathbf{z}'Var\left(C\mathbf{y} - C\mathbf{X}\beta_\tau|\mathbf{X}, W\right)\mathbf{z} \\ &= \mathbf{z}'\left(D + \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\right)\nu^2\Sigma\left(D + \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\right)\mathbf{z} \\ &= \nu^2\mathbf{z}'D\Sigma D'\mathbf{z} + \nu^2\mathbf{z}'\left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{z} \\ &\geq \nu^2\mathbf{z}'\left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{z} = Var(\mathbf{z}'\hat{\beta}_{\tau,GLS}|\mathbf{X}, W). \end{aligned} \tag{4.26}$$

□

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 can be shown using the similar principles.

4.4 Example: Mortgage Applications

To illustrate the value of the ordinary expectile regression and its GLS counterpart, we will give a brief example using a binary dependent variable. This will unite the estimators proposed in sections 3.2 on page 14 and 3.2.2 on page 16 with the feasible weights as in subsection 5.2 on page 32. Both the interpretation in section 4.2 on page 21 and the interpretation in section 4.3 on the preceding page are applicable.

The data for this demonstration are the Boston Home Mortgage Disclosure Act (HMDA) data from the famous paper by Munnell et al. [36, 41]. The research question is whether there is a statistically significant difference in mortgage application denial based on race. The dependent variable of interest is *deny*, a binary variable that takes a value of one if the mortgage application was denied and zero otherwise. Because this variable is binary, we will employ a linear probability model to study how $E(y_i|\mathbf{x}_i) = \Pr(y_i = 1|\mathbf{x}_i)$ varies with several covariates. The most interesting covariate is *black,* which takes a value of one if the individual is black and zero otherwise.

The linear probability model is appropriate for this demonstrative example. In the simple regression of *deny* on *black*, the predictor falls within the unit interval with probability one: thus the expectile weights are known *a priori* and the usual criticism of linear probability models does not apply. Additionally, the binary response model gives the expectile weights a concrete interpretation: predictors with non-central τ correspond to atypical relative odds of denial. Both OLS and GLS estimators have been used for binary response models of this type, see [5]. This example also serves as a clear motivator for future research into binary response models for expectile regression.

Simple Regression: *deny* on *black*

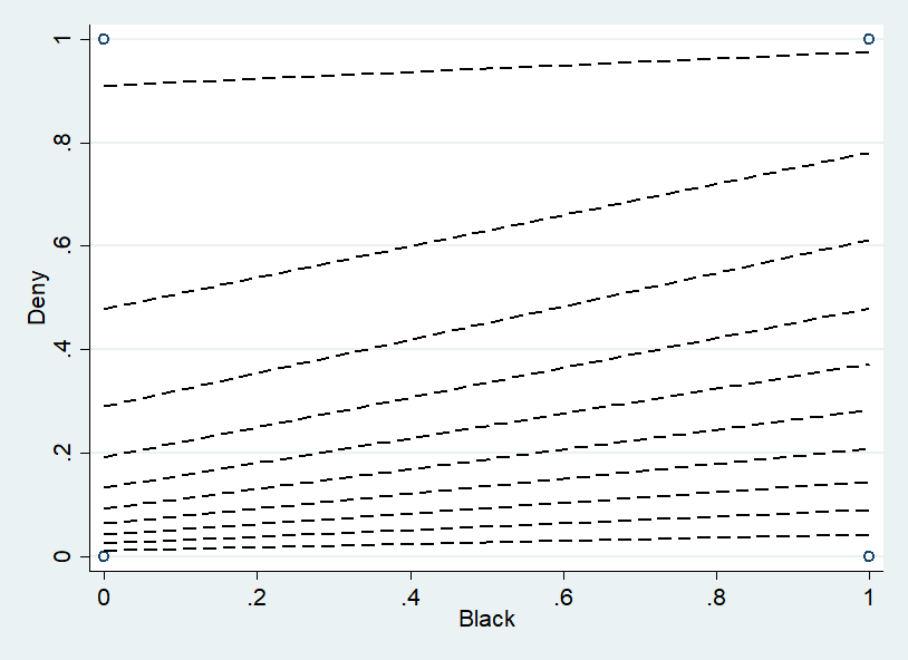

Figure 3: The simple regression of *deny* on *black*. All observations fall into one of the four corners, constraining a traditional quantile regression. Expectiles for τ equal to all multiples of $.1 \in (0, 1)$ are given along with $\tau = .99$. The expectile "decile" with the largest slope is $\tau = .8$, suggesting that the largest expected effect of *black* on *deny* occurs for individuals with approximately 4 times the average odds of denial, *ceteris paribus*. A subset of regression coefficients is shown in Table 3.

Table 1: Simple Expectile Regression Coefficients

	(1)	(2)	(3)	(4)	(5)
VARIABLES	deny	deny	deny	deny	deny
τ	.5	.6	.7	.8	.9
black	0.191***	0.239***	0.287***	0.323***	0.302***
	(0.0253)	(0.0295)	(0.0324)	(0.0327)	(0.0281)
Constant	0.0926***	0.133***	0.192***	0.290***	0.479***
	(0.00642)	(0.00880)	(0.0119)	(0.0157)	(0.0191)
Observations	2,380	2,380	2,380	2,380	2,380
R-squared	0.042	0.053	0.063	0.071	0.067

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

Table 3: 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.

4.4.1 Demonstration Results

For this data, the core question is how different covariates affect the probability of home mortgage application denial. In Figure 3, we have a scatterplot of *deny* relative to *black*. Obviously, both are binary variables and there are only four possible locations where data are found. This is problematic for a traditional quantile regression, which passes through data points. Using that methodology, there would be exactly three possible fit lines tracing a "Z" shape through these four points. Expectiles, however, fall between the data and vary with respect to τ as discussed elsewhere. In the simple regression model shown in that figure, fitted values fall in the unit interval with probability one and the weights w_i are known *a priori*.

From the figure, it is clear that the expectiles of *deny* vary differently with respect to *black*. In fact, it is clear that the lowest expectiles have very low slope and higher expectiles have the opposite. Table 3 lists a subset of regression coefficient estimates from the same figure. While, on average, individuals with *black* equal to one are approximately 19% more likely to be denied, the coefficient is as high as 32% at the $\tau = .8$ expectile. This indicates that the individuals most affected by racial disparity are those who were more likely to be denied regardless of race. The result should not be surprising.

For linear probability models, Goldberger [19] 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

$$Var(\epsilon_i|\mathbf{X}) = \mathbf{x}_i'\beta_\tau(1 - \mathbf{x}_i'\beta_\tau). \tag{4.27}$$

Setting weights equal to the inverse of this expression produces consistent estimates [35] and has been adopted widely for linear probability models [20, p. 727]. We adopt this for the expectile linear probability model and show results in Table 5. These are similar to those in Table 4, with the maximum coefficient for *black* reaching approximately 14%. In the GLS case, however, the

Table 4: Expectile Regression Coefficients

VARIABLES	(1) deny	(2) deny	(3) deny	(4) deny	(5) deny
$\tau =$	.1	.3	.5	.7	.9
black	0.0170***	0.0485***	0.0837***	0.124***	0.150***
	(0.00540)	(0.0138)	(0.0226)	(0.0316)	(0.0379)
pi_rat	0.140***	0.299***	0.449***	0.655***	0.895***
	(0.0445)	(0.0792)	(0.114)	(0.157)	(0.194)
hse_inc	0.0375	0.0251	-0.0480	-0.166	-0.288
	(0.0371)	(0.0688)	(0.110)	(0.167)	(0.237)
ltv_med	0.00798**	0.0182**	0.0314**	0.0513**	0.0735**
	(0.00338)	(0.00744)	(0.0127)	(0.0209)	(0.0334)
ltv_high	0.0391**	0.115***	0.189***	0.250***	0.268***
	(0.0167)	(0.0367)	(0.0502)	(0.0572)	(0.0622)
ccred	0.00608***	0.0174***	0.0308***	0.0478***	0.0622***
	(0.00119)	(0.00279)	(0.00458)	(0.00668)	(0.00853)
mcred	0.00732**	0.0137*	0.0209*	0.0336*	0.0545*
	(0.00355)	(0.00724)	(0.0113)	(0.0176)	(0.0297)
pubrec	0.0497***	0.131***	0.197***	0.237***	0.195***
	(0.0103)	(0.0253)	(0.0349)	(0.0399)	(0.0390)
denpmi	0.520***	0.708***	0.702***	0.669***	0.560***
	(0.126)	(0.0688)	(0.0451)	(0.0380)	(0.0359)
selfemp	0.0143***	0.0359***	0.0598***	0.0946***	0.145***
	(0.00516)	(0.0123)	(0.0205)	(0.0321)	(0.0467)
Constant	-0.0812***	-0.146***	-0.183***	-0.215***	-0.124*
	(0.0196)	(0.0237)	(0.0277)	(0.0387)	(0.0717)
Observations	2,380	2,380	2,380	2,380	2,380
R-squared	0.107	0.210	0.266	0.301	0.243

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

Table 4: 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 τ: it is the individuals with a high probability of denial who also experience the largest marginal effects.

Table 5: Expectile GLS Regression Coefficients

VARIABLES	(1) deny	(2) deny	(3) deny	(4) deny	(5) deny
$\tau =$	.1	.3	.5	.7	.9
black	0.0116***	0.0411***	0.0810***	0.137***	0.108***
	(0.00336)	(0.0115)	(0.0216)	(0.0328)	(0.0249)
pi_rat	0.0468***	0.174***	0.377***	0.617***	0.871***
	(0.00823)	(0.0287)	(0.0614)	(0.115)	(0.142)
hse_inc	0.00821	0.0332	0.0633	0.110	-0.262
	(0.00963)	(0.0202)	(0.0768)	(0.122)	(0.162)
ltv_med	0.00220***	0.00835**	0.0191***	0.0442**	0.0761**
	(0.000820)	(0.00353)	(0.00538)	(0.0175)	(0.0312)
ltv_high	0.0318***	0.0945***	0.152***	0.256***	0.130***
	(0.0109)	(0.0340)	(0.0526)	(0.0457)	(0.0358)
ccred	0.00367***	0.0133***	0.0274***	0.0508***	0.0632***
	(0.000554)	(0.00195)	(0.00387)	(0.00647)	(0.00707)
mcred	0.00239***	0.00899***	0.0191***	0.0293***	0.0701***
	(0.000762)	(0.00202)	(0.00644)	(0.0106)	(0.0249)
pubrec	0.0409***	0.129***	0.211***	0.239***	0.126***
	(0.00850)	(0.0239)	(0.0326)	(0.0384)	(0.0275)
denpmi	0.525***	0.735***	0.716***	0.838***	0.593***
	(0.128)	(0.0367)	(0.0288)	(0.0200)	(0.0288)
selfemp	0.00793***	0.0297***	0.0640***	0.115***	0.165***
	(0.00217)	(0.00746)	(0.0166)	(0.0283)	(0.0379)
Constant	-0.0223***	-0.0833***	-0.175***	-0.282***	-0.136**
	(0.00240)	(0.00971)	(0.0180)	(0.0340)	(0.0638)
Observations	2,133	2,130	2,132	2,153	2,331
R-squared	0.017	0.214	0.472	0.584	0.547

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

Table 5: Regression coefficients for the full spectrum: $\tau = .1, .3, .5, .7, .9$, with many covariates added. The GLS estimator is used with $\omega_i = (\mathbf{x}_i'\beta(1-\mathbf{x}_i'\beta))^{-1}$. Results largely resemble the previous table, with the coefficient on *black* reaching a value of approximately 14%.

maximum coefficient is obtained at $\tau = .7$, while the OLS coefficient is maximized at $\tau = .9$. There is room for discussion regarding which qualities might cause an individual to expect the largest difference as a result of the *black* variable. The number of observations in the GLS results varies because observations with predictors outside the unit interval have undefined variance per 4.27. As a result, these observations dropped as the model iterates to convergence.

In the figure and all three tables, the economic result is the same. The 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 τ or extremely high τ) or larger (for τ between .7 and .9, say) than usually reported. This adds value to the empirical discussion by revealing heterogeneity.

In this example, the non-central regressions deliver some insight into which individuals might be disproportionately affected by their race or unmeasured race-correlated attributes. Conditional on the full spectrum of covariates available, the individuals closest to the "margin" may expect to be 15% more likely to be denied on account of race. This is nearly twice the coefficient suggested by a standard OLS or GLS regression.

5 Feasibility

5.1 Feasibility of the Expectile WLS Estimator

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 on page 26. The estimator for expectile coefficients

$$\hat{\beta}_\tau = (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{y} \tag{5.1}$$

has a typical generalized least squares form with diagonal weight matrix given by

$$[W]_{ii} = \begin{cases} \tau & \text{if } y_i - \mathbf{x}_i'\beta_\tau \geq 0 \\ 1-\tau & \text{if } y_i - \mathbf{x}_i'\beta_\tau < 0. \end{cases} \tag{5.2}$$

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

$$\begin{aligned} \hat{\beta}_{\tau,FGLS} &= (\mathbf{X}'\hat{W}\mathbf{X})^{-1}\mathbf{X}'\hat{W}\mathbf{y} \\ [\hat{W}]_{ii} = \hat{w}_i &= \begin{cases} \tau & \text{if } y_i - \mathbf{x}_i'\hat{\beta}_{\tau,FGLS} \geq 0 \\ 1-\tau & \text{if } y_i - \mathbf{x}_i'\hat{\beta}_{\tau,FGLS} < 0. \end{cases} \end{aligned} \tag{5.3}$$

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 $\hat{\beta}_{\tau,FGLS}$ can be evaluated. Repeating this procedure with the new value $\hat{\beta}_{\tau,FGLS}$, the estimated coefficients will converge relatively quickly as the sub-Hessian for this problem is globally negative semidefinite (see [38]). Then the converged estimator is the correct (exact) *sample* linear expectile coefficient vector. This is similar overall to the procedure used by Zellner [49] for the mean regression.

It is clear from the definition in equation 5.3 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 $\hat{\beta}_{\tau,FGLS}$ is consistent, which was proven by [37]. In the heteroscedastic case, we have $\Sigma_{ii} = w_i\omega_{ii}$. Then consistency of $\hat{\beta}_{\tau,FGLS}$ requires a consistent estimator of w_i and one of ω_{ii}. See [8] or [38] 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 $\mathbb{F}_n$ 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,

$$\{i : \ y_i - \mathbf{x}_i'\hat{\beta}_{\tau,FGLS} \geq 0\} \equiv \{i : \ y_i - \mathbf{x}_i'\beta_\tau \geq 0\} \tag{5.4}$$

then the estimated weights are exact; $\hat{w}_i = w_i$. In that case, the optimal expectile estimator $\hat{\beta}_\tau = (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{y}$ *is feasible*. Using the location model as an example, we can see that this happens with positive probability.

Lemma 15. *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 τ^{th} sample expectile $\hat{\mu_\tau}$ by minimizing $R_n(\theta;\tau) = n^{-1}\sum_{i=1}^n \varsigma_\tau(y_i - \theta)$, where the "swoosh" function ς_τ is as in equation 5.5 or equation 2.4 on page 4,

$$\varsigma_\tau(u) = u^2|\tau - I(u < 0)|. \tag{5.5}$$

Because the solution of this strictly convex minimization problem (and the first-order condition) is unique, we have

$$\hat{\mu_\tau} = \arg\min_\theta n^{-1}\sum_{i=1}^n \varsigma_\tau(y_i - \theta).$$

$$\implies n^{-1}\sum_{i=1}^n (y_i - \hat{\mu_\tau})|\tau - I(y_i - \hat{\mu_\tau} < 0)| = 0 \tag{5.6}$$

or

$$n^{-1}(1-\tau)\sum_{i=1}^n (y_i - \hat{\mu_\tau})I(y_i - \hat{\mu_\tau} < 0) = n^{-1}\tau\sum_{i=1}^n (y_i - \hat{\mu_\tau})I(y_i - \hat{\mu_\tau} \geq 0) \tag{5.7}$$

The probability of the sample expectile $\hat{\mu_\tau}$ being less than or equal to a particular value x, is then the same as

$$\Pr(\left(\tau\int_x^\infty (y-x)d\mathbb{F}_n\right) \leq \left((1-\tau)\int_{-\infty}^x (x-y)d\mathbb{F}_n\right)), \tag{5.8}$$

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 $\hat{\mu_\tau} \leq x$ is *nondecreasing* (the set of possible values for $D \coloneqq \{\{y_i\}_{i=1}^n : \hat{\mu_\tau} \leq x\}$ is increasing). In particular, in means that the probability of the event $\Pr(\hat{\mu_\tau} \leq x) = \int_D f(\mathbf{y})d\mathbf{y}$ is monotone increasing if F has Lebesgue measure, implying that the distribution of $\hat{\mu_\tau}$ has positive density on $[\inf(Y), \sup(Y)]$. □

Proposition 16. *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 $\tau \in (0,1)$, μ_τ falls into the interior of this interval with probability one. Then, for any sample, the probability that there is at least one observation above μ_τ and at least one observation below μ_τ 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 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 the weighted average $\hat{\mu_\tau}$ is asymptotically dense, there is a nonempty set of possible samples producing $\max(y_i|y_i < \mu_\tau) < \hat{\mu_\tau} < \min(y_i|y_i \geq \mu_\tau)$ as $n \to \infty$.

Suppose that the distribution F does not have Lebesgue measure, but let F be continuous at μ_τ. Let there be some sequence $a_n \to \infty$ such that $a_n(\hat{\mu_{\tau,n}} - \mu_\tau) \rightsquigarrow Z$ as $n \to \infty$. Very broad conditions for this limiting behavior in expectiles were published recently [23]. So long as Z has density, clearly,

$$\int_{a_n(\max(y_i|y_i<\mu_\tau)-\mu_\tau)}^{a_n(\min(y_i|y_i\geq\mu_\tau)-\mu_\tau)} dZ > 0 \tag{5.9}$$

for any $n < \infty$ and asymptotically so long as $a_n\left(\min(y_i|y_i \geq \mu_\tau) - \max(y_i|y_i < \mu_\tau)\right) \not\to 0$. In one interesting example, where the distribution F is fully discrete or has no density in an open ball around μ_τ, 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

$$a_n\left(\min(y_i|y_i \geq \mu_\tau) - \mu_\tau\right) \to \infty \tag{5.10}$$

$$a_n\left(\mu_\tau - \max(y_i|y_i < \mu_\tau)\right) \to \infty. \tag{5.11}$$

which implies

$$\lim_{n\to\infty} \int_{a_n(\max(y_i|y_i<\mu_\tau)-\mu_\tau)}^{a_n(\min(y_i|y_i\geq\mu_\tau)-\mu_\tau)} dZ = \int_{-\infty}^{\infty} dZ = 1 \tag{5.12}$$

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.

5.2 Example: Expectile Binary Response

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 [22], [21] 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

$$E(Y|X = x) = \Pr(Y = 1|x) = G(x'\beta) \tag{5.13}$$

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 [2] for a thorough comparison or see Greene [20] 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, $\mathbf{x}_i'\beta$, to the *interior* of the unit interval[14]

$$G : \mathbb{R}^k \mapsto (0, 1) \tag{5.14}$$

and we note the obvious result.

Proposition 17. *Let $\mathbf{y}, \mathbf{X}$ 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 5.14. 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, β, 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 β 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 [5] 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.

6 Variance of Expectile Residuals

The estimated variance and mean squared error of expectile residuals have not been studied thoroughly. This has the potential to become a complicated subject. As an example, assume that $\epsilon|\mathbf{X} \sim (0, \sigma^2\Sigma)$ so that β is the "true" mean regression coefficient vector. Then

$$\begin{aligned} Var(y|\mathbf{X}) &= E\left((y - \mathbf{X}\beta)(y - \mathbf{X}\beta)'|\mathbf{X}\right) \\ &= E\left(\epsilon\epsilon'|\mathbf{X}\right) = \sigma^2\Sigma \end{aligned} \tag{6.1}$$

We have a shape parameter Σ and a scale parameter σ^2. The shape parameter Σ requires further assumptions to identify: see [33]. Tools are available for the mean regression, as in [46]. These tools can be adapted to expectile regression, but this subject is large and situationally dependent.

[14] 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 $\Pr(y_i = 1|\mathbf{x}_i) = 1$ or 0 and to observe an error term, as this would make the model logically incoherent.

We cannot address it properly here. As in previous sections, we will restrict our attention to the independent case where $E(\epsilon\epsilon'|\mathbf{X})$ is a diagonal matrix.

Note also that the expectile variance parameter σ_τ^2 is not the mean variance σ^2 nor is it the σ_i^2 of the location-scale model which is popular in the generalized quantile literature [6, 12], which varies with i for a given τ. 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, $Var(\epsilon_i|\mathbf{X})$.
2. The weighted variance or mean squared error (under the weighted distribution $\tilde{F}$ in 2.6 on page 5).
3. The mean squared error of residuals, $E(\epsilon_i^2|\mathbf{X})$.

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 3.8 on page 13.

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

$$\begin{aligned} E(\epsilon_i|\mathbf{X}) &= E(y_i - \mathbf{x}_i'\beta_\tau + \mathbf{x}_i'\beta_{.5} - \mathbf{x}_i'\beta_{.5}|\mathbf{X}) \\ &= E(y_i - \mathbf{x}_i'\beta_{.5}|\mathbf{X}) - \mathbf{x}_i'\beta_\tau + \mathbf{x}_i'\beta_{.5} \\ &= \mathbf{x}_i'\beta_{.5} - \mathbf{x}_i'\beta_\tau. \end{aligned} \tag{6.2}$$

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

$$\begin{aligned} Var(\epsilon_i|\mathbf{X}) &= Var(y_i - \mathbf{x}_i'\beta_\tau - E(\epsilon_i|\mathbf{X})|\mathbf{X}) \\ &= Var(y_i - \mathbf{x}_i'\beta_{.5}|\mathbf{X}) \end{aligned} \tag{6.3}$$

which is merely the OLS or GLS residuals. This requires no special treatment[15].

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

$$WVar(\epsilon_i|\mathbf{X}) = E_\tau(\epsilon_i^2|\mathbf{X}), \tag{6.4}$$

see equation 3.8 on page 13. Thus, when we take the interpretations from section 4 on page 17 seriously, we may employ estimators such as the standard weighted variance estimator

$$\frac{n}{n-k}\left(\sum_{i=1}^n w_i\right)^{-1}\sum_{i=1}^n w_i\hat{\epsilon_i^2}. \tag{6.5}$$

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

$$s^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-k} \tag{6.6}$$

[15]The fact that $Var(y_i - \mathbf{x}_i'\beta_\tau|\mathbf{X}) = Var(y_i - \mathbf{x}_i'\beta_{.5}|\mathbf{X})$ for all τ 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.

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

As an aside, the scale parameter σ_τ^2 for the mean squared error of $(\mathbf{y} - \mathbf{X}\beta_\tau)$ will vary across τ. This means that the residuals ϵ_i can become extremely large as the predictor $\mathbf{X}\beta_\tau$ moves far into the tail of an unbounded distribution. Even if $y_i|\mathbf{X}$ are conditionally i.i.d., we have $E(\epsilon\epsilon'|\mathbf{X}) = \sigma_\tau^2 I_n$ where the constant σ_τ^2 will vary depending on τ. From the classic literature, we know that σ_τ^2 is minimized when $\tau = .5$. The following subsections are devoted to estimating σ_τ^2.

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 σ^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 $\hat{\epsilon} = y - \mathbf{X}\hat{\beta}_\tau$. Then they are

$$\hat{\sigma_\tau^2} = \frac{\hat{\epsilon}'\hat{\epsilon}}{n}, \quad s_\tau^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-k}. \tag{6.7}$$

For $\tau = .5$, the latter was shown to be unbiased by Gauss [18]. That classical result was reformulated by Aitken using matrix algebra [1] 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

$$\begin{aligned} E(\epsilon\epsilon'|\mathbf{X}) &= E((y - \mathbf{X}\beta_\tau)(y - \mathbf{X}\beta_\tau)'|\mathbf{X})) \\ &= \text{diag}(\sigma_\tau^2). \end{aligned} \tag{6.8}$$

However, the properties of estimators in equation 6.7 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 σ_τ^2. This is a standard way to prove unbiasedness of s^2 for the mean regression. Notice: with the annihilator matrix M_τ we have

$$\hat{\epsilon} = \mathbf{y} - \mathbf{X}\hat{\beta_\tau} = M_\tau \mathbf{y} = M_\tau(\mathbf{X}\beta_\tau + \epsilon) = M_\tau \epsilon. \tag{6.9}$$

So we may write the sum of squared residuals as

$$\begin{aligned} E(\hat{\epsilon}'\hat{\epsilon}|\mathbf{X}, W) &= E(\epsilon' M_\tau' M_\tau \epsilon|\mathbf{X}, W) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} [M_\tau' M_\tau]_{ij} \, E(\epsilon_i \epsilon_j|\mathbf{X}, W) \\ &= \sum_{i=1}^{n} [M_\tau' M_\tau]_{ii} \, \sigma_\tau^2. \end{aligned} \tag{6.10}$$

This follows from the independence of $\epsilon_i' s$ and the definition of matrix multiplication. Nearly the same result was given by Aitken [1] 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_τ is not symmetric and $M_\tau' M_\tau$ does not simplify. Of course, the last expression above is $\sigma_\tau^2 \times \text{trace}(M_\tau' M_\tau)$ where

Limit Trace of $P_\tau' P_\tau$, $k = 1$

Figure 4: The trace of the inner product of the expectile projection matrix P_τ with itself: $\text{trace}(P_\tau' P_\tau)$. The trace is shown as a function of τ and of $F(\mu_\tau)$ i.e. what proportion of observations are given weight w_i equal to $1 - \tau$. In this case, $\mathbf{X} = 1_n$, so $k = 1$. The minimum, $\text{trace}(P_\tau' P_\tau) = k = 1$, occurs wherever $\tau = .5$ and not otherwise.

$$\begin{aligned} \text{trace}(M_\tau' M_\tau) &= \text{trace}((I - P_\tau)'(I - P_\tau)) \\ &= \text{trace}(I - P_\tau - P_\tau' + P_\tau' P_\tau). \end{aligned} \tag{6.11}$$

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

$$\begin{aligned} \text{trace}(P_\tau) &= \text{trace}(\mathbf{X}(\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W) \\ &= \text{trace}((\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{X}) \\ &= k \end{aligned} \tag{6.12}$$

and the same for $\text{trace}(P_\tau')$. Then

$$\begin{aligned} \text{trace}(P_\tau' P_\tau) &= \text{trace}(W\mathbf{X}(\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W) \\ &= \text{trace}((\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W W\mathbf{X}) \\ &= \text{trace}\left(\left(\sum_{i=1}^n w_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'\right)\left(\sum_{i=1}^n w_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^n w_i^2 \mathbf{x}_i \mathbf{x}_i'\right)\right). \end{aligned} \tag{6.13}$$

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 $\mathbf{X}$ and W, we have the following lemma.

Lemma 18. *Let* $\mathbf{y}, \mathbf{X}$ *be as above and* $\hat{\epsilon} = \mathbf{y} - \mathbf{X}\hat{\beta}_\tau$ *with* $\hat{\beta}_\tau = (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{y}$. *Then*

$$E(\hat{\epsilon}'\hat{\epsilon}|\mathbf{X}, W) = \sigma_\tau^2 \left(n - 2k + \textit{trace}(P_\tau' P_\tau)\right). \tag{6.14}$$

So, *neither* variance estimator in equation 6.7 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 τ, 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 4) 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 4 and 5. In the former, $k = 1$. In the latter, $k = 5$. We will discuss the properties of this function in 6.2 on page 38. We will also show that the sampling distribution for small n can differ significantly from is limiting distribution; see Figure 6 on page 40.

Notice that equation 6.14 implies the following result.

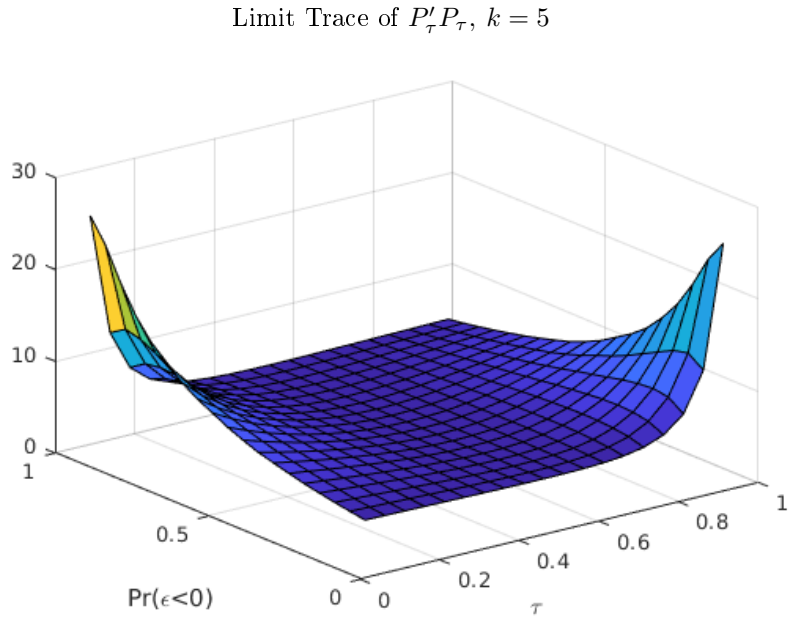

Figure 5: The trace of the inner product of the expectile projection matrix P_τ with itself: trace($P'_\tau P_\tau$). The trace is shown as a function of τ and of $F(\mu_\tau)$ i.e. what proportion of observations are given weight w_i equal to $1 - \tau$. In this case, $k = 5$. The minimum, trace($P'_\tau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

Lemma 19. *Let* $\mathbf{y}, \mathbf{X}$ *be as above and* $\hat{\epsilon} = \mathbf{y} - \mathbf{X}\hat{\beta}_\tau$ *with* $\hat{\beta}_\tau = (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{y}$. *Then the estimator*

$$\tilde{\sigma^2_{\tau,n}} = \frac{\left(\mathbf{y} - \mathbf{X}\hat{\beta}_\tau\right)' \left(\mathbf{y} - \mathbf{X}\hat{\beta}_\tau\right)}{n - 2k + trace(P'_\tau P_\tau)} \tag{6.15}$$

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

The proof is obvious, given the preceding lemma where $E(\hat{\epsilon}'\hat{\epsilon}|\mathbf{X}, W)$ is shown. As a practical matter it should make little difference: the usual estimators in 6.7 will be extremely close to each other and to the revised estimator in 6.15 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 7.4.

If evaluating 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.

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 [37] provide broad conditions for consistency and asymptotic normality of the estimator in the linear regression case. Working papers by Barry et al. [8] and Philipps [38] present asymptotic results for the weighted regression coefficients. Holzmann and Klar [23] investigate the asymptotic properties of the expectile more thoroughly for the location model.

Here, we show that the proposed estimator in equation 6.15 is consistent. We also show that the estimators in equation 6.7 are consistent. Under the assumption that the sequence $\{y_i, \mathbf{x}_i\}$ is independently drawn from a location-scale model, we have $\Pr(w_i = \tau)$ constant and $W, \mathbf{X}$ are independent. Then as $n \to \infty$ we have the following.

Lemma 20. *Let* $W, \mathbf{X}$ *be independent and* $n^{-1}\mathbf{X}'\mathbf{X} \xrightarrow{p} Q_X$ *for some matrix* Q_X *as* $n \to \infty$. *Then*

$$n^{-1}\mathbf{X}'W\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n} w_i\mathbf{x}_i\mathbf{x}'_i \xrightarrow{p} E(w_i\mathbf{x}_i\mathbf{x}'_i) = E(w_i)E(\mathbf{x}_i\mathbf{x}'_i). \tag{6.16}$$

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

Lemma 21. *Let* $n^{-1}\mathbf{X}'W\mathbf{X} \xrightarrow{p} E(w_i)E(\mathbf{x}_i\mathbf{x}'_i)$ *as* $n \to \infty$ *and* $E(\mathbf{x}_i\mathbf{x}'_i)$ *have rank* k . *Then*

$$\begin{aligned}
\text{trace}(P_\tau' P_\tau) &= \text{trace}\left(\left(\sum_{i=1}^n w_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'\right)\left(\sum_{i=1}^n w_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^n w_i^2 \mathbf{x}_i \mathbf{x}_i'\right)\right). \\
&\xrightarrow{p} \text{trace}\left(\left(nE(w_i)Q_X\right)^{-1}\left(nQ_X\right)\left(nE(w_i)Q_X\right)^{-1}\left(nE(w_i^2)Q_X\right)\right) \\
&= \frac{n^2 E(w_i^2)}{n^2 E\left(w_i\right)^2}\text{trace}\left(Q_X^{-1}Q_X Q_X^{-1}Q_X\right) \\
&= \frac{E(w_i^2)}{E\left(w_i\right)^2} \times k \qquad (6.17)
\end{aligned}$$

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 $\mathcal{O}_p(1)$ as both its numerator and denominator are $\mathcal{O}_p(n^2)$. The value of this ratio (both equation 6.17 and the obvious estimator) varies depending on only two factors: τ itself, and the proportion of observations such that $w_i = \tau$, which is simply equal to $\Pr(y_i \geq \mathbf{x}_i'\beta_\tau)$. It is interesting that the expectile variance estimate is influenced by which *quantile* the τ^{th} expectile happens to approximate.

Proposition 22. *Let $\sigma_\tau^2 = E(\epsilon_i^2)$, rank$(E(\mathbf{X}'W\mathbf{X})) = k$, $E(\mathbf{X}'W\epsilon) = 0$, and let $\hat{\beta}_\tau$ be a consistent estimator of β_τ, $\hat{\beta}_\tau \xrightarrow{p} \beta_\tau$. The estimator*

$$\sigma_{\tau,n}^{\tilde{2}} = \frac{\left(\mathbf{y} - \mathbf{X}\hat{\beta}_\tau\right)'\left(\mathbf{y} - \mathbf{X}\hat{\beta}_\tau\right)}{n - 2k + trace(P_\tau' P_\tau)} \xrightarrow{p} \sigma_\tau^2 \qquad (6.18)$$

i.e., $\sigma_{\tau,n}^{\tilde{2}}$ is a consistent estimator.

Proof. The proof is almost standard. Of course $\hat{\epsilon}_i = \epsilon_i - \mathbf{x}_i'(\hat{\beta}_\tau - \beta)$, so

$$\begin{aligned}
\hat{\sigma}_\tau^2 &= \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 - 2\left(\frac{1}{n}\sum_{i=1}^n \epsilon_i w_i \mathbf{x}_i'\right)(\hat{\beta}_\tau - \beta) + (\hat{\beta}_\tau - \beta)'\left(\frac{1}{n}\sum_{i=1}^n w_i \mathbf{x}_i \mathbf{x}_i'\right)(\hat{\beta}_\tau - \beta) \\
&\xrightarrow{p} \sigma_\tau^2
\end{aligned}$$

by the weak law of large numbers, the continuous mapping theorem, and the convergence in probability of $\hat{\beta}_\tau$ to β_τ. Then

$$\sigma_{\tau,n}^{\tilde{2}} = \frac{n}{n - 2k + \text{trace}(P_\tau' P_\tau)}\hat{\sigma^2}_\tau \xrightarrow{p} \sigma_\tau^2$$

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

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 6, we provide some simulation evidence that the sampling properties of $\text{trace}(P_\tau' P_\tau)$ are not a major nuisance. For

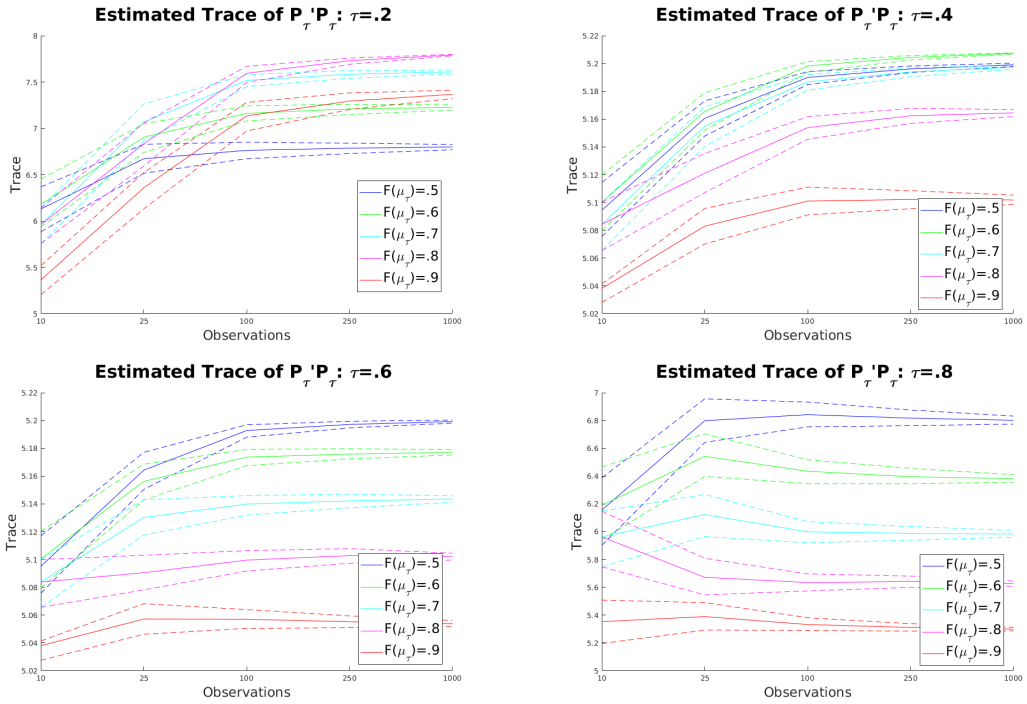

Figure 6: The median and 70% confidence intervals for trace$(P_\tau' P_\tau)$ for $k = 5$ and $n = 10$, 25, 250, or 1000. $[\mathbf{X}]_{ij}$ is distributed uniformly and the distribution of y is not specified, but $\Pr(y_i \geq \mathbf{x}_i' \hat{\beta}_\tau) = F(\mu_\tau)$ is taken as fixed.

five small-to-medium sample sizes ($n = 10$, 25, 250, or 1000) with a fixed $\Pr(y_i \geq \mathbf{x}_i'\hat{\beta}_\tau)$, the 70% confidence intervals of the trace$(P_\tau' P_\tau)$ are shown. The simulated data $\mathbf{X}$ 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 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.

6.3 Asymmetric Conditional MSE

It may be useful to estimate the *conditional* mean squared error $E(\epsilon_i^2 | \epsilon_i \gtrless 0)$ 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[16]. 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

$$\hat{\sigma_\tau^2} | \epsilon_i \geq 0 = \frac{1}{n_2} \sum_{i=1}^{n} \hat{\epsilon_i}^2 I(\epsilon_i \geq 0) \tag{6.19}$$

$$\hat{\sigma_\tau^2} | \epsilon_i < 0 = \frac{1}{n_1} \sum_{i=1}^{n} \hat{\epsilon_i}^2 I(\epsilon_i < 0) \tag{6.20}$$

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 (τ 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 ϵ_i^2, we would have

$$\begin{aligned} E(\sum_{i=1}^{n} \hat{\epsilon_i}^2 I(\hat{\epsilon}_i \geq 0) | \mathbf{X}, W) &= nE\left(\hat{\epsilon_i}^2 I(\hat{\epsilon}_i \geq 0) | \mathbf{X}, W\right) \\ &= E\left(\hat{\epsilon}_i^2 | \hat{\epsilon}_i \geq 0\right) \times \Pr(\hat{\epsilon}_i \geq 0) \end{aligned} \tag{6.21}$$

The expectation in the last line can be estimated by its sample moment and the probability can be estimated using the empirical CDF, so

$$\tilde{\sigma_\tau^2} | \epsilon_i \geq 0 \coloneqq \frac{1}{n - 2k + \text{trace}(P_\tau' P_\tau)} \left(\sum_{i=1}^{n} \hat{\epsilon_i}^2 I(\epsilon_i \geq 0) \right) \times \frac{n}{n_2} \tag{6.22}$$

[16] 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 σ^2. However, we prefer to consider the possibility that this assumption may be violated.

is a reasonable estimator for equation 6.19 with a degrees of freedom penalty. Clearly this estimator[17] is consistent in the same way as equation 6.18 and a similar estimator can be made for the mean squared error of negative error terms.

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. [6]. The R^2 statistic for generalized m-quantiles was introduced very recently [12]. 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 [15] for a discussion of the bias of R^2 or see chapter 8 of the famous text by Maddala [32] for a theoretical overview. The un-adjusted R^2 statistic for non-central estimators is similar to the usual[18] weighted least squares R^2, such as the version given by Kvalseth [30] or the suggested variation (pseudo-R^2) by Willett [47]. The generalized version proposed by Anderson and Sprecher [4] can also be used[19]. As with

[17] 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.

[18] The pseudo R-squared statistic for weighted regression may be used for expectile regression:

$$R^2_{ER} = 1 - \frac{WSSE}{WSST} \tag{6.23}$$

but suffers from the usual criticisms as far as model selection is concerned. For additional discussion regarding model selection in expectile regression models, see [39] or [51].

Notice that the formulation given by Willett

$$= 1 - \left[\frac{(\mathbf{y} - \mathbf{X}\hat{\beta_\tau})' W (\mathbf{y} - \mathbf{X}\hat{\beta_\tau})}{\mathbf{y}' W \mathbf{y} - n\bar{y}_\tau^2}\right]$$

extends to expectiles only if the weights we prefer for the denominator are the weights that produce $\bar{y}_\tau^2$; we may prefer to keep the expression in the form

$$= 1 - \left[\frac{(\mathbf{y} - \mathbf{X}\hat{\beta_\tau})' W (\mathbf{y} - \mathbf{X}\hat{\beta_\tau})}{(\mathbf{y} - \bar{y}_\tau 1_n)' W (\mathbf{y} - \bar{y}_\tau 1_n)}\right]$$

[19] A difficulty occurs where $WSST = \sum_{i=1}^n w_i (y_i - \bar{y_\tau})^2$. Anderson and Sprecher suggest comparing the residual sum of squares from the full model with a constant-only model;

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

For regression expectiles, the two models *may not* produce the same weights. This corresponds to the formulation in equations 6.24 and 6.25, where the function ς_τ is included explicitly. Willett'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.

generalized least squares models, there are many options.

The adjusted R^2 statistic, denoted $\bar{R}^2$, was proposed in a textbook by Theil [43] and has become a standard tool for mean regression. The adjusted R^2 for expectiles proposed by Aragon et al. is

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^n \varsigma_\tau(y_i - \mathbf{x}_i'\hat{\beta}_\tau)/(n-\nu)}{\sum_{i=1}^n \varsigma_\tau(y_i - \bar{y_\tau})/(n-1)} \tag{6.24}$$

where the numerator on the right incorporates the weighted sum of squared errors and the denominator incorporates the weighted total sum of squares. We may write these as

$$\begin{aligned} WSSE &= \sum_{i=1}^n \varsigma_\tau(y_i - \mathbf{x}_i'\hat{\beta}_\tau) = \sum_{i=1}^n \hat{w}_i(y_i - \hat{y}_i)^2 \\ WSST &= \sum_{i=1}^n \varsigma_\tau(y_i - \bar{y_\tau}) = \sum_{i=1}^n \tilde{w}_i(y_i - \bar{y_\tau})^2. \end{aligned} \tag{6.25}$$

Note that the estimated weights in the $WSSE$ and $WSST$ need not be the same; see footnote 19. Here, $\tilde{w}_i$ is equal to τ if $y_i \geq \bar{y_\tau}$ and $1-\tau$ otherwise. A degrees-of-freedom adjustment is already in place in equation 6.24, where the numerator is divided by $n-\nu$ and the denominator by $n-1$. The choice of ν 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 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 6.24, 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 3, it is shown that the expected value of $E(\hat{\epsilon}' W \hat{\epsilon}|\mathbf{X}, W)$ is

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

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

$$\tilde{R}^2 = 1 - \frac{\sum_{i=1}^n \varsigma_\tau(y_i - \mathbf{x}_i'\hat{\beta}_\tau)/(\text{trace}(\hat{W}) - \text{trace}(\hat{W}P_\tau))}{\sum_{i=1}^n \varsigma_\tau(y_i - \bar{y_\tau})/(\text{trace}(\tilde{W}) - \text{trace}(\tilde{W}P_\tau^1))}. \tag{6.26}$$

The expected values of $WSST$ is also made clear in Appendix A3. This is arguably the correct 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 19.

As a measure of goodness of fit under an asymmetric loss function, the formulation in 6.26 uses that loss function ς_τ and is entirely appropriate. The other possibility is based on pseudo-R^2 statistics, such as from [47]. The fundamental question is whether we prefer a measure of goodness of fit per the loss function ς_τ 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,

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

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 $\tilde{F}$, which comes

from the full model. The weights in P_τ and P_τ^1 would also be the estimated weights from the full model. This reduces to the pseudo-R^2 of Willett [47] when the degrees of freedom adjustment is omitted and has the same interpretation as the generalized R^2 of Anderson and Sprecher [4] when the estimated weights $\hat{w}_i$ as in equation 5.3 are taken seriously–when the interpretation is that the latent process is truly from this particular $\tilde{F}$.

7 Conclusions

Under modified assumptions, the Gauss-Markov theorem does extend to expectile regression. Thus, we import expectile regression to the classical framework. We find that the expectile regression estimator is the best linear unbiased estimator under a set of assumptions that requires asymmetric variance and a weighted orthogonality condition. Interestingly, the first (linearity) and third (full rank) Gauss-Markov assumptions require no modification. We also show that expectile GLS is the BLUE under asymmetric heteroscedasticity. The generalized (weighted) estimator has been studied only recently and deserves additional attention. For the location-scale model where generalized quantiles attained by different loss functions produce the same sets of regression lines, the τ^{th} expectile GLS estimator is obviously the BLUE for whichever regression line (under any loss function) corresponds to the τ^{th} expectile.

The expectile regression (generalized, when heteroscedasticity is present) is the BLUE in three alternative model designs that have useful interpretations. It is the BLUE when (1) residuals are intended to be some constant other than zero, on average; or (2) we model an observation with atypical odds of positive and negative errors; or (3) data are missing not-at-random and asymmetrically. However, estimators for residual variance and mean squared error are not as simple as in the mean regression: they differ depending on the choice of interpretation and the usual degrees of freedom penalty may be incorrect. Our unbiased sample variance estimator is smaller than the standard estimator, which is biased in every case except the mean regression.

It is interesting that the expectile weights are feasible in some cases. Our demonstration in Section 4 has this property, at least for the simple regression. In that demonstration, we find that an atypical individual (closer than average to the margin) has nearly twice the average expected effect of the variable *black* on his or her probability of being denied a mortgage. Other authors may consider exploiting expectile regression to find hidden results such as these.

This work sheds light on some new ideas, but leaves many questions unexplored. Asymptotic results relating to generalized expectile regression (expectile GLS) are available only very recently (see [8]) and not widely known or studied. 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.

We would summarize our contribution as follows. First, we show that expectile regression is a useful tool for the social sciences. Second, we have drawn new connections between disparate parts of the literature. Expectiles fit into the classical framework. Third, our work casts light on a major target for future research. The entire family of mean regression estimators–not merely GMM models or binary response estimators–may be extended to expectile regression. We cannot overstate the amount of fundamental work that is missing from this field. The proverbial "low-hanging fruit"

are plentiful. As far as this document is concerned, any errors are the sole responsibility of the author.

References

[1] Alexander C Aitken. On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48, 1936.

[2] John H Aldrich, Forrest D Nelson, and E Scott Adler. *Linear probability, logit, and probit models*. Number 45. Sage, 1984.

[3] Ralph C Allen and Jack H Stone. The gauss markov theorem: a pedagogical note. *The American Economist*, 45(1):92–94, 2001.

[4] Richard Anderson-Sprecher. Model comparisons and r 2. *The American Statistician*, 48(2):113–117, 1994.

[5] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

[6] Yves Aragon, Sandrine Casanova, Raymond Chambers, and Eve Leconte. Conditional ordering using nonparametric expectiles. 2005.

[7] Francisco E Balderrama and Raymond Rodríguez. *Decade of betrayal: Mexican repatriation in the 1930s*. UNM Press, 2006.

[8] Amadou Barry, Karim Oualkacha, and Arthur Charpentier. Weighted asymmetric least squares regression for longitudinal data using gee. *arXiv preprint arXiv:1810.09214*, 2018.

[9] Alexander Basilevsky. *Applied matrix algebra in the statistical sciences*. Courier Corporation, 2013.

[10] Richard T Behrens and Louis L Scharf. Signal processing applications of oblique projection operators. *IEEE Transactions on Signal Processing*, 42(6):1413–1424, 1994.

[11] Fabio Bellini, Bernhard Klar, Alfred Müller, and Emanuela Rosazza Gianin. Generalized quantiles as risk measures. *Insurance: Mathematics and Economics*, 54:41–48, 2014.

[12] Annamaria Bianchi, Enrico Fabrizi, Nicola Salvati, and Nikos Tzavidis. Estimation and testing in m-quantile regression with applications to small area estimation. *International Statistical Review*, 86(3):541–570, 2018.

[13] Jens Breckling and Ray Chambers. M-quantiles. *Biometrika*, 75(4):761–771, 1988.

[14] Zehua Chen. Conditional lp-quantiles and their application to the testing of symmetry in non-parametric regression. *Statistics & probability letters*, 29(2):107–115, 1996.

[15] Jan Solomon Cramer. Mean and variance of r2 in small and moderate samples. *Journal of econometrics*, 35(2-3):253–266, 1987.

[16] Abdelaati Daouia, Stéphane Girard, and Gilles Stupfler. Extreme m-quantiles as risk measures: From l1 to lp optimization. *Bernoulli*, pages to–appear, 2019.

[17] Friedhelm Eicker. Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 59–82, 1967.

[18] Carl-Friedrich Gauss. *Theoria combinationis observationum erroribus minimis obnoxiae*, volume 1. Henricus Dieterich, 1823.

[19] Arthur Stanley Goldberger et al. *Econometric theory*. New York: John Wiley & Sons., 1964.

[20] William H Greene. *Econometric analysis*. Pearson Education India, 2003.

[21] James J Heckman and Thomas E MaCurdy. A simultaneous equations linear probability model. *canadian Journal of Economics*, pages 28–37, 1985.

[22] James J Heckman and James M Snyder Jr. Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators. Technical report, National bureau of economic research, 1996.

[23] Hajo Holzmann, Bernhard Klar, et al. Expectile asymptotics. *Electronic Journal of Statistics*, 10(2):2355–2371, 2016.

[24] Peter J Huber. *Robust statistical procedures*, volume 68. Siam, 1996.

[25] Thomas Kneib. Beyond mean regression. *Statistical Modelling*, 13(4):275–303, 2013.

[26] Koenker. *Quantile Regression (Econometric Society monographs; no. 38)*. Cambridge University Press, 2005.

[27] Roger Koenker. When are expectiles percentiles? *Econometric Theory*, 9(3):526–527, 1993.

[28] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

[29] Roger Koenker, Victor Chernozhukov, Xuming He, and Limin Peng. *Handbook of quantile regression*. CRC press, 2017.

[30] Tarald O Kvålseth. Cautionary note about r2. *The American Statistician*, 39(4):279–285, 1985.

[31] Jongkwan Lee, Giovanni Peri, and Vasil Yasenov. The employment effects of mexican repatriations: Evidence from the 1930's. Technical report, National Bureau of Economic Research, 2017.

[32] GS Maddala. Econometrics, 1977. *McGraw Hills Pub. Co., New York*.

[33] Jan R Magnus. Maximum likelihood estimation of the gls model with unknown parameters in the disturbance covariance matrix. *Journal of econometrics*, 7(3):281–312, 1978.

[34] Andreĭ Markov. *Wahrscheinlichkeits-rechnung*. 1912.

[35] Robert G McGillivray. Estimating the linear probability function. *Econometrica (pre-1986)*, 38(5):775, 1970.

[36] Alicia H Munnell, Geoffrey MB Tootell, Lynn E Browne, and James McEneaney. Mortgage lending in boston: Interpreting hmda data. *The American Economic Review*, pages 25–53, 1996.

[37] Whitney K Newey and James L Powell. Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, pages 819–847, 1987.

[38] Collin S Philipps. Quasi-maximum likelihood estimation for conditional expectiles. *Forthcoming*, 2019.

[39] Elmar Spiegel, Fabian Sobotka, Thomas Kneib, et al. Model selection in semiparametric expectile regression. *Electronic Journal of Statistics*, 11(2):3008–3038, 2017.

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

[41] James H Stock, Mark W Watson, et al. *Introduction to econometrics*, volume 104.

[42] M Subrahmanyam. A property of simple least squares estimates. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 355–356, 1972.

[43] Henri Theil. Economic forecasts and policy. 1961.

[44] Elisabeth Waldmann, Fabian Sobotka, and Thomas Kneib. Bayesian geoadditive expectile regression. *arXiv preprint arXiv:1312.5054*, 2018.

[45] Linda Schulze Waltrup, Fabian Sobotka, Thomas Kneib, and Göran Kauermann. Expectile and quantile regression–david and goliath? *Statistical Modelling*, 15(5):433–456, 2015.

[46] Halbert White et al. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *econometrica*, 48(4):817–838, 1980.

[47] John B Willett and Judith D Singer. Another cautionary note about r 2: Its use in weighted least-squares regression analysis. *The American Statistician*, 42(3):236–238, 1988.

[48] Qiwli Yao and Howell Tong. Asymmetric least squares regression estimation: a nonparametric approach. *Journal of Nonparametric Statistics*, 6(2-3):273–292, 1996.

[49] Arnold Zellner. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, 57(298):348–368, 1962.

[50] Xian-Da Zhang. *Matrix analysis and applications*. Cambridge University Press, 2017.

[51] Jun Zhao and Yi Zhang. Variable selection in expectile regression. *Communications in Statistics-Theory and Methods*, 47(7):1731–1746, 2018.

[52] Johanna F Ziegel. Coherence and elicitability. *Mathematical Finance*, 26(4):901–918, 2016.

Appendix

A1: 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 [20].

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 ϵ_i with $y_i - \mathbf{x}_i'\beta_\tau$, we have an assumption on the population moment

$$E(w_i(y_i - \mathbf{x}_i'\beta_\tau)|\mathbf{X}) = 0.$$

This also implies orthogonality given in equation 3.4:

$$E(\mathbf{x}_i w_i(y_i - \mathbf{x}_i'\beta_\tau)|\mathbf{X}) = 0_k. \tag{7.1}$$

The sample counterpart[20] to this is as below, which leads to the estimator. We choose our estimator $\hat{\beta}_\tau$ in order to ensure that equation 3.4 holds in-sample.

$$\begin{aligned} 0_k &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i w_i(y_i - \mathbf{x}_i'\hat{\beta}_\tau) \\ &= \sum_{i=1}^{n} \mathbf{x}_i w_i(y_i - \mathbf{x}_i'\hat{\beta}_\tau) \\ &= \sum_{i=1}^{n} \mathbf{x}_i w_i y_i - \sum_{i=1}^{n} \mathbf{x}_i w_i \mathbf{x}_i'\hat{\beta}_\tau) \\ &= \mathbf{X}'W\mathbf{y} - \mathbf{X}'W\mathbf{X}\hat{\beta}_\tau \end{aligned}$$

Clearly, this implies $\hat{\beta}_\tau = (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{y}$ as would be the case with any weighted least squares problem. This is a *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 $(\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W$ 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 β_τ. First decompose the estimator as follows

$$\begin{aligned} \hat{\beta}_\tau &= (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{y} \\ &= (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W(\mathbf{X}\beta_\tau + \epsilon) \\ &= \beta_\tau + (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\epsilon. \end{aligned} \tag{7.2}$$

[20] 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 $\mathbb{F}_n(y) = n^{-1}\sum_{i=1}^{n} I(y_i \leq y)$, so too can the sample moment conditions be expressed as standard moment conditions with respect to $\tilde{\mathbb{F}}_n$, a weighted empirical distribution conforming to the definition in 2.6.

The question is whether $E(\hat{\beta}_\tau - \beta_\tau) = 0$. We will take W as known and use the tower rule.

$$\begin{aligned} E(\hat{\beta}_\tau - \beta_\tau) &= E(\beta_\tau + (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\epsilon - \beta_\tau) \\ &= E\left((\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\epsilon\right) \\ &= E\left(E\left((\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\epsilon|\mathbf{X},W\right)\right) \\ &= E\left((\mathbf{X}'W\mathbf{X})^{-1}E\left(\mathbf{X}'W\epsilon|\mathbf{X},W\right)\right) \\ &= E\left((\mathbf{X}'W\mathbf{X})^{-1}E\left(\sum_{i=1}^{n}\mathbf{x}_i w_i\epsilon_i|\mathbf{X},W\right)\right) = 0 \end{aligned} \tag{7.3}$$

Then $\hat{\beta}_\tau$ 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 section5.1 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, $\hat{\beta}_\tau$ is a sort of "oracle" estimator because it has perfect foreknowledge of the weights w_i. Otherwise, the unbiased version of $\hat{\beta}_\tau$ is not feasible to estimate and its feasible counterpart with estimated weights may be inferior.

Next we state the variance of our estimator $\hat{\beta}_\tau$. From equations 7.2 and 7.3 we can see that

$$\begin{aligned} Var(\hat{\beta}_\tau|\mathbf{X},W) &= E\left((\hat{\beta}_\tau - \beta_\tau)(\hat{\beta}_\tau - \beta_\tau)'|\mathbf{X},W\right) \\ &= E\left((\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\epsilon\epsilon'W\mathbf{X}(\mathbf{X}'W\mathbf{X})^{-1}|\mathbf{X},W\right) \\ &= (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'WE(\epsilon\epsilon'|\mathbf{X},W)W\mathbf{X}(\mathbf{X}'W\mathbf{X})^{-1} \end{aligned} \tag{7.4}$$

Supposing that the un-weighted non-central variance[21] of $\epsilon|\mathbf{X}$ is $E(\epsilon\epsilon'|\mathbf{X},W) = \Sigma$, we have our solution. This sandwich-type formula[22] would simplify under conditions such as, for instance, $E(W\epsilon\epsilon'|\mathbf{X}) = \nu^2 I_n$ which we assumed in Assumption 4. Then, if that assumption holds[23],

$$Var(\hat{\beta}_\tau|\mathbf{X},W) = \nu^2(\mathbf{X}'W\mathbf{X})^{-1}. \tag{7.5}$$

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, $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, when $\tau = .5$. In practice, we suggest deference to the heteroscedasticity-robust covariance matrix in equation 7.4.

[21] We have $E(\epsilon_i|\mathbf{X}) \neq 0$, so $Var(\epsilon|\mathbf{X}) \neq E(\epsilon\epsilon'|\mathbf{X})$. The identity in equation 3.8 is not applicable because we are interested in the variance around $\mathbf{X}'\beta_\tau$ under the usual distribution of errors, not the *weighted* variance, and the variance or weighted variance around the usual measure of central tendency $E(\mathbf{y}|\mathbf{X})$.

[22] Replacing $E(\epsilon\epsilon'|\mathbf{X},W)$ with the estimator $\hat{\Sigma} = \text{diag}(\hat{\epsilon_i}^2)$ produces the usual sandwich type heteroscedasticity-robust estimator of Eicker [17] or White [46]. That estimator, with the corresponding estimates $\hat{W} = \text{diag}(\hat{w_i})$, is Newey and Powell's estimator [37] of the asymptotic covariance matrix, which those authors prove to be consistent under general conditions.

[23] A standard estimator for equation 7.5 is

$$\hat{\nu}^2\left(\mathbf{X}'\hat{W}\mathbf{X}\right)^{-1} = \frac{1}{n-k}\sum_{i=1}^{n}\hat{w}_i\hat{\epsilon}_i^2\left(\mathbf{X}'\hat{W}\mathbf{X}\right)^{-1}.$$

See the discussion in Section 6.

A2. 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 $P_{.5} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $M_{.5} = I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ are symmetric and idempotent. In the asymmetrically weighted case, we have

$$P_\tau = \mathbf{X}(\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W \neq W\mathbf{X}(\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}' = P_\tau'$$

and

$$M_\tau = I - P_\tau \neq I - P_\tau' = M_\tau'$$

so neither matrix is symmetric. However,

$$\begin{aligned} P_\tau P_\tau &= \mathbf{X}(\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{X}(\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W = \mathbf{X}(\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W = P_\tau \\ M_\tau M_\tau &= (I - P_\tau)(I - P_\tau) = II - IP_\tau - P_\tau I + P_\tau P_\tau = I - P_\tau = M_\tau \end{aligned}$$

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

$$\begin{aligned} P_\tau &= 1_n(1_n'W1_n)^{-1}1_n'W \\ &= \left(\sum_{i=1}^n w_i\right)^{-1} 1_n 1_n'W \\ &= \left(\sum_{i=1}^n w_i\right)^{-1} \begin{pmatrix} w_1 & w_2 & \cdots & w_n \\ w_1 & w_2 & & w_n \\ \vdots & \vdots & \ddots & \vdots \\ w_1 & w_2 & \cdots & w_n \end{pmatrix} \end{aligned}$$

note that left multiplication by P_τ maps the vector y as follows

$$\begin{aligned} P_\tau y &= 1_n(\underbrace{1_n'W1_n}_{=\sum_{i=1}^n w_i})^{-1} \underbrace{1_n'Wy}_{=\sum_{i=1}^n w_i y_i} \\ &= 1_n \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i} \\ &= 1_n \bar{y_\tau} \end{aligned}$$

However, if we write $P_\tau' y$, we have

$$\begin{aligned} P_\tau' y &= W 1_n (1_n' W 1_n)^{-1} 1_n' y \\ &= \left(\sum_{i=1}^n w_i \right)^{-1} \begin{pmatrix} w_1 & w_1 & \cdots & w_1 \\ w_2 & w_2 & & w_2 \\ \vdots & \vdots & \ddots & \vdots \\ w_n & w_n & \cdots & w_n \end{pmatrix} y \\ &= \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n w_i}. \end{aligned}$$

Similarly,

$$\begin{aligned} P_\tau^{1'} P_\tau^1 &= W 1_n (1_n' W 1_n)^{-1} 1_n' 1_n (1_n' W 1_n)^{-1} 1_n' W \\ &= \left(\sum_{i=1}^n w_i \right)^{-2} \begin{pmatrix} w_1 & w_1 & \cdots & w_1 \\ w_2 & w_2 & & w_2 \\ \vdots & \vdots & \ddots & \vdots \\ w_n & w_n & \cdots & w_n \end{pmatrix} \begin{pmatrix} w_1 & w_2 & \cdots & w_n \\ w_1 & w_2 & & w_n \\ \vdots & \vdots & \ddots & \vdots \\ w_1 & w_2 & \cdots & w_n \end{pmatrix} \\ &= n \left(\sum_{i=1}^n w_i \right)^{-2} \begin{pmatrix} w_1^2 & w_1 w_2 & \cdots & w_1 w_n \\ w_2 w_1 & w_2^2 & & w_2 w_n \\ \vdots & \vdots & \ddots & \vdots \\ w_n w_1 & w_n w_2 & \cdots & w_n^2 \end{pmatrix} \end{aligned}$$

whereas,

$$\begin{aligned} P_\tau^1 P_\tau^{1'} &= 1_n (1_n' W 1_n)^{-1} 1_n' W W 1_n (1_n' W 1_n)^{-1} 1_n' \\ &= \left(\sum_{i=1}^n w_i \right)^{-2} \begin{pmatrix} w_1 & w_2 & \cdots & w_n \\ w_1 & w_2 & & w_n \\ \vdots & \vdots & \ddots & \vdots \\ w_1 & w_2 & \cdots & w_n \end{pmatrix} \begin{pmatrix} w_1 & w_1 & \cdots & w_1 \\ w_2 & w_2 & & w_2 \\ \vdots & \vdots & \ddots & \vdots \\ w_n & w_n & \cdots & w_n \end{pmatrix} \\ &= \left(\sum_{i=1}^n w_i \right)^{-2} \sum_{i=1}^n w_i^2 \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \end{aligned}$$

And the difference is clear.

A3. Trace of $M_\tau' W M_\tau$

Here, we show that $\text{trace}(M_\tau' W M_\tau) = \text{trace}(W) - \text{trace}(W P_\tau)$.

The expected value of the weighted sum of squared errors is

$$\begin{aligned} E(WSSE|\mathbf{X}, W) &= E(\hat{\epsilon}' W \hat{\epsilon}|\mathbf{X}, W) \\ &= E(\epsilon' M_\tau' W M_\tau \epsilon|\mathbf{X}, W) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} [M_\tau' W M_\tau]_{ij} \, E(\epsilon_i \epsilon_j|\mathbf{X}, W) \\ &= \sum_{i=1}^{n} [M_\tau' W M_\tau]_{ii} \, \sigma_\tau^2 \end{aligned}$$

which relies on the trace of $M_\tau' W M_\tau$.

$$\begin{aligned} M_\tau' W M_\tau &= (I - P_\tau)' W (I - P_\tau) \\ &= W - W P_\tau - P_\tau' W + P_\tau' W P_\tau. \end{aligned}$$

This is straightforwards. The two negative matrices have the same trace:

$$\begin{aligned} \text{trace}(W P_\tau) &= \text{trace}((\mathbf{X}' W \mathbf{X})^{-1} \mathbf{X}' W W \mathbf{X}) \\ &= \text{trace}\left(\left(\sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \left(\sum_{i=1}^{n} w_i^2 \mathbf{x}_i \mathbf{x}_i'\right)\right) \\ &\lesssim k \end{aligned}$$

and

$$\begin{aligned} \text{trace}(P_\tau' W P_\tau) &= \text{trace}(W \mathbf{X} (\mathbf{X}' W \mathbf{X})^{-1} \mathbf{X}' W \mathbf{X} (\mathbf{X}' W \mathbf{X})^{-1} \mathbf{X}' W) \\ &= \text{trace}((\mathbf{X}' W \mathbf{X})^{-1} \mathbf{X}' W \mathbf{X} (\mathbf{X}' W \mathbf{X})^{-1} \mathbf{X}' W W \mathbf{X}) \\ &= \text{trace}(W P_\tau). \end{aligned}$$

So $\text{trace}(M_\tau' W M_\tau) = \text{trace}(W) - \text{trace}(W P_\tau)$. Then

$$E(WSSE|\mathbf{X}, W) = (\text{trace}(W) - \text{trace}(W P_\tau))\sigma_\tau^2$$

In the special case where $\mathbf{X} = 1_n$, we have

$$\begin{aligned} E(WSST|\mathbf{X}, W) &= E(\hat{\epsilon}' W \hat{\epsilon}|\mathbf{X}, W) \\ &= E((y - \bar{y}_\tau)' W (y - \bar{y}_\tau)|\mathbf{X}, W) \\ &= (\text{trace}(W) - \text{trace}(W P_\tau^1))\sigma_\tau^2 \end{aligned}$$

where

$$\begin{aligned} \text{trace}(W P_\tau^1) &= \text{trace}((1_n' W 1_n)^{-1} 1_n' W W 1_n) \\ &= \frac{\sum_{i=1}^{n} w_i^2}{\sum_{i=1}^{n} w_i} \leq 1. \end{aligned}$$

Note that this value is strictly contained in the unit interval because $w_i^2 < w_i$ for $\tau \in (0, 1)$

Contents