Comparison of Tobit and Quantile Regression with Skewed Normal Censored Data:

Frandsen's Favorite Estimator Wins Again

Carver Coleman, George Garcia, Nicholas Jernigan, Brandon Ly, Zac Pond

I. Introduction:

Without valid data, econometricians are left helpless in their attempt to better understand the world. Unfortunately, the complex nature of data often results in difficulties measuring some variable of interest. One such example in which this arises is censored data. Censored data occurs when, for one of many reasons, your dependent variable is only accurately measured within some detection interval. Any time your dependent variable falls outside this interval, you simply observe either the upper or lower threshold.

Applied work yields countless situations in which censored data arise. These situations include survival analyses, onset studies, and any data collection approach where technological limitations create some detection threshold. As shown in Appendix A, this creates problems for least squares estimation, as estimates are attenuated by a factor equivalent to the proportion of uncensored observations. Thus, the econometrician is left to find an alternate approach.

The standard solution to this problem is the Censored Regression (Tobit) Model, which rests upon the strong assumption that disturbances are normally distributed. Armed with this assumption, one can then formulate a likelihood function to estimate coefficients using Maximum Likelihood Estimation (MLE). Conditional upon the validity of this assumption, the efficiency property of MLE ensures Tobit will be unrivaled in performance. However, not surprisingly, researchers have identified numerous classes of data in which the normality assumption likely does not hold (Wilson 2018). Hence, in these situations, the astute econometrician is left to consider alternative models that may yield more accurate results.

One such model to be considered is Quantile Regression (QR). Rather than estimating the effects on average outcomes as in least squares, QR estimates the effect of treatment on any chosen percentile of your dependent variable. Several different variations of QR, such as weighted QR and non-parametric QR, have been used to consistently estimate parameters in the presence of censoring (Powell 1986, Wang 2009, & Gaunnan 2005). The general proof of QR's consistency amidst censored data is given in Appendix B. Moreover, these properties do not depend on the normality of disturbances, making QR a valid alternative to Tobit.

In this paper, we compare the statistical properties and performance of Tobit and QR in a series of simulation studies. As previously mentioned, we are specifically interested in comparing the estimators in situations where the Tobit normality assumption does not hold. Thus, we deliberately generate 343 unique, distorted normal distributions with various combinations of skewness, spread, and censoring. After comparing both estimators using this vast set of normality assumption violations, we find that QR significantly outperforms Tobit in almost all situations. These findings provide further justification for using QR to account for censored data when the normality assumption seems unlikely.

II. Methodology

Data

To determine the effect of skewness on the bias and efficiency of the Quantile and Tobit estimators, we first consider the following linear equation:

(1) $$y_i^* = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where β_1 is the true effect of x_i on y_i^*, ε_i is a residual term, and $1 \leq i \leq n$ for some $n \in \mathbb{N}$. For simplicity, we generate x_i as a random variable distributed uniformly on the interval $[-1,1]$. The residual ε_i is then generated independent of x_i as a random variable with a skew normal distribution centered at 0 with standard variance 1 (this, of course, implies that ε_i is homoskedastic). Thus, the probability distribution function $f(\varepsilon)$ of the residual is defined by

$$f(\varepsilon) = \frac{2}{\omega}\phi\left(\frac{\varepsilon}{\omega}\right)\Phi\left(\alpha\left(\frac{\varepsilon}{\omega}\right)\right) \tag{2}$$

where $\phi(\cdot)$ is the standard normal probability density function, $\Phi(\cdot)$ is the standard normal cumulative distribution function, α is a skewness parameter, and $\omega > 0$ is a scale parameter.

Now, we introduce censorship into our above equation. Let $y_i = \begin{cases} y_i^* & \text{if } y_i^* < c \\ c & \text{if } y_i^* \geq c \end{cases}$.

We define the cutoff c as

$$c = F^{-1}(\tau) \tag{3}$$

where $F^{-1}(\cdot)$ is the sample inverse cumulative distribution of y_i^* and $\tau > .5$ is a quantile of this distribution.[1] Notice that we consider only right-censorship of (1). Left censorship would lead to analogous results.

Data for the simulation were generated using the *sn* and *stats* packages in R (Azzalini, R Core Team et al).

Tobit

If we assume a normally distributed epsilon, $\varepsilon_i |\, X_i, c \sim N(0, \sigma^2)$ where σ^2 is the variance of ε_i, then we obtain the following likelihood function for observation *i*:

$$L_i = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i^* - \beta_0 - \beta_1 x_i)^2}{\sigma^2}\right) \text{ if } y_i^* < c \\ 1 - \Phi\left(\frac{c - \beta_0 - \beta_1 x_i}{\sigma}\right) \text{ if } y_i^* \geq c \end{cases} .$$

We can then use maximum likelihood to derive our Tobit estimator:

[1] In practice, we used the quantile function from the *stats* package in R to determine the value for *c*

$$\hat{b}^{MLE} = \underset{b}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \ln L_i(y_i, x_i | b)$$

where $b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$.

Quantile Regression

For Quantile Regression, we first define the quantile function:

$$Q_Y(\tau) = q : \Pr(Y \leq q) = \tau$$

where Y is the vector of y_i , $1 \leq i \leq n$. If we assume the error term's τ-quantile is independent of x_i as follows,

$$Q_{\varepsilon|x}(\tau)=0,$$

then we obtain our quantile regression estimator:

$$\hat{b}^{QR} = \underset{b}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - \beta_1 x_i)$$

where $\rho_\tau(\varepsilon) = \big(\tau - 1(\varepsilon < 0)\big)\varepsilon$.

Parameters

Our experiment tests how Tobit and Quantile regression respond under skew normal data and censoring. Notice our Tobit model assumes normality in the residual's distribution, and consequently, the *y*-distribution, while our quantile regression model only assumes independence of the residual and x_i, but not normality.

Our parameters include $n, \beta_1, \tau, \alpha, \omega,$ and *c* defined in equations (1), (2), and (3). We set $\beta_0 = 0$ for simplicity. We manipulate α and ω to determine the skewness of the residual's distribution. Changes in α affect the "slantness" of the distribution, while changes in ω affect the scale of the distribution (see Figures 1-6 below). The cutoff level *c* is unconditional on x_i and is the percentile of our dependent variable y at which any observation lying above this percentile is censored (see equation (3)). Thus, a cutoff quantile of 0.8 signifies that the highest 20% of y-values will uniformly be capped to the y-value of an observation lying at the 80th percentile of our distribution (Figure 7 below). We must be cautious, however, in choosing the right τ given c. By default, we run our quantile regression with $\tau = 0.5$ for low standard errors in the quantile regression estimators. However, a low enough c coupled with a low ω and high α can cause the right end of our data to lack variation at the quantile $\tau = 0.5$. We therefore compensate by generating an automatically adjusted τ to satisfy the assumption that the percent of observations located below c is greater than τ for all x_i (see Figure 7 for details).

Figures 1-6 PDFs (above) and CDFs (below) of y_i given changes in α, ω, and c

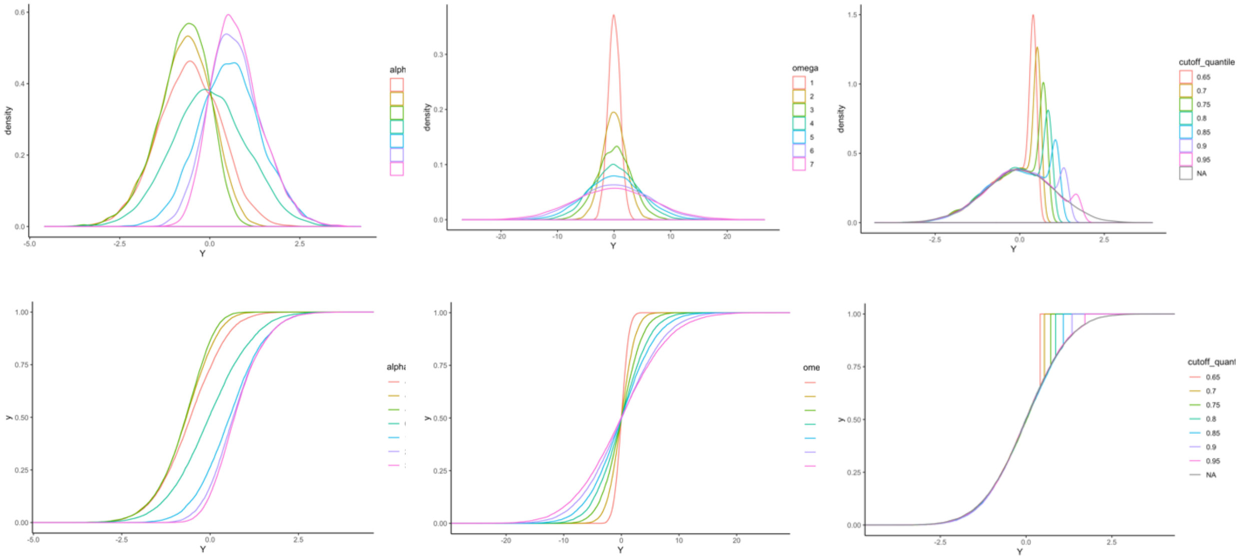
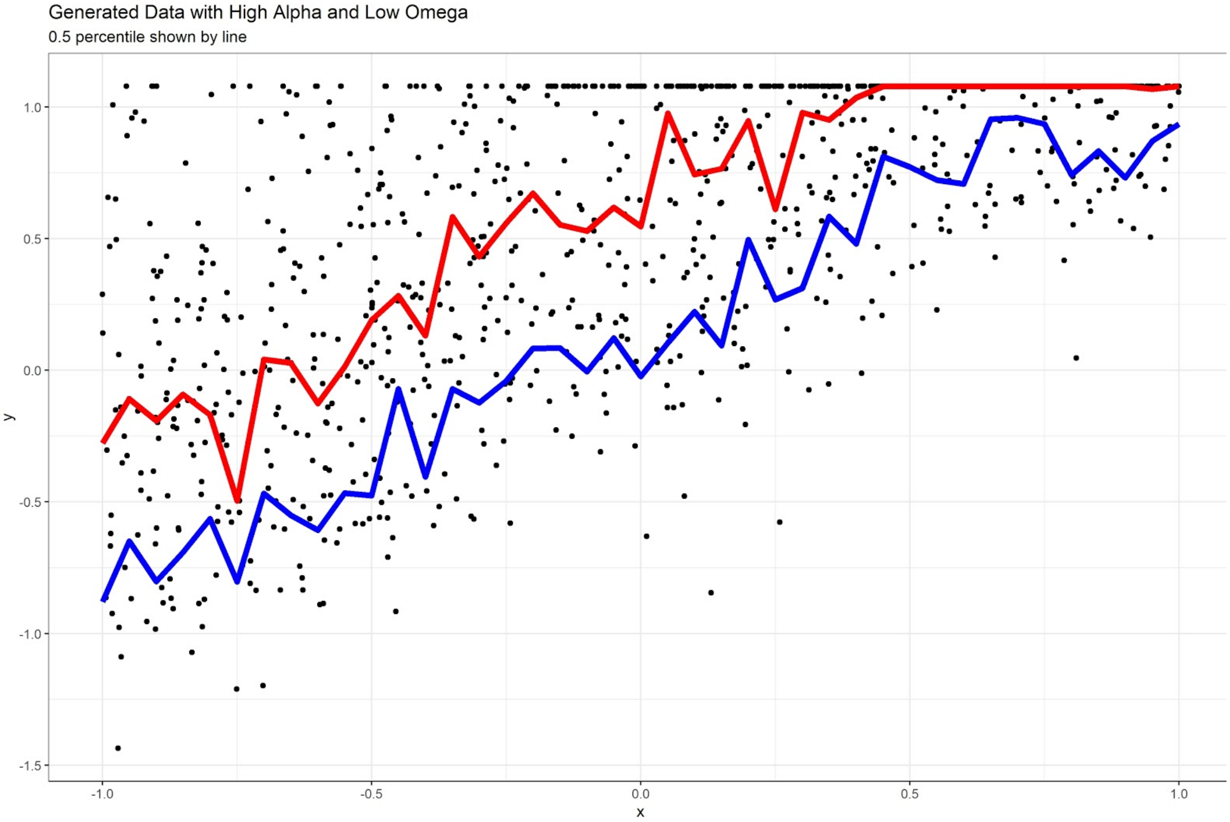


Figure 7



Note: The red line shows the 0.5 quantile (50% of the data is below the line) for the generated data with a low ω-value, high α-value, and a 0.7 quantile cutoff. The assumption in question is that the percentage of data under the cutoff must be greater than or equal to τ for all x. Notice that a 0.5 τ does not satisfy this assumption. The blue line shows the 0.15 quantile for the same generated data. Notice that it, however, never reaches the cutoff level, and would thus be valid to be used in the quantile regression. This process is automated by calculating the percentage of data below the cutoff in the right wing of the data. The τ is set to just below this percentage.

Notice, however, that if we simply set $\beta_1 = 0$, we could avoid the trouble of having to generate τ sufficiently low in response to c. Unfortunately, setting $\beta_1 = 0$ leads to another complication: unbiasedness in the censored regression model (see Appendix B). Consequently, in the case of $\beta_1 = 0$, OLS might be a better estimator than Tobit or quantile regression. We therefore set $\beta_1 = .5$, which still allows for enough variation in the right-tail of our data's distribution after censoring (as the value of β_1 becomes greater, a larger proportion of the data, conditional on x_i being large, becomes censored).

For our simulations, we set *n* = 1000 and run 500 simulations for each combination of our parameters α, ω, and *c*. We consider 7 distinct values for each parameter,

$$\alpha \in \{-3, -2, -1, 0, 1, 2, 3\},$$

$$\omega \in \{1, 2, 3, 4, 5, 6, 7\},$$

$$c = F^{-1}(\tau), \tau \in \{.65, .7, .75, .8, .85, .9, .95\},$$

for a total of 7^3 = 343 different combinations. After generating each dataset, we run Tobit and quantile regressions over each dataset and compare the average bias and mean squared error (MSE) for each parameter combination. Bias for each simulation is calculated as follows:

$$\text{Bias} = \widehat{b_1} - \beta_1$$

where $\widehat{b_1}$ is our regressor coefficient and β_1 = 0.5 by design. We average across all 500 calculated biases to obtain the mean bias for a given α, ω, and *c* combination. MSE is calculated as follows:

$$\text{MSE} = \left(\widehat{b_1} - \beta_1\right)^2$$

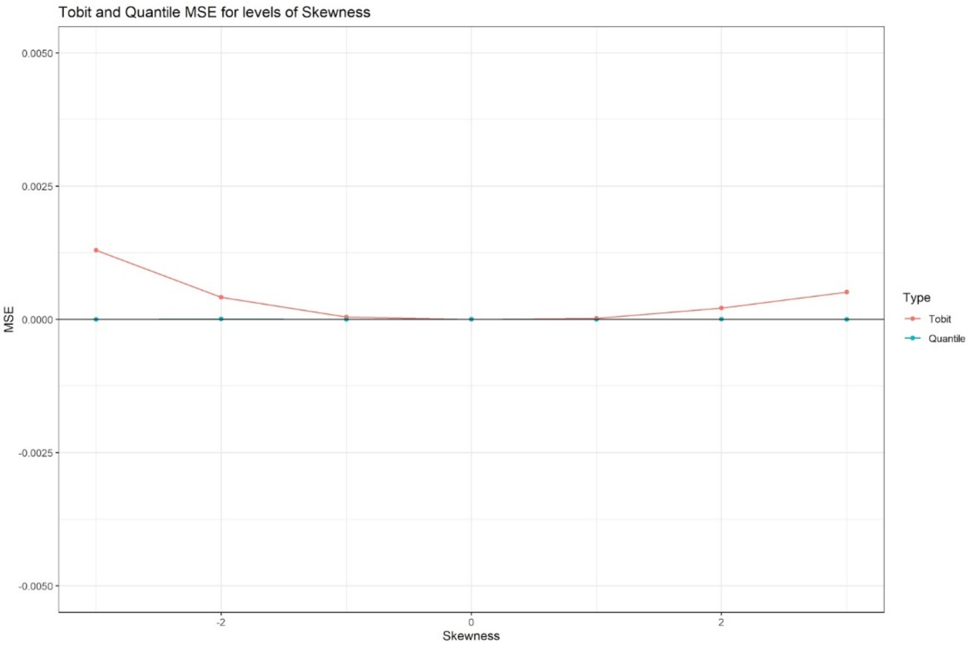
Likewise, we average across all 500 calculated MSEs to obtain the mean MSE for a given α, ω, and *c* combination. All code can be found at www.github.com/1carvercoleman/Censoring-Simulation

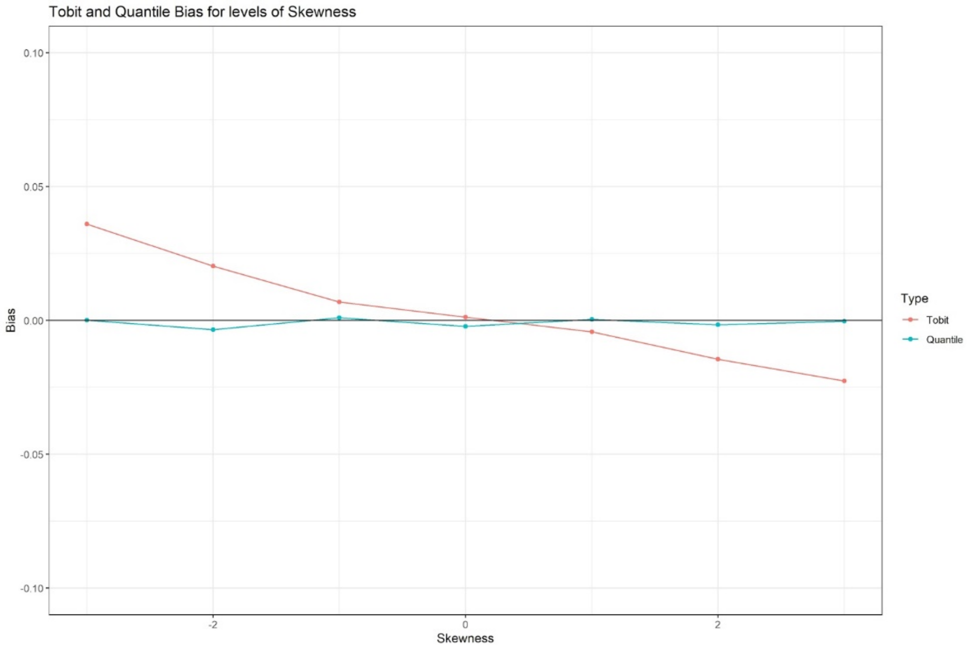
III Results.

The first finding was that the Tobit estimator is not robust to changes in skewness. Both bias and MSE were significantly affected as skewness was adjusted. This is not surprising considering the assumptions the Tobit estimator requires to function properly. As we change the level of skewness from zero, we are effectively changing the distribution from normal to something else. Tobit requires normality to give robust results, our experiment demonstrates that as we take normality away, we bias the estimates Tobit gives us. However, when the assumption of normality is kept, Tobit performed just as well as the quantile regression estimator. We conducted a series of welch two sample t-tests and when the skewness was zero, no matter how we changed spread or the cutoff, the estimates between QR and Tobit were not significantly different from each other (never having a p-value $< .05$).

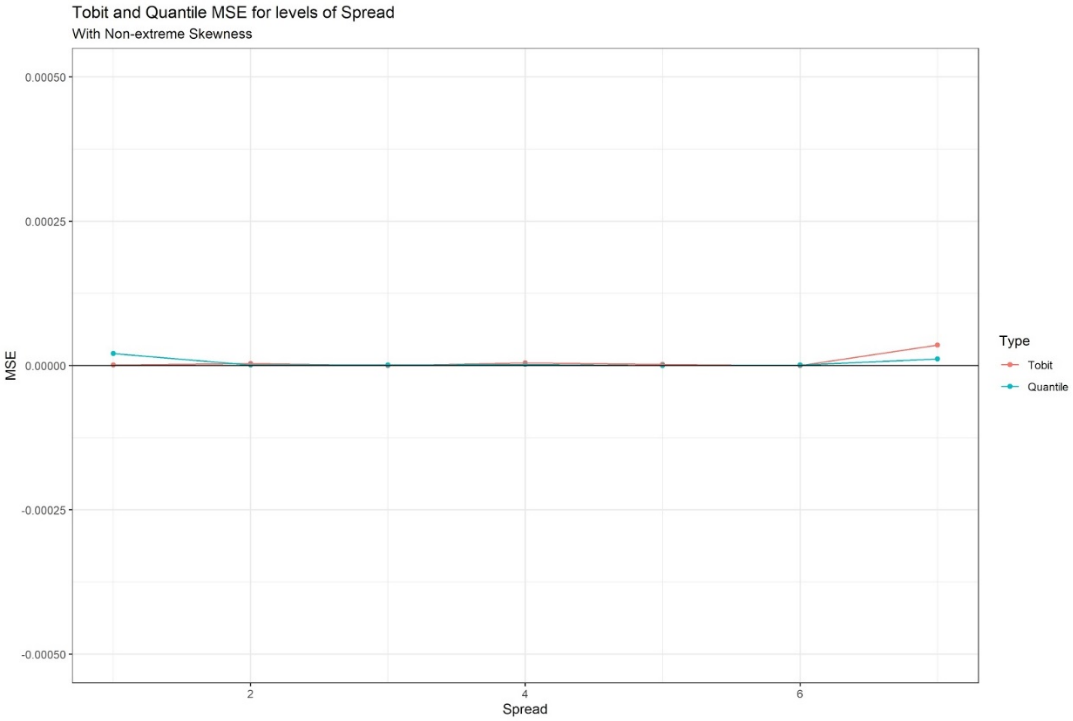
The Tobit estimator was designed specifically to estimate relationships in censored, normally distributed data. In that setting, Tobit is theoretically the best estimator. The interesting finding in this case is that even when the assumption of normality is kept, QR did not perform significantly worse than Tobit. Since in any practical application the assumption of perfect normality is unlikely to be kept, we assume QR is a better estimator. When the assumption of normality was kept in our experiment, QR and Tobit gave similar estimates, suggesting that QR must not be worse than Tobit even when normality holds.

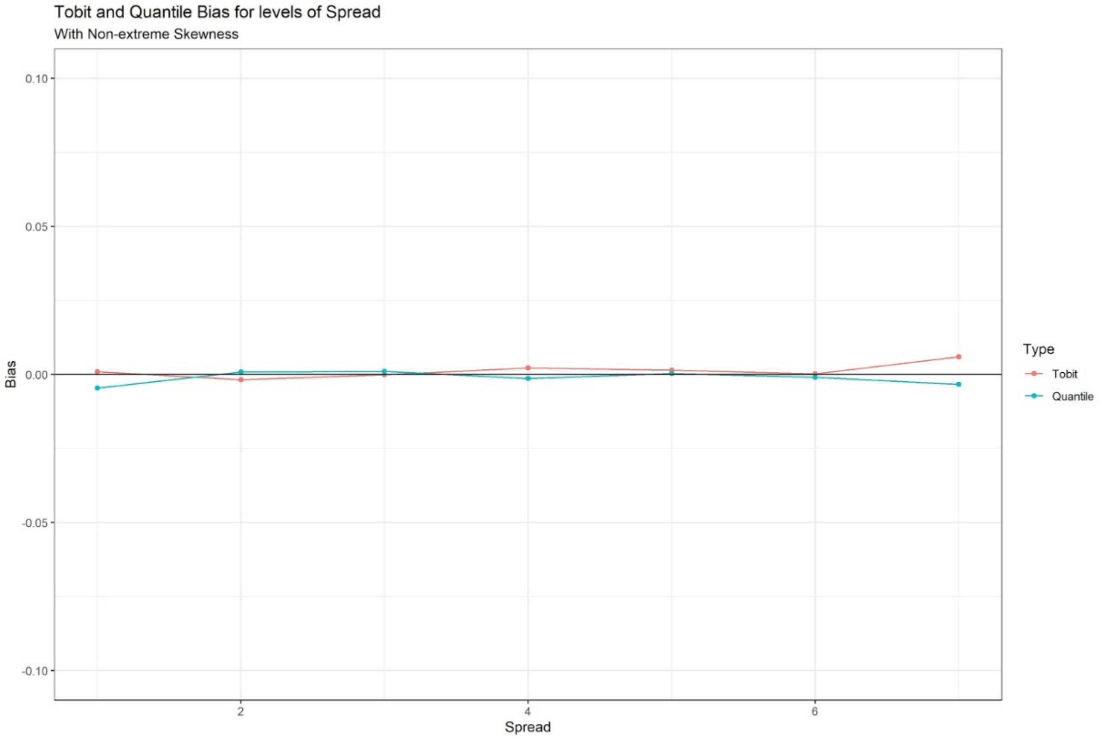
To contrast with Tobit, QR estimators were robust to changes in skewness. In every experiment where the skewness deviated from zero, the Tobit and QR estimates were significantly different from each other. QR was always less biased and had less MSE than Tobit. This is consistent with the assumptions, or rather lack of assumptions, that QR needs in order to be robust. Unlike Tobit, QR makes no assumptions about skewness or heteroskedasticity, making it robust to both. These findings vindicate the theory which suggests that in most real-life settings when data is collected, QR will give more accurate parameters than Tobit since it is unlikely that collected data will have normally distributed error terms.

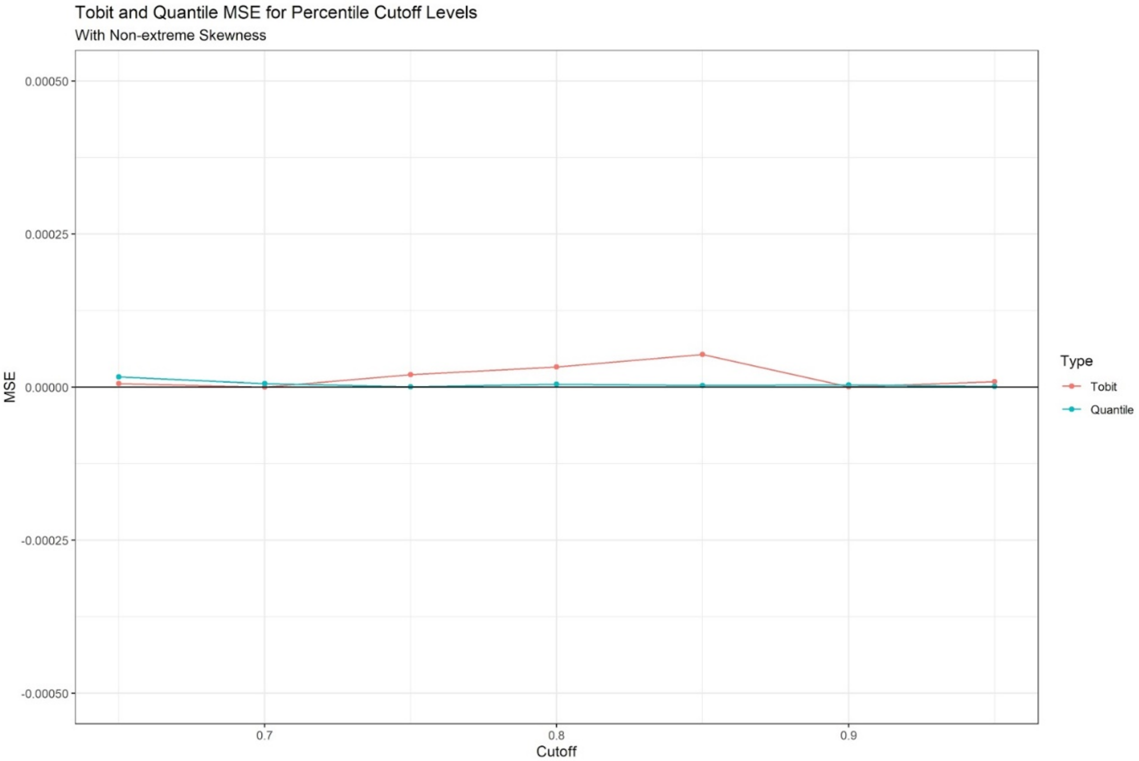

Tobit and Quantile MSE for levels of Skewness
0.0050
0.0025
0.0000
-0.0025
-0.0050
MSE
-2
0
2
Skewness
Type
Tobit
Quantile

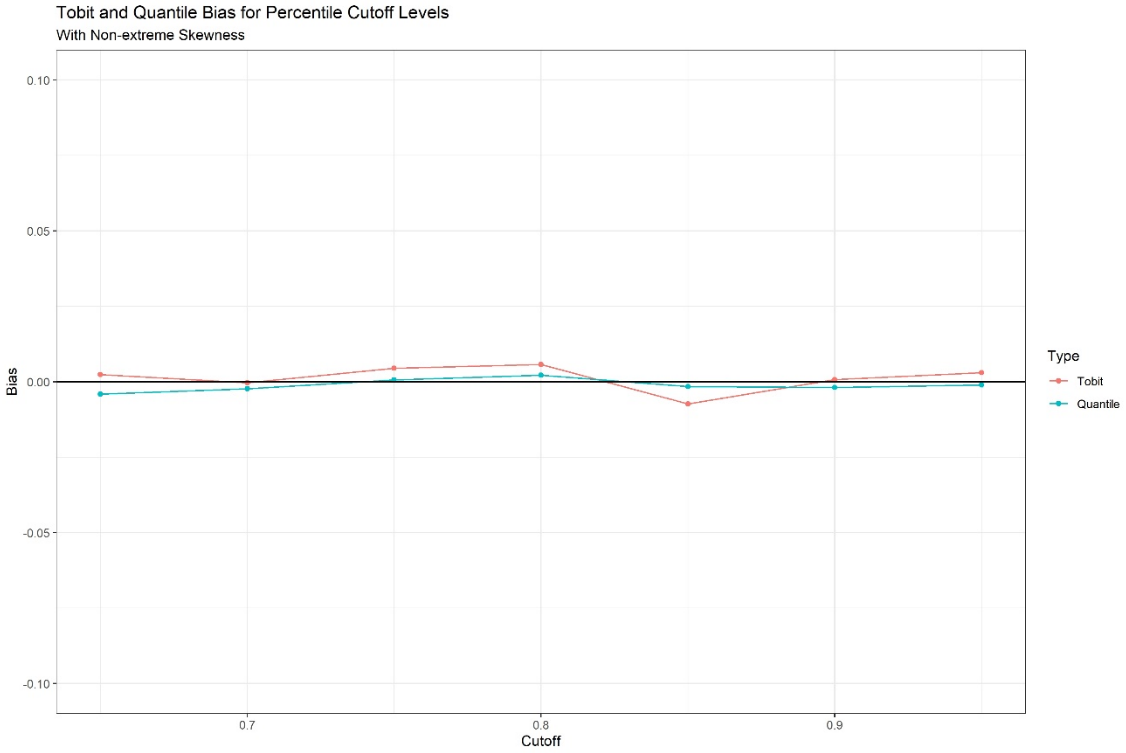

Tobit and Quantile Bias for levels of Skewness
0.10
0.05
0.00
-0.05
-0.10
Bias
-2
0
2
Skewness
Type
Tobit
Quantile

Our second and more mysterious finding was that spread did not affect the performance of QR or Tobit. While it appears on our graphs below that as spread decreases, Tobit gains an advantage over QR, when we conducted our welch two sample t-test, these differences were not statistically significant. A possible explanation for this lack of effect is that as spread decreases, the amount of unaccounted variance decreases, which would be strange if it significantly affected the results of either QR or Tobit.




Tobit and Quantile Bias for levels of Spread
With Non-extreme Skewness
0.10
0.05
0.00
-0.05
-0.10
Bias
2
4
6
Spread
Type
Tobit
Quantile


Tobit and Quantile MSE for Percentile Cutoff Levels
With Non-extreme Skewness
0.00050
0.00025
0.00000
-0.00025
-0.00050
MSE
0.7
0.8
0.9
Cutoff
Type
Tobit
Quantile



Above are the graphs sharing our results for how changes in the cutoff level affected the MSE and bias of Tobit and quantile regressions. The graphs do not show a consistent pattern for the effect of cutoff on MSE or bias. Consistent with this observation, when we conducted our welch two sample t-tests, we found that there were no cutoff changes that made the estimates from QR and Tobit significantly different from each other. Cutoff is similar to spread in that the ability to handle changes in either parameter never varied significantly between the two estimators.

Correlation Coefficients on Bias	Quantile Regression	Tobit
Skewness	.0006583	-.0937795
Spread	.0022126	.0030588
Cutoff	.0030629	-.0005127

The last results from our experiment that summarize our findings are the coefficients in the correlation matrix when you compare the difference between Tobit and QR in how skewness, spread, and cutoff affected bias. The only significant coefficient was the effect skewness had on Tobit. As discussed above, this was expected since skewness violates the assumption of normality that Tobit requires to be robust. The potentially surprising finding from our experiment is that both QR and Tobit were equally robust to changes in spread and cutoff. The only situation where each estimator deviated was changes in skewness, where QR was consistently superior to Tobit. There were no situations where Tobit performed significantly better than QR. This suggests that QR is superior to Tobit in almost all applied situations if all we are considering are changes in skewness, spread, and cutoff.

IV Conclusion.

Our results confirm the need for normality in order to effectively do a robust Tobit regression. As skewness was adjusted, Tobit lost its accuracy, while QR remained robust. We found that neither cutoff or spread could consistently bias QR or Tobit. We also demonstrated that where Tobit struggles to be consistent and unbiased, QR perform well consistently. Even in the conditions where Tobit's key assumptions held, the estimates given by QR were not significantly different than the ones given by Tobit. Considering this, our results suggest that QR is often better than Tobit regression for censored data.

Appendix

Appendix A: Partial Effects in the Censored Regression Model

Consider the censored regression model defined in equations (1) through (3) and let $f(\varepsilon)$ and $F(\varepsilon)$ denote the density and cdf of ε_i. By construction, ε_i is a continuous random variable with mean 0 and variance σ^2, and $f(\varepsilon \mid x) = f(\varepsilon)$. Then

$$\partial E\ [y \mid \boldsymbol{x}]\ /\partial x\ = \beta_1 \times Prob[\ y * < \ c].$$

(see Greene's *Econometric Analysis* for proof of Theorem 19.4)

Notice that $\beta_1 = 0$ implies that $\partial E\ [y \mid \boldsymbol{x}]\ /\partial x\ = \beta_1 = 0$. Thus, censoring does not lead to biasedness in OLS estimator in the case of $\beta_1 = 0$.

Appendix B: Proof that Quantile Regression can be used to estimate a censored model

Consider the model defined in equations (1)-(3). We first show that $Q_{Y^*|x}(\tau) = \beta_0(\tau) + \beta_1 x_i$ for $\beta_0(\tau) = \beta_0 + Q_\varepsilon(\tau)$

By definition this is true of $\Pr\left(Y_i^* \leq \beta_0(\tau) + \beta_1 x_i | x_i\right) = \tau$, thus

$$\begin{aligned}\Pr\left(Y_i^* \leq \beta_0(\tau) + \beta_1 x_i | x_i\right) &= \Pr\left(\beta_0 + \beta_1 x_i + \varepsilon_i \leq \beta_0 + Q_\varepsilon(\tau) + \beta_1 x_i | x_i\right) \\ &= \Pr\left(\varepsilon_i \leq Q_\varepsilon(\tau) | x_i\right) \\ &= \tau \text{ by definition.}\end{aligned}$$

We now show that $Q_{Y|x_i}(\tau) = \min\{\beta_0(\tau) + \beta_1 x_i, c\}$.

Case 1: $\beta_0(\tau) + \beta_1 x_i < c$. In that case, the event $\{Y \leq \beta_0(\tau) + \beta_1 x_i\}$ is equivalent to $\{Y^* \leq \beta_0(\tau) + \beta_1 x_i\}$ which we showed in part (a) has probability τ. So in this case $Q_{Y|x_i}(\tau) = \beta_0(\tau) + \beta_1 x_i$ by definition.

Case 2: $\beta_0(\tau) + \beta_1 x_i \geq c$. In this case, the event $\{Y \leq \beta_0(\tau) + \beta_1 x_i\}$ has probability one, since Y is never higher than c. Consider $\Pr(Y \leq q)$ for any q less than c. Then $\Pr(Y \leq q) = \Pr(Y^* \leq q) < \tau$ by definition (since we know $q < Q_{Y^*|x}(\tau) = \beta_0(\tau) + \beta_1 x_i$. Therefore,

$$\inf\{q : \Pr(Y \leq q | x_i) \geq \tau\} = c.$$

Since $Q_{Y|x_i}(\tau) = \min\{\beta_0(\tau) + \beta_1 x_i, c\}$, it follows we can estimate β_1 by:

$$\left(\hat{b}_0(\tau), \beta_1\right) = \arg\min_{b_0, b_1} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau\left(y_i - \min\{b_0 + b_1 x_i, c\}\right).$$

References

Azzalini, Adelchi. *Package 'sn'* (2005). https://cran.r-project.org/web/packages/sn/sn.pdf.

Gannoun, Ali, et al. "Non-parametric quantile regression with censored data." *Scandinavian Journal of Statistics* 32.4 (2005): 527-550.

Greene, William H. "Limited Dependent Variables—Truncation, Censoring, and Sample Selection". *Econometric Analysis*. Upper Saddle River: Pearson Education, Inc., 2012. 849.

Powell, James L. "Censored regression quantiles." *Journal of econometrics* 32.1 (1986): 143-155.

R Core Team et al. "stats-package: The R Stats Package." *rdrr.io*. https://rdrr.io/r/stats/stats-package.html

Wang, Huixia Judy, and Lan Wang. "Locally weighted censored quantile regression." *Journal of the 'American Statistical Association* 104.487 (2009): 1117-1128.

Wilson, Theodore, Tom Loughran, and Robert Brame. "Substantial bias in the Tobit estimator: making a case for alternatives." *Justice Quarterly* 37.2 (2020): 231-257.