MIT Sloan School of Management

Working Paper 4451-03
December 2003

Censored Regressors and Expansion Bias

Roberto Rigobon, Thomas M. Stoker

Censored Regressors and Expansion Bias

Roberto Rigobon   Thomas M. Stoker*

November 2003

Abstract

We show how using censored regressors leads to *expansion bias,* or estimated effects that are proportionally too large. We show the necessity of this effect in bivariate regression and illustrate the bias using results for normal regressors. We study the bias when there is a censored regressor among many regressors, and we note how censoring can work to undo errors-in-variables bias. We discuss several approaches to correcting expansion bias. We illustrate the concepts by considering how censored regressors can arise in the analysis of wealth effects on consumption, and on peer effects in productivity.

1. Introduction

When the values of the dependent variable of a linear regression model are censored, the OLS estimates of the regression coefficients are biased. This familiar fact is a standard lesson covered in textbooks on econometrics. It has stimulated a great deal of work on consistent estimators of coefficients when there is a censored dependent variable.

In view of this, it seems surprising that very little attention has been paid to implications of censoring of an independent variable, or regressor, in estimation of a linear model. Indeed, it would seem that researchers encounter censored regressors as often or even more often than situations of censored dependent variables. Consider how often variables are observed in ranges, including unlimited top and bottom categories. For instance, observed household income is often recorded in increments of one thousand or five thousand dollars, but would have a top-coded response of, say, "$100,000 and above." Here we are interested in what difference it makes if we estimate a regression with the top-coded income data when the correct specification has income level (no top coding) as the appropriate regressor.

We show that using a censored regressor results in *expansion bias* in OLS estimates of regression coefficients, namely, estimated effects will be too large in absolute value.[1] For instance, if income is

* Sloan School of Management, MIT, 50 Memorial Drive, Cambridge, MA 02142 USA. We have received valuable comments from Vincent Hogan, Dale Jorgenson and Whitney Newey.

[1] Expansion bias is the opposite of *attenuation bias*, familiar from problems such as errors-in-variables or simple censored dependent variable models.

top-coded (or bottom-coded, or both), a positive income effect will be overestimated. This fact is straightforward to show, and several approaches for correcting it can be proposed.

When there are many regressors, the bias associated with using a censored regressor is more complicated, but some useful understanding of it can be developed. We supplement general derivations with exact formulae for the case when all regressors are normally distributed. We study an interesting side issue, which is what bias occurs when a regressor that is measured with error is also censored. In that case the two sources of bias work in opposite directions, and can cancel each other out. We develop this relationship as well to show some useful trade-offs in these bias problems.

We note at least three approaches for correcting expansion bias in empirical work. When the form of the distribution of the regressors is unknown, there is a semiparametric approach that appears to be the most efficient method.

We discuss how censored regressors can arise in two specific application areas; the estimation of wealth effects on consumption, and the estimation of peer group effects on productivity. Our discussion is intended to give concrete illustration to the ideas, and we plan to carry out applications of these kinds as part of future research.

Expansion bias from censored regressors is a straightforward problem, but ignoring that problem can lead to substantial mismeasurement of effects. Section 2 shows how expansion bias arises, and considers several related topics including what can be said when there are many regressors. Section 3 discusses corrections for expansion bias, Section 4 discusses the application areas, and Section 5 gives some concluding remarks.

2. Censored Regressors and Expansion Bias

2.1. The Basic Problem

We start with bivariate regression to see the simplest form of expansion bias. Suppose that the true model is

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad i = 1, ..., n \tag{1}$$

where x_i is the (uncensored) regressor and ε_i is the disturbance. We assume that the distribution of (x_i, ε_i) has finite second moments and obeys $E(\varepsilon_i | x_i) = 0$. Suppose that x_i^{cen} denotes the censored version of x_i with bottom-coding and top-coding, as

$$x_i^{cen} = x_i \cdot 1\left[\xi^- \leq x_i \leq \xi^+\right] + \xi^- \cdot 1\left[x_i < \xi^-\right] + \xi^+ \cdot 1\left[\xi^+ < x_i\right] \tag{2}$$

where ξ^-, ξ^+ are scalars, $\xi^- < \xi^+$. In words, x_i^{cen} is x_i when in the range $\xi^- \leq x_i \leq \xi^+$, but is equal to the respective lower limit ξ^- or upper limit ξ^+ when x_i falls out of that range. The question of interest is what happens when we use x_i^{cen} to estimate β; namely if we estimate the model

$$y_i = a + b x_i^{cen} + u_i \quad i = 1, ..., n, \tag{3}$$

then what relation does the OLS coefficient $\hat{b}$ have to β?

A picture makes the answer clear. Figure 1 shows a scatterplot of data without censoring of the regressor. The small circles are the resulting data points when the regressor is censored at upper and lower bounds. The estimated regression using the censored regressor clearly has a steeper slope that the one using the uncensored regressor, because of the "pile-up" of observations at each limit. This is what we call *expansion bias*. Moreover, it is clear that expansion bias would result if there was only one-sided censoring (top-coding or bottom-coding only); that each censoring limit contributes to the amount of expansion bias.

A more formal analysis reflects these features as well. We begin by stating the conclusion as:

Proposition 1. *Provided that* $0 < \Pr\left\{x < \xi^-\right\} + \Pr\left\{x > \xi^+\right\} < 1$, *we have*

$$plim\ \hat{b} = \beta \cdot (1 + \Lambda) \tag{4}$$

where $\Lambda > 0$.

The proof of Proposition 1 is direct, and given in the following subsection (it may be skipped in a quick reading). It shows that the relative bias is

$$\Lambda = \Lambda^- + \Lambda^+, \tag{5}$$

where Λ^- and Λ^+ are positive terms that arise from censoring at the bottom limit and the top limit respectively, exactly in line with Figure 1. This is the main result on censoring with bivariate regression. After the proof, we show the size of the bias when the uncensored regressor is normally distributed, and then we develop the bias in the context of many regressors in Section 2.2.

2.1.1. Proof of Proposition 1

Begin by defining the parts of x_i that are omitted by censoring.

$$x_i^- = \left(x_i - \xi^-\right) \cdot 1\left[x_i < \xi^-\right] \quad \text{and} \quad x_i^+ = \left(x_i - \xi^+\right) \cdot 1\left[\xi^+ < x_i\right], \tag{6}$$

so that the true model (1) appears as

$$y_i = \alpha + \beta x_i^{cen} + \beta x_i^- + \beta x_i^+ + \varepsilon_i \quad i = 1, ..., n \tag{7}$$

The omitted variable bias formula then gives

$$\text{plim}\ \hat{b} = \beta \cdot \left(1 + \Lambda^- + \Lambda^+\right) \tag{8}$$

where Λ^-, Λ^+ are the auxiliary coefficients from omitting x_i^-, x_i^+ respectively, namely

$$\Lambda^- = \frac{Cov\left(x^-, x^{cen}\right)}{Var\left(x^{cen}\right)} \quad \text{and} \quad \Lambda^+ = \frac{Cov\left(x^+, x^{cen}\right)}{Var\left(x^{cen}\right)}. \tag{9}$$

Consider Λ^- first, and assume that $\Pr\left\{x < \xi^-\right\} > 0$ (since if $\Pr\left\{x < \xi^-\right\} = 0$, then $\Lambda^- = 0$). We have

$$\begin{aligned} Cov\left(x^-, x^{cen}\right) &= E\left(x^- \cdot x^{cen}\right) - E\left(x^-\right) E\left(x^{cen}\right) \\ &= E\left(\xi^-\left(x - \xi^-\right) 1\left[\left(x - \xi^-\right) < 0\right]\right) - E\left(x^-\right) E\left(x^{cen}\right) \\ &= \left[\xi^- - E\left(x^{cen}\right)\right] \cdot E\left(x^-\right) \\ &= E\left(\xi^- - x^{cen}\right) \cdot E\left(x^-\right) \end{aligned} \tag{10}$$

Each of the expectations $E\left(\xi^- - x^{cen}\right)$ and $E\left(x^-\right)$ is an integral with nonpositive integrand, and each integrand is strictly negative over a range of positive probability. Therefore, each of the expectations is strictly negative, and their product $Cov\left(x^-, x^{cen}\right)$ is strictly positive. We conclude that

$$\Lambda^- = \frac{Cov\left(x^-, x^{cen}\right)}{Var\left(x^{cen}\right)} > 0 \tag{11}$$

so that omitting x_i^- results in expansion bias.

Similarly for Λ^+, we assume $\Pr\left\{x > \xi^+\right\} > 0$, and derive

$$Cov\left(x^+, x^{cen}\right) = E\left(\xi^+ - x^{cen}\right) \cdot E\left(x^+\right) \tag{12}$$

Here, each of the two expectations is an integral with nonnegative integrand, and each integrand is strictly positive over a range of positive probability. Therefore, each of these expectations is strictly positive, as is their product $Cov\left(x^+, x^{cen}\right)$, and we conclude that

$$\Lambda^+ = \frac{Cov\left(x^+, x^{cen}\right)}{Var\left(x^{cen}\right)} > 0 \tag{13}$$

so that omitting x_i^+ also results in expansion bias. This shows Proposition 1, where $\Lambda = \Lambda^- + \Lambda^+$.[2]

2.1.2. Expansion Bias with a Normal Regressor

The expansion bias Λ depends on various expectations over truncated distributions. Assuming a particular distributional form will often allow the bias to be computed directly. In econometrics, the most familiar formulae for truncated and censored expectations are from a normal distribution. In this section we illustrate the bias assuming that the uncensored regressor x is normally distributed with mean μ_x and variance σ_x^2. To examine the bias at different levels of censoring, we can parametrize using the censoring limits ξ^-, ξ^+, or equivalently, the probabilities of censoring $p^- = \Pr\left(x < \xi^-\right)$, $p^+ = \Pr\left(x > \xi^+\right)$. We choose the latter, because of a sort of 'scale free' intuition.

[2] It should be noted that Λ^- (or Λ^+) is slightly larger than the bias term that would arise if there was only censoring of low values (or high values). That bias term is in the form (11) (or (13) respectively) with the same numerator but slightly larger denominator (since x^{cen} is only censored on one side). More on this in the discussion of Figure 2.

The bias formulae themselves are complicated and do not admit to obvious interpretation. Because of that, we give a brief derivation and show the formulae in Appendix A. Here we illustrate the bias graphically.

Figure 2 gives two depictions of expansion bias. The solid line displays the two-sided bias Λ of (5) under the assumption of symmetric censoring, with the same probability $p^- = p^+ \equiv p$ of censoring in the high and low region, and is plotted against the total probability of censoring $2p$. The dashed line is the expansion bias from one-sided censoring, or top-coded data, which is plotted with the same total censoring probability. For instance, plotted over $2p = .2$ is the two-sided bias from censoring 10% in each tail, and the one-sided bias from censoring 20% in the upper tail. For comparison, the diagonal $(2p, 2p)$ is included as the dotted line.

We see that the bias is roughly linear in $2p$ for low censoring levels, up to around 30% of the data censored. After that the bias rises nonlinearly, but a bias that doubles the coefficient value involves a lot of censoring; 60% or more of the data.

The two-sided bias is greater than the one-sided bias over the whole range of probabilities. So, in this sense, censoring on both sides induces more than twice the bias of censoring on one side only. This is likely due to the fact that with two-sided censoring, the censored points are in two separated groups and therefore have more influence on the estimated regression.

2.2. Expansion Bias and Several Regressors

We now study expansion bias when there are one or more regressors. We extend the model (1) to include a k-vector or regressors z_i as

$$y_i = \alpha + \beta x_i + \gamma' z_i + \varepsilon_i \quad i = 1, ..., n \tag{14}$$

where we assume that the distribution of $\left(x_i, z_i', \varepsilon_i\right)$ is nonsingular, has finite second moments and obeys $E\left(\varepsilon_i | x_i, z_i\right) = 0$. Again, we are interested in what happens when x_i^{cen} is used instead of x_i; so if we estimate the model

$$y_i = a + b x_i^{cen} + c' z_i + u_i \quad i = 1, ..., n, \tag{15}$$

then how are the OLS coefficients $\hat{a}$, $\hat{b}$, $\hat{c}$ biased as estimators of α, β, γ?

We can develop a deeper understanding of expansion bias by viewing the OLS estimation in (15) as a pooled regression. As discussed further in Section 3, no bias is generated by observations that are not censored, or 'truncated' observations. To use this fact, it is valuable to define the following approximation device:

$$x_i^{(\eta)} = \begin{cases} \eta x_i + (1-\eta)\,\xi^- & \text{if } x_i < \xi^- \\ x_i & \text{if } \xi^- \leq x_i \leq \xi^+ \\ \eta x_i + (1-\eta)\,\xi^+ & \text{if } \xi^+ < x_i \end{cases} \tag{16}$$

where η transforms between the uncensored and censored regressor, as $x_i^{(1)} = x_i$ and $x_i^{(0)} = x_i^{cen}$.

We write the true model (14) in terms of $x_i^{(\eta)}$, and approximate the estimation of (15) as a pooled regression of the three samples, the ‘low censored’ with $x_i < \xi^-$, the ’truncated’ with $\xi^- \leq x_i \leq \xi^+$, and the ‘high censored’ with $\xi^+ < x_i$. For ‘low censored’ observations, we have

$$y_i = \left[\alpha + \beta\xi^- \left(1 - \frac{1}{\eta}\right)\right] + \left[\beta + \beta\left(\frac{1}{\eta} - 1\right)\right] x_i^{(\eta)} + \gamma' z_i + \varepsilon_i \ ; \quad \text{if } x_i < \xi^-, \tag{17}$$

for ’truncated’ observations we have

$$y_i = \alpha + \beta x_i^{(\eta)} + \gamma' z_i + \varepsilon_i \ ; \quad \text{if } \xi^- \leq x_i \leq \xi^+ \tag{18}$$

and for ’high censored’ observations we have

$$y_i = \left[\alpha + \beta\xi^+ \left(1 - \frac{1}{\eta}\right)\right] + \left[\beta + \beta\left(\frac{1}{\eta} - 1\right)\right] x_i^{(\eta)} + \gamma' z_i + \varepsilon_i \ ; \text{ if } \xi^+ < x_i \tag{19}$$

Denote the OLS coefficients from regression y_i on a constant, $x_i^{(\eta)}$ and z_i as $\hat{a}^{(\eta)}$, $\hat{b}^{(\eta)}$, $\hat{c}^{(\eta)}$. The bias in those estimates is given as

$$\text{plim} \begin{pmatrix} \hat{a}^{(\eta)} \\ \hat{b}^{(\eta)} \\ \hat{c}^{(\eta)} \end{pmatrix} - \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = [\Omega_\eta]^{-1} \left[p^- \Omega_\eta^- \begin{pmatrix} \beta\xi^- \left(1 - \frac{1}{\eta}\right) \\ \beta\left(\frac{1}{\eta} - 1\right) \\ 0 \end{pmatrix} + p^+ \Omega_\eta^+ \begin{pmatrix} \beta\xi^+ \left(1 - \frac{1}{\eta}\right) \\ \beta\left(\frac{1}{\eta} - 1\right) \\ 0 \end{pmatrix} \right] \tag{20}$$

where $p^- = \Pr\left\{x < \xi^-\right\}$ is the probability of ‘low censoring,’ $p^+ = \Pr\left\{x > \xi^+\right\}$ is the probability of ’high censoring,” and

$$\Omega_\eta = E\left[\left(1, x^{(\eta)}, z'\right)' \left(1, x^{(\eta)}, z'\right)\right] = \left[p^- \Omega_\eta^- + (1 - p^- - p^+)\Omega_\eta^{trun} + p^+ \Omega_\eta^+\right], \tag{21}$$

where Ω_η^-, Ω_η^{trun}, Ω_η^+ are the conditional second moment matrices over the ‘low censoring’ region $x < \xi^-$, the truncated region $\xi^- \leq x_i \leq \xi^+$, and the ‘high censoring’ region $x > \xi^+$, respectively. The bias (20) is a matrix-weighted average of the biases in the low and high censoring regions; with expansion bias terms $\beta\left(\frac{1}{\eta} - 1\right)$ from each region. As $\eta \rightarrow 0$, these bias terms explode, and so for the approximation device to be useful, we must verify that the weights Ω_η^-, Ω_η^+ shrink at the same rate.

This is easy to do. By spelling out the moments as $\eta \rightarrow 0$, it is immediately clear that

$$E\left[\left(y, 1, x^{(\eta)}, z'\right)' \left(y, 1, x^{(\eta)}, z'\right)\right] \rightarrow E\left[\left(y, 1, x^{cen}, z'\right)' \left(y, 1, x^{cen}, z'\right)\right] \tag{22}$$

so that

$$\begin{pmatrix} \hat{a}^{(\eta)} \\ \hat{b}^{(\eta)} \\ \hat{c}^{(\eta)} \end{pmatrix} \rightarrow \begin{pmatrix} \hat{a} \\ \hat{b} \\ \hat{c} \end{pmatrix} \tag{23}$$

and

$$\Omega_\eta \to E\left[\left(1, x^{cen}, z^{'}\right)^{'}\left(1, x^{cen}, z^{'}\right)\right] \equiv \Omega. \tag{24}$$

The bias components can be computed directly, for instance

$$p^{+}\Omega_\eta^{+}\left(\begin{array}{l} \beta\xi^{+}\left(1-\frac{1}{\eta}\right) \\ \beta\left(\frac{1}{\eta}-1\right) \\ 0 \end{array}\right) = p^{+}\beta\left(1-\eta\right)\left(\begin{array}{l} \mu_x^{+}-\xi^{+} \\ \eta\left(M_{xx}^{+}-\mu_x^{+}\xi^{+}\right)+\left(1-\eta\right)\xi^{+}\left(\mu_x^{+}-\xi^{+}\right) \\ M_{xz}^{+}-\mu_z^{+}\xi^{+} \end{array}\right)$$

$$\to p^{+}\beta\left(\begin{array}{l} \mu_x^{+}-\xi^{+} \\ \xi^{+}\left(\mu_x^{+}-\xi^{+}\right) \\ M_{xz}^{+}-\mu_z^{+}\xi^{+} \end{array}\right) \equiv \Omega B^{+} \tag{25}$$

where '$^{+}$ 'indicates expectation over the 'high censoring region'; namely $\mu_x^{+} = E\left(x \mid x > \xi^{+}\right)$, $\mu_z^{+} = E\left(z \mid x > \xi^{+}\right)$, $M_{xx}^{+} = E\left(x^2 \mid x > \xi^{+}\right)$, $M_z^{+} = E\left(zz^{'} \mid x > \xi^{+}\right)$; and the weight Ω_η^{+} shrinks to 0 because the variance of $x^{(\eta)}$ shrinks to 0 over that region. By the symmetries in the formulae, we have

$$p^{-}\Omega_\eta^{-}\left(\begin{array}{l} \beta\xi^{-}\left(1-\frac{1}{\eta}\right) \\ \beta\left(\frac{1}{\eta}-1\right) \\ 0 \end{array}\right) = p^{-}\beta\left(1-\eta\right)\left(\begin{array}{l} \mu_x^{-}-\xi^{-} \\ \eta\left(M_{xx}^{-}-\mu_x^{-}\xi^{-}\right)+\left(1-\eta\right)\xi^{-}\left(\mu_x^{-}-\xi^{-}\right) \\ M_{xz}^{-}-\mu_z^{-}\xi^{-} \end{array}\right)$$

$$\to p^{-}\beta\left(\begin{array}{l} \mu_x^{-}-\xi^{-} \\ \xi^{-}\left(\mu_x^{-}-\xi^{-}\right) \\ M_{xz}^{-}-\mu_z^{-}\xi^{-} \end{array}\right) \equiv \Omega B^{-} \tag{26}$$

So, as above, the overall bias in $\left(\hat{a}, \hat{b}, \hat{c}^{'}\right)$ is the sum of biases generated by the low and high censoring region,

$$\text{plim}\left(\begin{array}{c} \hat{a} \\ \hat{b} \\ \hat{c} \end{array}\right) - \left(\begin{array}{c} \alpha \\ \beta \\ \gamma \end{array}\right) = \beta\cdot\Omega^{-1}\left(\begin{array}{l} p^{-}\left(\mu_x^{-}-\xi^{-}\right)+p^{+}\left(\mu_x^{+}-\xi^{+}\right) \\ p^{-}\xi^{-}\left(\mu_x^{-}-\xi^{-}\right)+p^{+}\xi^{+}\left(\mu_x^{+}-\xi^{+}\right) \\ p^{-}\left(M_{xz}^{-}-\mu_z^{-}\xi^{-}\right)+p^{+}\left(M_{xz}^{+}-\mu_z^{+}\xi^{+}\right) \end{array}\right) = B^{-}+B^{+}. \tag{27}$$

where B^{-} and B^{+} are defined in (25) and (26).

We can easily interpret these formulae in the 'no correlation' case. Suppose $\mu_z = 0$ and that z is *mean independent* of x. Then $\mu_z^{-} = \mu_z^{+} = M_{xz}^{-} = M_{xz}^{+} = 0$, and Ω is block diagonal (partitioned according to $(1, x^{cen}),\ z$). Consequently, there is 0 bias in the z coefficient $\hat{c}$, and $\left(\hat{a}, \hat{b}\right)$ behave as though z were not in the equation for estimation. If further, $E\left(x^{cen}\right) = 0$ (which implies $\xi^{-} < 0 < \xi^{+}$), then we have

$$\text{plim } \hat{a} - \alpha = \beta\cdot\left(p^{-}\left(\mu_x^{-}-\xi^{-}\right)+p^{+}\left(\mu_x^{+}-\xi^{+}\right)\right), \tag{28}$$

and

$$\text{plim } \hat{b} - \beta = \beta \cdot \frac{\left(p^{-}\xi^{-}\left(\mu_x^{-} - \xi^{-}\right) + p^{+}\xi^{+}\left(\mu_x^{+} - \xi^{+}\right)\right)}{Var\left(x^{cen}\right)}. \tag{29}$$

The bias in the intercept $\hat{a}$ consists of a negative 'low censoring' term and and a positive 'high censoring' term. For $\hat{b}$ there is positive expansion bias term from both low and high censoring (as consistent with Proposition 1). In a 'symmetric censoring' case with $p^{-} = p^{+}$, $-\xi^{-} = \xi^{+}$, and $-\mu_x^{-} = \mu_x^{+}$, the intercept bias is exactly offset, plim $\hat{a} - \alpha = 0$. The slope $\hat{b}$ contains equal expansion bias terms from low and high censoring; plim $\hat{b} - \beta = \beta \cdot 2 \cdot p^{+}\ \xi^{+}\left(\mu_x^{+} - \xi^{+}\right)\big/ Var\left(x^{cen}\right)$.

For the case with correlation, the exact bias formulae are too complex to admit clear interpretation. However, we can learn from expanding the bias in terms of the censoring probabilities, as follows. Here we examine the 'high censoring bias' B^{+}, and an analogous development can be carried out for B^{-}. Suppose there is a single variable z with $\mu_z = 0$, and we now take $\mu_x = 0$. Now, through a very tedious calculation, the bias B^{+} can be written as terms linear in p^{+} and terms of higher polynomial order in p^{+}. Dropping the higher order terms, we have

$$B^{+} \cong \beta \cdot \frac{p^{+}}{M_{zz}M_{xx} - \left(M_{xz}\right)^2} \begin{pmatrix} \left(M_{zz}M_{xx} - \left(M_{xz}\right)^2\right)\left(\mu_x^{+} - \xi^{+}\right) \\ M_{zz}\xi^{+}\left(\mu_x^{+} - \xi^{+}\right) - M_{xz}\left(M_{xz}^{+} - \xi^{+}\mu_z^{+}\right) \\ M_{xx}\left(M_{xz}^{+} - \xi^{+}\mu_z^{+}\right) - M_{xz}\xi^{+}\left(\mu_x^{+} - \xi^{+}\right) \end{pmatrix} \tag{30}$$

where $M_{xx} = E\left(x^2\right), M_{xz} = E\left(xz\right)$ and $M_{zz} = E\left(z^2\right)$ are unconditional (and uncensored) second moments. For p^{+} small, this should be a very good approximation (since $\left(p^{+}\right)^k$, $k \geq 2$ is much smaller than p^{+}). Let $\sigma_x = \sqrt{M_{xx}}$, $\sigma_z = \sqrt{M_{zz}}$ denote the unconditional standard deviations of x and z and let ρ_{xz} denote their correlation coefficient.

The approximate bias term for the intercept $\hat{a}$ is $\beta p^{+}\left(\mu_x^{+} - \xi^{+}\right)$,which is the same form as in (28). For $\hat{b}$ and $\hat{c}$, the biases are:

$$\text{plim } \hat{b} - \beta \ \propto \ \beta p^{+}\left\{\xi^{+}\left(\mu_x^{+} - \xi^{+}\right) - \rho_{xz}\frac{\sigma_x}{\sigma_z}\left[M_{xz}^{+} - \xi^{+}\mu_z^{+}\right]\right\} \tag{31}$$

and

$$\text{plim } \hat{c} - \gamma \ \propto \ \beta p^{+}\left\{\frac{\sigma_x}{\sigma_z}\left[M_{xz}^{+} - \xi^{+}\mu_z^{+}\right] - \rho_{xz}\xi^{+}\left(\mu_x^{+} - \xi^{+}\right)\right\} \tag{32}$$

where the positive proportions are from (30) and are therefore approximate.

To interpret these terms, suppose that x and z are positively correlated overall ($\rho_{xz} > 0$), positively correlated within the 'high censored' region (within covariance $C_{xz}^{+} > 0$), and that $\xi^{+} > 0$. Given this, we have $\xi^{+}\left(\mu_x^{+} - \xi^{+}\right) > 0$ and $M_{xz}^{+} - \xi^{+}\mu_z^{+} = C_{xz}^{+} + \left(\mu_x^{+} - \xi^{+}\right)\mu_z^{+} > 0$. So, the approximate biases are each differences of positive terms. They can be interpreted as follows. Since x and z are positively correlated, in OLS estimation z will proxy some of the role of the censored values of x. Therefore, for $\hat{b}$, we see that (31) consists of the positive expansion bias term of (29), less a positive term due to z's

	Bias of:	Correlation				
Truncation		-90%	-50%	0%	50%	90%
1%	$\widehat{a}$	0.1%	0.1%	0.1%	0.1%	0.1%
	$\widehat{b}$	-0.4%	0.7%	0.8%	0.7%	-0.4%
	$\widehat{c}$	-0.4%	-0.1%	0.0%	0.1%	0.4%
20%	$\widehat{a}$	3.3%	4.2%	4.3%	4.2%	3.3%
	$\widehat{b}$	-11.5%	12.3%	14.5%	12.7%	-10.3%
	$\widehat{c}$	-8.0%	-2.6%	-0.2%	2.3%	7.7%
40%	$\widehat{a}$	7.2%	11.9%	12.5%	12.0%	7.4%
	$\widehat{b}$	-23.2%	26.1%	32.2%	27.2%	-21.4%
	$\widehat{c}$	-14.9%	-6.5%	-0.5%	5.6%	14.5%
60%	$\widehat{a}$	14.2%	29.7%	32.3%	30.1%	14.3%
	$\widehat{b}$	-27.2%	52.2%	65.9%	54.2%	-25.6%
	$\widehat{c}$	-20.2%	-11.7%	-0.9%	10.3%	19.9%

Table 1: Coefficient Biases: Two Regressors with One Censored.

role as a proxy for the censored values of x. Similar terms arise for $\hat{c}$ in (32) in the reverse positions; a positive bias arises because z proxies the censored values of x, less a term arising from the expansion bias of the coefficient for the uncensored x values.

It is interesting to note that the net biases can go either way depending on the correlation between x and z. For a small correlation, the expansion bias in $\hat{b}$ will be evident, and a negative bias in $\hat{c}$ arises in response to that. For a large correlation, the positive expansion bias in $\hat{b}$ can be wholly reversed, and with a positive bias arising for $\hat{c}$. In that case, it appears that z is doing a better job of proxying for x than the censored x^{cen} is.

To clarify the intuition of the previous derivation we performed a Monte Carlo exercise. We assumed that x, z and ε are normally distributed,[3] and computed OLS estimates for different degrees of (top-coding) censoring and different correlations between x and z. Summary results are given in Table 1[4] and the full results are displayed in Figures 3 to 5.

As can be seen in the Figure 3 and Table 1, the bias on the intercept $\hat{a}$ is always positive (for top-coding) and it is sizeable. For example, if 20 percent of the observations are censored there is a bias of roughly 4 percent. For the coefficient $\hat{b}$ of the censored regressor x^{cen}, the results are in line with the intuition we developed before. Namely, at low and moderate levels of correlation between x and z, there is a clear expansion bias, but as the correlation becomes high, the bias turns negative. The highest bias occurs at zero correlation, and it decreases with as absolute value of the correlation increases. For the coefficient $\hat{c}$ of the other regressor z, bias and correlation have a monotone relationship. Positive bias arises for positive correlation and negative bias for negative correlations, with larger levels of censoring implying larger biases in absolute terms. This also is in line with the intuition above. In any case, we feel these results suggest that the biases introduced by the censoring of a regressor can be large and

[3] We set $\alpha = 1$, $\beta = 0.5$, and $\gamma = 0.75$, took the variances of x and z to be the same and equal to half the value of the variance of ε.

[4] In line with (15), $\hat{a}$ is the estimated intercept, $\hat{b}$ is the estimated coefficient on the censored variable x and $\hat{c}$ is the estimated coefficient of the other variable z.

economically significant.

2.3. Expansion Bias and Errors-in-Variables

We close this section with a curiosity about censoring regressors that are measured with error. Errors-in-variables cause attenuation bias whereas censoring regressors causes expansion bias, the opposite. So it is natural to consider correcting for errors-in-variables by censoring the variable measured with error.

Consider the bivariate model

$$y_i = \alpha + \beta w_i + \varepsilon_i, \quad i = 1, ..., n \tag{33}$$

where

$$x_i = w_i + \nu_i \tag{34}$$

and ε_i, ν_i and w_i are mutually uncorrelated. If we estimate

$$y_i = a + b^* x_i + \eta_i \quad i = 1, ..., n, \tag{35}$$

then we have the standard result that

$$\text{plim } \hat{b}^* = \beta \left(1 - \lambda\right), \qquad \lambda = \frac{\text{Var}\left(\nu\right)}{\text{Var}\left(x\right)}. \tag{36}$$

However, if we further censor x as above, and estimate

$$y_i = a + b x_i^{cen} + u_i \quad i = 1, ..., n, \tag{37}$$

then we have

$$\text{plim } \hat{b} = \beta \left(1 - \lambda\right) \left(1 + \Lambda\right) \tag{38}$$

Provided $\left(1 - \lambda\right) \left(1 - \Lambda\right) = 1$, or

$$\Lambda = \frac{\lambda}{1 - \lambda} \tag{39}$$

then the errors-in-variables bias will have been corrected by the censoring.

It is not clear how practically useful this fact is, since to know how much to censor would require knowing how much error variance there is (and then it could be corrected directly). But we find it an interesting connection, and it suggests some other questions. Does correcting a certain attenuation bias require a great deal of censoring or very little? Is one of these bias problems of a different order of magnitude than the other?

Figure 6 sheds light on these questions for a normal regressor. It shows the level of censoring required to correct each level of attenuation bias. Specifically, it is assumed that there is two-side symmetric censoring (p low and p high censoring), and it plots the total censoring probability $2p$ against the relative error variance λ underlying the zero bias condition (39). It shows that the probability of censoring $2p$ needs to be a larger than the error variance λ but not a great deal larger. For instance, slightly less than 40% total censoring would be needed to correct the bias implied by 30% relative error variance. In any case, the impact of censoring a certain percentage of the data can be thought of a slightly smaller than the impact of an error variance of the same level, in the opposite direction.

In any case, it is clear that errors-in-variables bias can be counteracted by the censoring the regressor. This does raise questions about survey design. Suppose that a variable is measured with error and that error is likely higher for the top range of the variable - for example, hours spent in traffic, gallons of beer consumed, etc. In this circumstance it is possible some degree of top-coding would make sense as part of survey design. In any case, there may be some practical impact to the interplay of censoring with errors-in-variables.

3. Correcting Expansion Bias

There are at least three broad approaches to correcting expansion bias that are useful to discuss. The first approach is to focus on the distribution of the regressors, assuming a specific form that allows the bias terms to be identified. The bias terms could then be estimated (or simulated), and the OLS coefficients adjusted by the bias estimates. This approach has the usual proviso about distributional assumptions in econometric work – namely there is usually not a great deal of guidance as to what structure should be assumed, especially in multivariate settings. However, it is worthwhile to note that since the regressors are observed, assumptions on their distributional structure are (potentially) testable. For instance, to validate the assumption that the regressors are multivariate normal, one could apply goodness-of-fit tests to check whether the censored regressor is distributed as a censored univariate normal.

The second approach is to eliminate the problematic data. That is, estimation can be done on the truncated sample with $\xi^- < x_i^{cen} < \xi^+$, where all observations with censored data have been deleted from the sample. As we noted in Section 2.2, the OLS estimators from the truncated sample will be unbiased and consistent estimators of the true coefficients. Because the censored data is informative, one would expect an efficiency loss from this approach, but it is so simple to do that it is probably good empirical practice to always compute the estimates from the truncated sample for comparison. This approach also illustrates an important difference between the censoring of a regressor and the more familiar problem of censoring a dependent variable. Dropping all censored observations of a dependent variable does not avoid the bias in OLS coefficients.

The third approach is to augment the specification of the regression equation to remove the source of the bias. Consider the case of many regressors. The true model is

$$y_i = \alpha + \beta x_i^{cen} + \gamma' z_i + \beta x_i^- + \beta x_i^+ + \varepsilon_i \quad i = 1, ..., n \tag{40}$$

where

$$x_i^- = \left(x_i - \xi^-\right) \cdot 1\left[x_i < \xi^-\right] \quad \text{and} \quad x_i^+ = \left(x_i - \xi^+\right) \cdot 1\left[\xi^+ < x_i\right]. \tag{41}$$

Expansion bias arises from omitting x^- and x^+. For correction, we can add to the equation the parts of x^- and x^+ that are correlated with the included regressors; as in

$$\begin{aligned} y_i &= \alpha + \beta x_i^{cen} + \gamma' z_i + \beta E\left(x^- | x^{cen} = x_i^{cen}, z = z_i\right) \\ &\qquad + \beta E\left(x^+ | x^{cen} = x_i^{cen}, z = z_i\right) + e_i \end{aligned} \tag{42}$$

where $e_i \equiv \varepsilon_i - \beta E\left(x^- | x^{cen} = x_i^{cen}, z = z_i\right) - \beta E\left(x^+ | x^{cen} = x_i^{cen}, z = z_i\right)$ is uncorrelated with x^{cen} and z.

Let's examine these correction terms in a bit more detail. First, we have

$$\begin{aligned} E\left(x^- | x^{cen} = x_i^{cen}, z = z_i\right) &= 0 \quad &&\text{if } x_i^{cen} \neq \xi^- \\ &\equiv f^-(z_i) &&\text{if } x_i^{cen} = \xi^- \end{aligned} \tag{43}$$

and

$$\begin{aligned} E\left(x^+ | x^{cen} = x_i^{cen}, z = z_i\right) &= 0 \quad &&\text{if } x_i^{cen} \neq \xi^+ \\ &\equiv f^+(z_i) &&\text{if } x_i^{cen} = \xi^+ \end{aligned} \tag{44}$$

So, the terms are nonzero only for the censored observations, but for those points, they are functions of z_i that are determined by the joint distribution of x and z. As before, if we assume a specific form for the joint distribution of x and z, we could derive and/or estimate the functions $f^-(\cdot)$ and $f^+(\cdot)$.

Treating those functions as unknown, we have a fully semiparametric model

$$y_i = \alpha + \beta x_i^{cen} + \gamma' z_i + \beta f^-(z_i)\, d_i^- + \beta f^+(z_i)\, d_i^+ + e_i, \;\; i = 1, .., n \tag{45}$$

where $d_i^- \equiv 1\left[x_i^{cen} = \xi^-\right]$ and $d_i^+ \equiv 1\left[x_i^{cen} = \xi^+\right]$ indicate the censored points. As long as $p^- > 0$, $p^+ > 0$, $1 - p^- - p^+ > 0$, the model is identified. In heuristic terms, we can identify α, β and γ with the truncated data, the function $f^-(z_i)$ from $y_i - \alpha - \beta\xi^- - \gamma' z_i$ for the low censored points, and the function $f^+(z_i)$ from $y_i - \alpha - \beta\xi^+ - \gamma' z_i$ for the high censored points. The model (45) is is very similar to partially linear models as originally studied by Robinson(1988), and estimation can be approached by series expansion, within-differencing or a variety of other techniques. We do not go into the details of these methods here, but we will analyze them in subsequent research.[5]

It is useful to consider whether there are 'quick-fix' methods to recommend for empirical practice. Since $f^-(z_i)$ and $f^+(z_i)$ are unknown functions, an initial idea is to approximate them by linear functions. This amounts to allowing the regression parameters to vary between the truncated data, the low censored data and the high censored data. This is accomplished by adding d_i^-, $d_i^- z_i$, d_i^+, $d_i^+ z_i$ to the list of regressors. This should help remove some of the bias in the coefficients of 1, x_i^{cen} and z_i.

Are linear approximations likely to be good in this problem? The functions $f^-(z_i)$ and $f^+(z_i)$ are unknown and so there is no real substitute for a fully semiparametric approach to estimating them. However, it is useful to return once again to the case of a multivariate normal distribution and see how

[5] Yatchew (2003) contains coverage of partially linear models and many references. Schmalensee and Stoker (1999) analyzes U.S. gasoline demand using partially linear models, with bivariate nonparametric structure.

'linear' $f^{-}(z_i)$ and $f^{+}(z_i)$ are. Taking z to be a single variable for simplicity, under the assumption that (x,z) are joint normal, $f^{-}(z_i)$ and $f^{+}(z_i)$ are easily derived. We present the formulae in Appendix A.

Figure 7 plots these functions assuming that x and z are standard normal, with a correlation of .5. The function $f^{-}(z_i)$ is based on 10% low censoring, and $f^{+}(z_i)$ is based on 20% high censoring.[6] The functions are not wildly nonlinear over the substantive range of z density, but they are not linear either. So here some nonlinear terms would be necessary for a good approximation. In general, a semiparametric method would allow more arbitrary shapes to be captured, and one could perform tests for the function shapes implied by joint normality.

4. Situations where Censored Regressors are Endemic

This paper is our first writing on censored regressors, and we are in the process of developing applications where censored regressors are prevalent and where empirical findings may have been greatly impacted by their presence. We now describe some of the areas we plan to study as part of future research, to give more concrete illustration to the issues we have been discussing.

4.1. The Effects of Wealth on Consumption

In recent years, many developed countries have witnessed tremendous expansions in consumption expenditures at the same time as substantial increases in household wealth levels.[7] This has fueled great interest in the measurement of the effects of wealth on consumption decisions.

Consider first the issues associated with studying how consumption reacts to changes in financial wealth, say with a stylized model:

$$C_i = \alpha + \beta \cdot INC_i + \gamma \cdot FW_i + \varepsilon_i \quad i = 1, ..., n \tag{46}$$

where C_i is consumption, INC_i is income and FW_i is a measure of financial wealth, such as the value of stock and bond holdings. Now, it is very common that income is top-coded, producing the kind of censored regressor bias that we have discussed. But also, measures of financial wealth are often bottom-coded reflections of actual wealth, because of problems in measuring or capturing debt levels. That is, using only positive components of wealth, such as actual stock and bond holdings, generates a censored regressor problem of the same style as that generated by top-coded income. Since income

[6] We choose different values to illustrate how the shapes of the functions can vary. It is easy to verify that the same amount of low and high censoring (here $-\xi^{-} = \xi^{+}$) gives the functions as negative reflections of one another, $-f^{-}(-z) = f^{+}(z)$.

[7] During the 1990's there were multiyear expansions in consumption in the US and the UK (among others). During the same time, the total wealth of Americans grew more than 15 trillion dollars, with a 262% increase in corporate equity and a 14% increase in housing and other tangible assets (see Poterba (2000) for an excellent survey). Housing prices increased in both countries as well.

and measured wealth tend to be positively correlated, wealth (and income) effects will tend to be overestimated.

In fact, published estimates of the elasticity of consumption with regard to financial wealth seem large. With aggregate data, estimates in the range of 4% but up to 10% can be found, varying with the type of asset included and the time period under consideration.[8] With individual data, estimates tend to be larger,[9] such as 8%. In any case, we plan to investigate whether the censored character of income and financial wealth can help account for the magnitude of these estimates.

These issues are greatly exacerbated when one adds consideration of housing wealth, as in.

$$C_i = \alpha + \beta \cdot INC_i + \gamma \cdot FW_i + \delta \cdot HW_i + \varepsilon_i \quad i = 1, ..., n \tag{47}$$

where HW_i is housing wealth. Even when housing wealth is accurately measured, the censoring of income and/or financial wealth will cause the effect of housing wealth to be overestimated. In fact, estimates of the marginal propensity to consume out of housing wealth elasticity are in the range of 18%, which again seems quite large.[10] It would be interesting to understand how much of this effect could be due to censored regressor problems.

It is worth mentioning that estimates of wealth effects are of substantial interest to economic policy. A key issue of monetary policy is how much aggregate demand is affected by changes in interest rates. In addition to the direct effects on consumption, it is obvious that interest rates will affect housing wealth as well as financial wealth. A substantial impact of wealth on consumption, either through enhanced borrowing or cashing out of capital gains, will be a big part of whether interest rates are effective or not. In any case, understanding these connections is crucial for the design of effective monetary policy.[11]

4.2. Minimum Wage and Peer Effects

Here we consider how peer effects in productivity are studied, as in Borjas (1994), Card et. al. (1998), and others. In this work, it is typical to use the average wage of the peer group as a proxy for the ability level of the peer group. This introduces error-in-variables structure in the standard way. However, if minimum wage rules are binding for some groups, then there is a censored regressor problem. As

[8] Laurence Meyer and Associates (1994) find an elasticity of 4.2 percent, Brayton and Tinsley (1996) find 3 percent, Ludvigson and Steindel (1999) estimate an overall elasticity of 4 percent (as well as some estimates as high as 10 percent).

[9] Using the PSID, Parker (1999) concludes that the elasticity of household expenditures to household net worth is approximately 8 percent (although the PSID has few households with large financial wealth holdings). Juster, Lupton, Smith and Stafford (1999) and Starr-McCluer (1999) find slightly smaller coefficient, but still larger than the ones obtained from aggregate data.

[10] See Aoki, et. al. (2002a, 2002b) and Attanasio, et. al. (1994), among others. Somewhat smaller estimates are given in Engelhardt (1996) and Skinner (1996).

[11] See Muellbauer and Murphy (1990), King (1990), Pagano (1990) Attanasio and Weber (1994) and Attanasio et. al. (2003), for various arguments on the connection between consumption and housing prices. In terms of whether assets prices should be targeted as part of monetary policy, see Bernanke and Gertler (1999,2001), Cecchetti et. al. (2000) and Rigobon and Sack (2003).

discussed in Section 2.3, the 'censored regressor' bias will be in the opposite direction to the "errors-in-variables' bias.

For concreteness, consider the model:

$$Y_i = \alpha + \beta H_i + \varepsilon_i, \qquad i = 1, ..., n \tag{48}$$

where Y_i is the variable of interest[12] and H_i stands for ability of the individual of the peer group. Suppose W_i is the true wage of the peer group that is correlated with ability, as in

$$W_i = \gamma H_i + \eta_i \tag{49}$$

ε_i and η_i are standard innovations, assumed to be uncorrelated.

Now, if true wage W_i is observed and used as a proxy for ability H_i, as in

$$Y_i = \alpha + \pi^* W_i + \nu_i^*, \tag{50}$$

then there is a pure errors-in-variables problem. Proper instruments could be used to obtain consistent estimates. However, suppose that there is a minimum wage constraint, so that observed wage is in fact

$$W_i^{cen} = \underline{W} \cdot 1\left[W_i < \underline{W}\right] + W_i \cdot 1\left[W_i \geq \underline{W}\right] \tag{51}$$

In this situation, the regression

$$Y_i = \alpha + \pi W_i^{cen} + \nu_i, \tag{52}$$

has errors-in-variables and a censored regressor. As discussed in Section 2.3, the attenuation bias from errors-in-variables can be counteracted by the expansion bias from the censored regressor.

5. Conclusion

The fact that censored regressors generate expansion bias was a big surprise to both authors. We noticed the phenomena in some simulations, and were able to understand the source pretty easily. In fact, it is a completely straightforward point, as Figure 1 can be explained to students with only rudimentary knowledge of econometric methods. Nevertheless, we don't feel that it is a minor problem for practical applications. Quite the contrary, we feel that problems of censored regressors are likely as prevalent or more prevalent than problems of censored dependent variables in typical econometric applications.

We feel that we were able to make some progress in understanding the structure of expansion bias. Sure, the 'many regressor' formulae are complicated – in what problem are they easy? – but we were

[12] Examples of Y_i include own wages, hours worked, decisions of financing, decisions of participation in financial markets or labor markets, etc., or any decision that is affected by peer considerations. As with out discussion of consumption, applications would include control variables, which we abstract from for this discussion.

able to see how the censoring issues transmit across regressors. We were able to get a sense of the severity of the biases using formulae and simulations from normal distributions, and a side comparison to the biases that arise from errors in variables.

But the real value of this material will be decided by how well expansion bias can be isolated in actual applications. Here, we are quite hopeful. There is a correction method which is easy to implement, understand and interpret – drop the censored points and see what happens. And, there is a straightforward semiparametric approach for using all the data, including the censored points. While we are not yet at the point of having a well tried "set of instructions" for implementing the semiparametric approach, we expect to generate one as a byproduct of doing applications ourselves.

References

[1] Aoki, K., J. Proudman and G. Vlieghe (2002a) “House prices, consumption, and monetary policy: a financial accelerator approach” *Bank of England Quarterly Bulletin.*

[2] Aoki, K., J. Proudman and G.Vlieghe (2002b) “Houses as collateral: has the link between house prices and consumption in the UK changed?’, *Economic Policy* Review Vol. 8 (1), Federal Reserve Bank of New York,

[3] Attanasio, O., L. Blow, R. Hamilton, and A. Leicester (2003) “Consumption, House Prices, and Expectations” Institute for Fiscal Studies, Mimeo.

[4] Attanasio, O., and G. Weber (1994) “The UK Consumption Boom of the late 1980s: aggregate implications of microeconomic evidence” in *The Economic Journal*, Vol. 104, Issue 427, November, pp 1269-1302.

[5] Bernanke, B., and M. Gertler, “Monetary Policy and Asset Price Volatility,” *Federal Reserve Bank of Kansas City Economic Review*, LXXXIV (1999), 17–51.

[6] Bernanke, B., and M. Gertler, “Should Central Banks Respond toMovements in Asset Prices?” *American Economic Review Papers and Proceedings*, XCI (2001), 253–257.

[7] Borjas, G. (1994) “Long-Run Convergence of Ethnic Skill Differentials” NBER 4641

[8] Card, D., J. DiNardo, and E. Estes (1998) “The More Things Change: Immigrants and the Children of Immigrants in the 1940s, the 1970s, and the 1990s” NBER 6519

[9] Cecchetti, S. G., H. Genberg, J. Lipsky, and S. Wadhwani, *Asset Prices and Central Bank Policy* (London: International Center for Monetary and Banking Studies, 2000).

[10] Green, W.H. (2003). *Econometric Analysis*, 5th ed. New Jersey: Prentice Hall.

[11] King, M. (1990) “Discussion” in *Economic Policy*, Vol. 11, pp 383-387.

[12] Muellbauer, J. and A. Murphy (1990) “Is the UK balance of payments sustainable?” in *Economic Policy*, Vol. 11, pp 345-383.

[13] Pagano C.(1990) “Discussion” in *Economic Policy*, Vol. 11, pp 387-390.

[14] Poterba, J. M. (2000) “Stock Market Wealth and Consumption," *Journal of Economic Perspectives,* Volume 14, Number 2, Spring 2000, pp. 99–118.

[15] Robinson, P.M. (1988). "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931-954.

[16] Engelhardt, G. (1996). “House Prices and Home Owner Saving Behavior," *Regional Science and Urban Economics,* 26, pp. 313–36.

[17] Skinner, J. (1996). “Is Housing Wealth a Sideshow?’ in *Advances in the Economics of Aging.* D. Wise, ed. Chicago: University of Chicago Press, pp. 241–68.

[18] Brayton, F. and P. Tinsley. (1996). “A Guide to the FRB/US: A Macroeconomic Model of the United States.’ Federal Reserve Board of Governors, Washington DC, Working Paper 1996-42.

[19] Lawrence H. Meyer and Associates. (1994). *The WUMM Model Book.* St. Louis: L. H. Meyer and Associates.

[20] Ludvigson, S. and C. Steindel. (1999). “How Important is the Stock Market Effect on Consumption?’ *Federal Reserve Bank of New York Economic Policy Review.* July, 5:2, pp. 29–52.

[21] Juster, F. T., Joseph L., J. P. Smith, and F. Stafford. (1999). “Savings and Wealth: Then and Now.’ Mimeo, University of Michigan, Institute for Survey Research.

[22] Starr-McCluer, M. (1999). “Stock Market Wealth and Consumer Spending.’ Mimeo, Federal Reserve Board of Governors.

[23] Parker, J. (1999). “Spendthrift in America? On Two Decades of Decline in the U.S. Saving Rate?’ in *NBER Macroeconomics Annual 1999.* B. Bernanke and J. Rotemberg, eds. Cambridge: MIT Press.

[24] Schmalensee, R. and T. Stoker (1999), "Household Gasoline Demand in the United States," *Econometrica*, 67, 645-662.

[25] Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician*, Cambridge: Cambridge University Press.

A. Appendix: Formulae for Normal Regressors

We first present the formulae for expansion bias when the uncensored regressor is normally distributed, namely $x \sim \mathcal{N}\left(\mu_x, \sigma_x^2\right)$. These formulae follow from standard expressions for the mean and variance of a truncated normal distribution (c.f. Green (2003), chapter 22, among many others), and we denote the standard normal density as $\phi(\cdot)$ and the standard normal c.d.f. as $\Phi(\cdot)$. It is clear that we can parameterize the bias formulae equivalently in terms of the censoring points ξ^-, ξ^+ or the probabilities $p^- = \Pr\left\{x < \xi^-\right\}$, $p^+ = \Pr\left\{x > \xi^+\right\}$; as

$$p^- = \Phi\left(\frac{\xi^- - \mu_x}{\sigma_x}\right);\ 1 - p^+ = \Phi\left(\frac{\xi^+ - \mu_x}{\sigma_x}\right) \tag{53}$$

are fully invertible to

$$\xi^- = \sigma_x \Phi^{-1}\left(p^-\right) + \mu_x; \quad \xi^+ = \sigma_x \Phi^{-1}\left(1 - p^+\right) + \mu_x \tag{54}$$

We choose the p^-, p^+ parameterization to facilitate some points in the text.

Recall that

$$x_i^{cen} = x_i \cdot 1\left[\xi^- \leq x_i \leq \xi^+\right] + \xi^- \cdot 1\left[x_i < \xi^-\right] + \xi^+ \cdot 1\left[\xi^+ < x_i\right], \tag{55}$$

$$x_i^- = \left(x_i - \xi^-\right) \cdot 1\left[x_i < \xi^-\right] \quad \text{and} \quad x_i^+ = \left(x_i - \xi^+\right) \cdot 1\left[\xi^+ < x_i\right]. \tag{56}$$

We some initial results

$$E\left(x^-\right) = -\sigma_x \phi\left[\Phi^{-1}\left(p^-\right)\right] - p^- \sigma_x \Phi^{-1}\left(p^-\right) \tag{57}$$

$$E\left(x^+\right) = \sigma_x \phi\left[\Phi^{-1}\left(1 - p^+\right)\right] - p^+ \sigma_x \Phi^{-1}\left(1 - p^+\right) \tag{58}$$

$$\begin{aligned} E\left(x^-\right)^2 \quad = \quad & \left\{\left(1 - \frac{\phi\left[\Phi^{-1}\left(p^-\right)\right]^2}{\left(p^-\right)^2} - \frac{\phi\left[\Phi^{-1}\left(p^-\right)\right]\Phi^{-1}\left(p^-\right)}{p^-}\right)\right. \\ & \left.\left(-\frac{\phi\left[\Phi^{-1}\left(p^-\right)\right]}{p^-} - \Phi^{-1}\left(p^-\right)\right)^2\right\} \cdot \sigma_x^2 p^- \end{aligned} \tag{59}$$

$$\begin{aligned} E\left(x^+\right)^2 \quad = \quad & \left\{\left(1 - \frac{\phi\left[\Phi^{-1}\left(1 - p^+\right)\right]^2}{\left(p^+\right)^2} + \frac{\phi\left[\Phi^{-1}\left(1 - p^+\right)\right]\Phi^{-1}\left(1 - p^+\right)}{p^+}\right)\right. \\ & \left.\left(\frac{\phi\left[\Phi^{-1}\left(1 - p^+\right)\right]}{p^+} - \Phi^{-1}\left(1 - p^+\right)\right)^2\right\} \cdot \sigma_x^2 p^+ \end{aligned} \tag{60}$$

$$E\left(x^{cen}x^{-}\right)=\left(\sigma_x\Phi^{-1}\left(p^{-}\right)+\mu_x\right)\left\{-\sigma_x\phi\left[\Phi^{-1}\left(p^{-}\right)\right]-p^{-}\sigma_x\Phi^{-1}\left(p^{-}\right)\right\} \tag{61}$$

$$E\left(x^{cen}x^{+}\right)=\left(\sigma_x\Phi^{-1}\left(1-p^{+}\right)+\mu_x\right)\left\{\sigma_x\phi\left[\Phi^{-1}\left(1-p^{+}\right)\right]-p^{+}\sigma_x\Phi^{-1}\left(1-p^{+}\right)\right\} \tag{62}$$

The bias is computed by substituting (57)-(62) in the following:

$$E\left(x^{cen}\right)=\mu-E\left(x^{-}\right)-E\left(x^{+}\right) \tag{63}$$

$$Cov\left(x^{-},x^{cen}\right)=E\left(x^{cen}x^{-}\right)-E\left(x^{cen}\right)E\left(x^{-}\right) \tag{64}$$

$$Cov\left(x^{+},x^{cen}\right)=E\left(x^{cen}x^{+}\right)-E\left(x^{cen}\right)E\left(x^{+}\right) \tag{65}$$

$$\begin{aligned} Var\left(x^{cen}\right) &= \mu_x^2+\sigma_x^2-E\left(x^{-}\right)^2-E\left(x^{+}\right)^2 \\ &\quad -2E\left(x^{cen}x^{-}\right)-2E\left(x^{cen}x^{+}\right)-\left[E\left(x^{cen}\right)\right]^2 \end{aligned} \tag{66}$$

$$\Lambda^{-}=\frac{Cov\left(x^{-},x^{cen}\right)}{Var\left(x^{cen}\right)} \text{ and } \Lambda^{+}=\frac{Cov\left(x^{+},x^{cen}\right)}{Var\left(x^{cen}\right)}. \tag{67}$$

with the expansion bias given as $\Lambda^{-}+\Lambda^{+}$.

For the regression correction terms of Section 3, assume that x and z are multivariate normal with means μ_x, μ_z, variances σ_x^2, σ_z^2 and covariance σ_{xz}. Define $\delta=\sigma_{xz}/\sigma_z^2$ and $\omega^2=\sigma_x^2-\delta^2\sigma_z^2$. Then it is straightforward to show that

$$\begin{aligned} f^{-}(z) &= \left(\mu_x-\xi^{-}+\delta\left(z-\mu_z\right)\right)\left(\Phi\left(\frac{\xi^{-}-\mu_x-\delta\left(z-\mu_z\right)}{\omega}\right)\right) \\ &\quad -\omega\phi\left(\frac{\xi^{-}-\mu_x-\delta\left(z-\mu_z\right)}{\omega}\right) \end{aligned} \tag{68}$$

and

$$\begin{aligned} f^{+}(z) &= \left(\mu_x-\xi^{+}+\delta\left(z-\mu_z\right)\right)\left(1-\Phi\left(\frac{\xi^{+}-\mu_x-\delta\left(z-\mu_z\right)}{\omega}\right)\right) \\ &\quad +\omega\phi\left(\frac{\xi^{+}-\mu_x-\delta\left(z-\mu_z\right)}{\omega}\right) \end{aligned} \tag{69}$$

where $\phi(\cdot)$ is the normal density and $\Phi(\cdot)$ is the normal c.d.f.

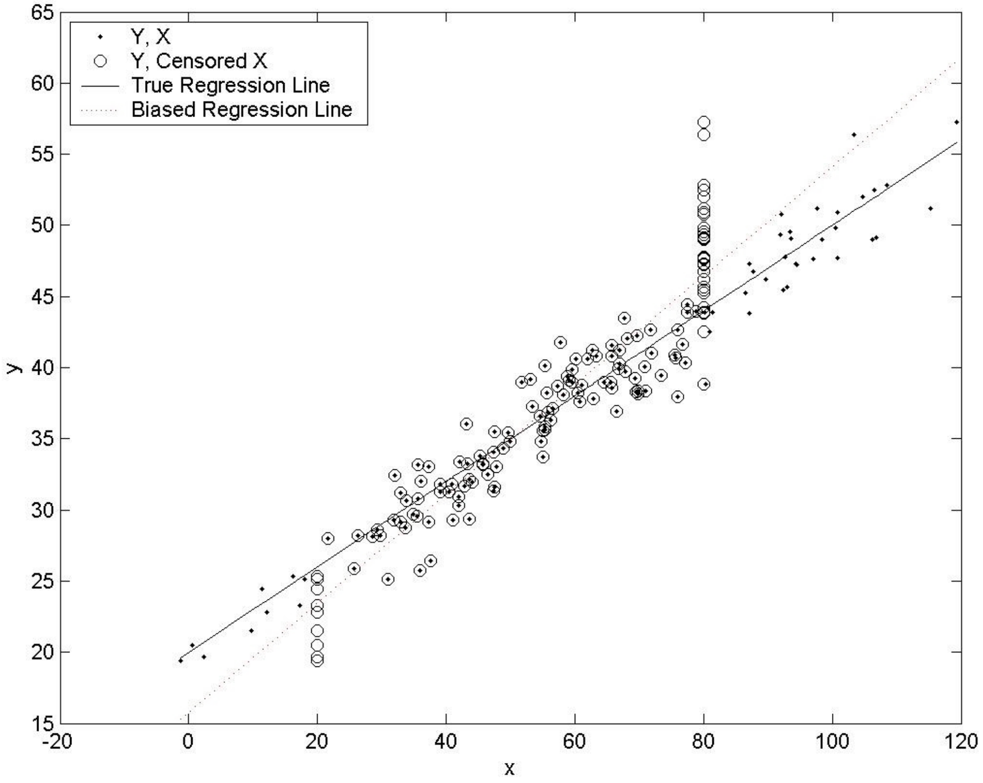

Figure 1: Expansion Bias

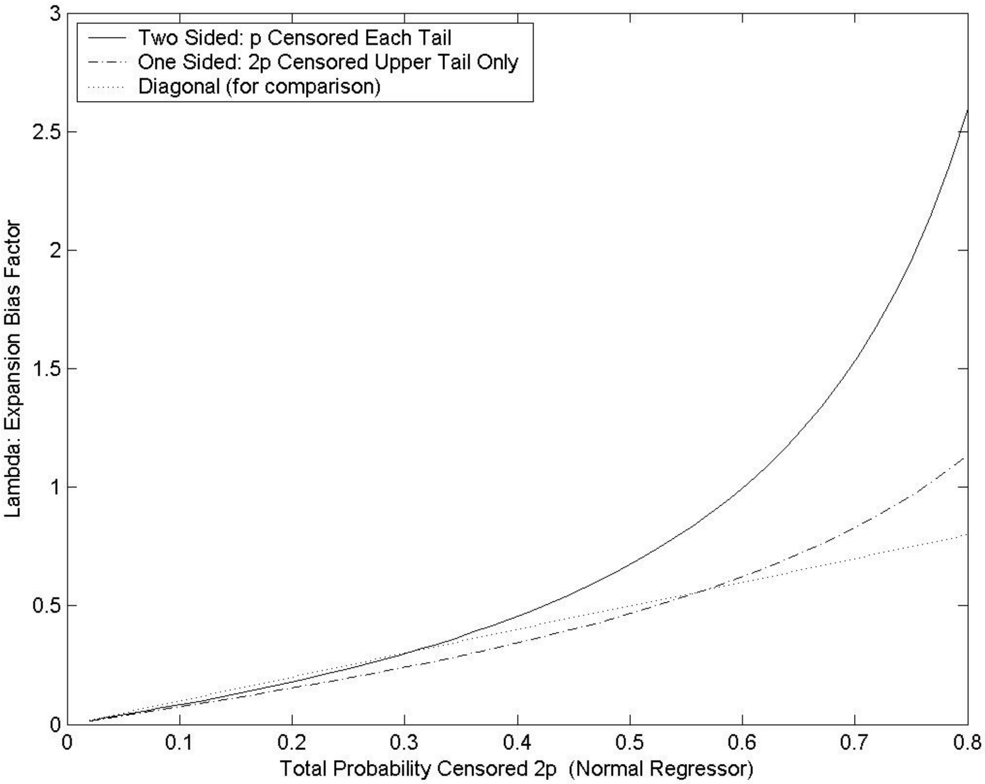

Figure 2: Expansion Bias by Probability Censored

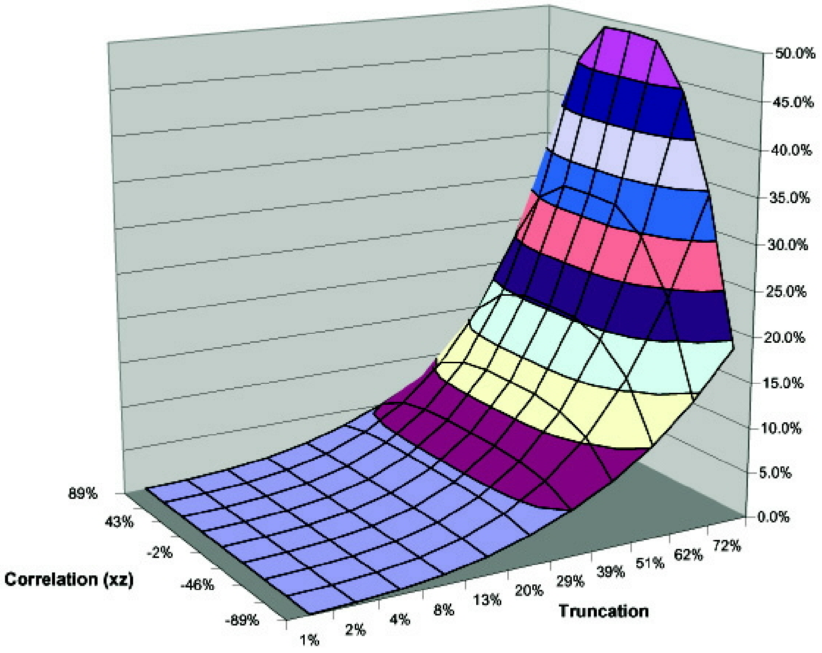

Figure 3: Bias: Intercept

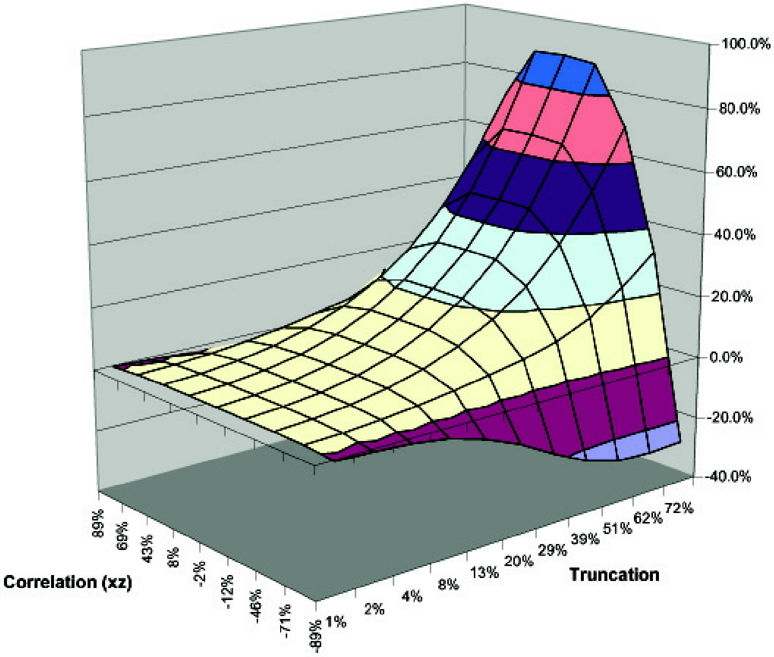

Figure 4: Bias: Coefficient of Censored Regressor

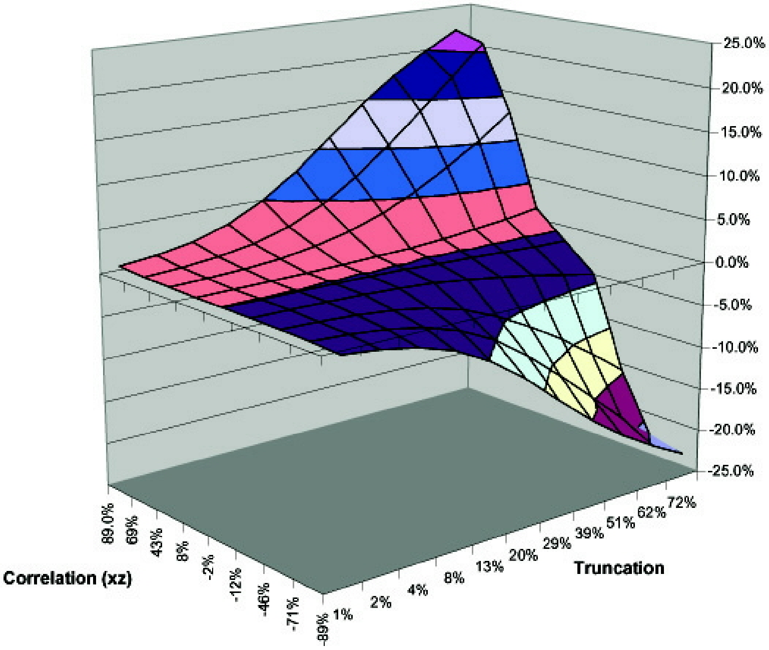

Figure 5: Bias: Coefficient of Uncensored Regressor

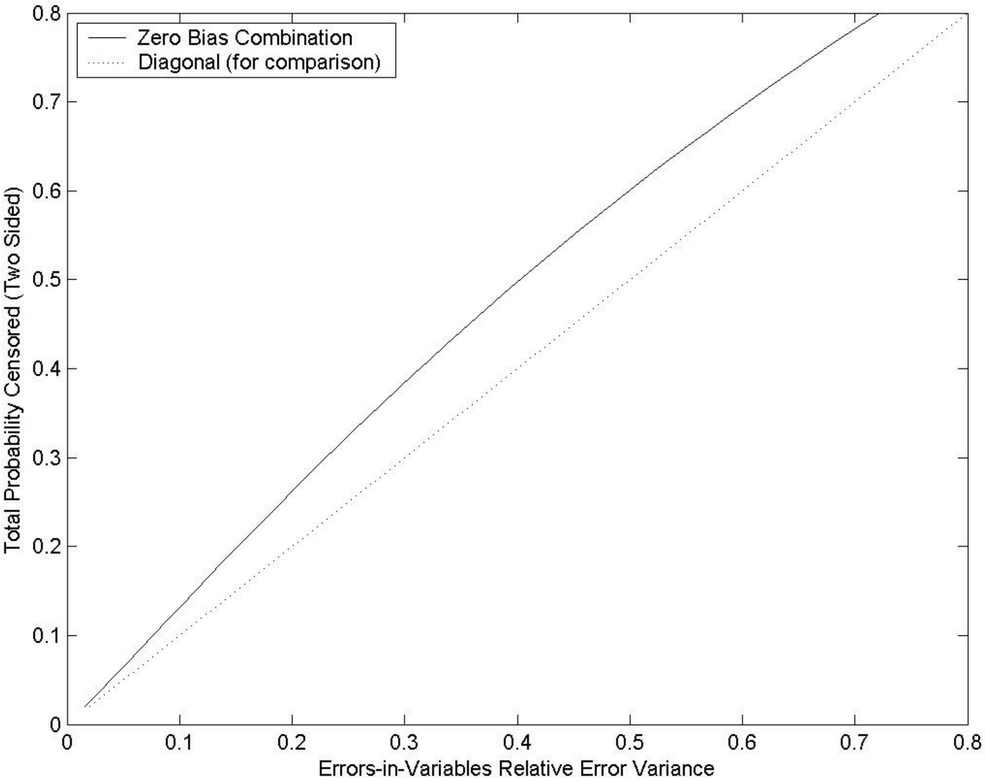

Figure 6: Errors-in-Variables and Expansion Bias: Normal Regressor

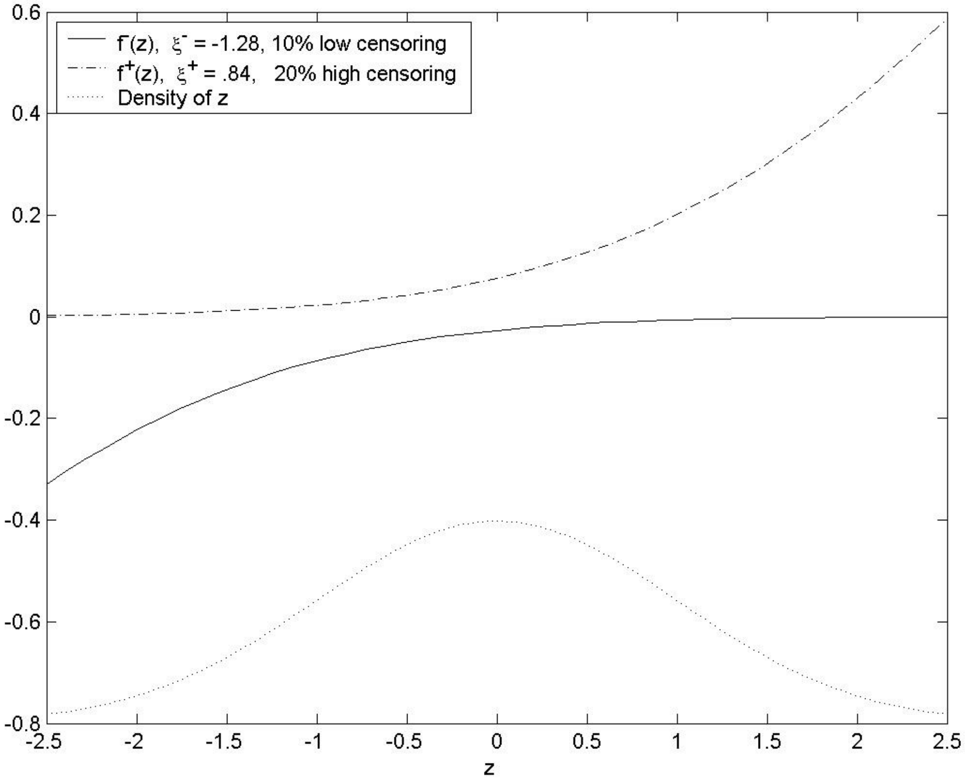

Figure 7: Regression Correction Terms: Normal Regressors