Bias from Censored Regressors

Roberto Rigobon Thomas M. Stoker*

October 2006, revised October 2007

Abstract

We study the bias that arises from using censored regressors in estimation of linear models. We present results on bias in OLS regression estimators with exogenous censoring, and IV estimators when the censored regressor is endogenous. Bound censoring such as top-and bottom-coding result in expansion bias, or effects that are too large. Independent random censoring results in bias that varies with the estimation method; attenuation bias in OLS estimators and expansion bias in IV estimators. We note how large biases can result when there are several regressors, and how that problem is particularly severe when a 0-1 variable is used in place of a continuous regressor.

1. Introduction

When the values of the dependent variable of a linear regression model are bounded and censored, the OLS estimates of the regression coefficients are biased. This well-known fact has stimulated a great deal of work on how to estimate coefficients when there is a censored dependent variable. Coefficients will also be biased if the values of regressors are censored, but there has been very little study of this phenomenon in the literature. This is a little odd, because in practice one encounters censoring in regressors or independent variables as often, or more often, than censoring in dependent variables. Moreover, as we show in this paper, the biases implied by censored regressors can be large and very insidious for practical work. In many cases estimated effects are systematically larger than the true effects, which we refer to as expansion bias. When

* Sloan School of Management, MIT, 50 Memorial Drive, Cambridge, MA 02142 USA. We are grateful for valuable comments from several seminar audiences, and want to specifically thank Elie Tamer, Mitali Das, Jinyong Hahn, Jerry Hausman, Whitney Newey and the reviewers.

one is trying to discover what are the most important influences in an empirical problem, having estimates that are too large can be at least as troublesome as having estimates that are too small or of the wrong sign.

The estimation of a model with censored regressors can often be approached as an estimation problem with missing data, as covered in Little (1992) and Little and Rubin (2002) among many others. That is, censored values can be treated as missing. As such, many procedures exist for data missing at random; which often apply to data censored at random, including various imputation strategies. When censoring is exogenous, estimation can proceed with complete cases only – where all observations with censored values are omitted. When the censoring process is modeled parametrically, likelihood methods are applicable – for instance, bound censoring (top-coding and bottom-coding) is a nonignorable data coarsening that could be approached as in Heitjan and Rubin (1990, 1991) Alternatively, one may approach censoring via partial identification as Manski and Tamer (2002) do for interval data; see also Magnac and Maurin's (2004, 2007) extension of Lewbel's(2000) results on identifying binary response models to interval data. Various semiparametric estimation methods for missing data and for nonstandard measurement error have been studied in recent work, some of which can be applied to situations with censored regressors, such as Ai (1997), Chen, Hong and Tamer (2005), Chen, Hong and Tarossa (2004), Horowitz and Manski (1998,2000), Black, Berger and Scott (2000), Liang, Wang, Robins and Carroll (2004), Tripathi (2003,2004), Mahajan (2006) and Ichimura and Martinez-Sanchis (2005), among many others. Ridder and Moffit (2003) survey another related literature, that on data combination.

Despite this, it is still a common practice to ignore the censored character of regressors in empirical work in economics. There are several reasons for this. With exogenous censoring, the set of complete cases is often a very small fraction of the original data, so that estimation based only on complete cases involves substantially lower precision than with the full sample. The censoring may occur in control variables that are of secondary interest, such as using a 0-1 variable in place of a continuous regressor. Finally, various kinds of imputations for censored or missing data values can be viewed as forms of censoring themselves, such as replacing top-coded values with an estimate of a tail mean. While imputations no doubt improve the situation there can still be errors introduced into estimation. Finally, some studies include a dummy variable that indicates censoring as an additional regressor; but this practice is very flawed (see Rigobon

and Stoker (2007) for a detailed criticism) To understand the implications of these issues, a bias analysis is in order. To our knowledge, there exists no systematic study of the biases induced by censored regressors in the literature. That is the purpose of this paper.

We present many results on bias from estimating with censored regressors in a linear model. We have results for various censoring structures, including our two primary examples of independent random censoring and censoring to an upper or lower bound. We cover OLS estimators for the case where the censored regressor is exogenous, and we cover IV estimators for the case where the censored regressor is endogenous, including the bias that occurs with random assignment. We derive explicit formulae and illustrate the bias graphically for models with a single regressor, and we cover the often severe transmission of bias that occurs when there are several regressors. While the majority of the exposition focuses on single-value censoring, we close with some results on the biases that arise when a 0-1 variable is used in place of a continuous regressor.

It is useful to keep in mind various ways that censoring arises in observed data. One source is where 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, and would have a top-coded response of, say, "$100,000 and above." Nonreponses are sometimes recorded at a bound value. For instance, household financial wealth (e.g. stock holdings) may be recorded with a lower bound of zero, and some of those zero values may be genuine zeros or may be nonresponses.

A second source of censoring occurs because observed data does not match the economic concept of interest, and is a censored version of it. Suppose we are interested in the impact of cash constraints on firm investment behavior. One cannot observe how cash-constrained a firm actually is, only imperfect reflections of it, such as the fact that dividends are paid or new debt was recently issued. Consider observed dividends as a measure of cash constraints. If dividends are large and positive, the firm is very likely not cash constrained. However, dividends are never observed to be negative, and zero values could represent a firm with either minimal cash constraints, or a firm that is struggling with major cash constraints. As such, observed dividends are a censored version of "lack of cash constraints," which is the concept of interest.

Also germane to our discussion is the use of a dummy (0-1) regressor in place of a continuous regressor. Consider the classical problem of returns to education. Suppose the mean of individual

log-wages is specified as a function of various controls and the number of years of education, which is not observed. If a regression analysis of log-wages includes a 0-1 variable indicating whether the individual has a college degree, the 0-1 variable is a censored version of the true regressor. The same kind of censoring would exist if there are separate discernible returns to high school, college, possibly varying with major or specialization and post-graduate study. Such severe forms of censoring raises issues for the interpretation of the estimates of coefficients of the 0-1 regressor, as well as coefficients of other variables in the equation. We analyze this situation below, noting how biases can force the coefficient of a 0-1 variable to have the wrong sign, as well as generate large expansion bias in coefficients of correlated regressors.

The bias induced by censored regressors varies with the type of censoring and the estimation procedure. For example, with bound censoring, the bias typically induces OLS estimates and IV estimates to be too large. In contrast, with independent random censoring, the bias typically induces OLS estimates to be too small, but IV estimates to be too large. We note how the transmission of bias with correlated regressors can result in enormous bias in estimation, including the case of 0-1 regressors. Many of our results are straightforward, but again, we are not aware of any results of this kind reported previously in the literature. The order of the presentation is as follows; we begin with results for simple regression models for intuition and explicit formulae, we then cover bias transmission in multivariate models, and finally examine bias from 0-1 censoring.

2. Bias from Censored Regressors

We are interested in bias from using a censored regressor in estimating a linear model. We assume that the true model is an equation of the form

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

where x_i is a single regressor of interest, and w_i is a k-vector of predictor variables. We assume that the distribution of $\left(x_i, w_i', \varepsilon_i\right)$ is nonsingular and has finite second moments. For studying regression estimators, we assume $E\left(\varepsilon_i | x_i, w_i\right) = 0$.

The problem is that we do not observe x_i, but rather a censored version of it. Consider

a censoring process described by an indicator d_i, and denote the probability of censoring as $p = \Pr\{d = 1\}$. When the regressor is censored, it is set to the value ξ. That is, we observe

$$x_i^{cen} = (1 - d_i)\, x_i + d_i \xi \tag{2}$$

where x_i^{cen} is the censored version of x_i. The question of interest is what happens if we estimate the model

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

How are the estimates $\hat{a},\ \hat{b},\ \hat{f}$ biased as estimators of $\alpha,\ \beta,\ \phi$?

While the censoring process can be quite general, for our analysis of regression, we exclude the problems that arise from censoring the dependent variable y_i. We assume that the regressor x_i is *exogenously censored*, by assuming that

$$E\left(\varepsilon_i | d_i, x_i, w_i\right) = 0 \tag{4}$$

We will drop this assumption in Section 2.3, when x_i is assumed to be endogenous.

There are two primary examples we will allude to throughout the text. First is independent random censoring, where d_i is taken as independent of x_i and w_i. Second is bound censoring, as in *top-coding* with censoring above an upper bound, where $d_i = 1\left[x_i > \xi\right]$, or *bottom-coding* with censoring below a lower bound, where $d_i = 1\left[x_i < \xi\right]$. Many of our results will extend immediately to the case of double bounding, where there is top-coding with upper bound ξ_1 together with bottom coding with lower bound ξ_0. In the terminology of Little and Rubin (2002), our notion of independent random censoring occurs when a censoring value assigned to observations that are MCAR ("missing completely at random"). Top-coding and bottom-coding involve censoring determined by the value of the regressor, so that they are analogous to NMAR processes ("not missing at random.") Similarly, in line with Horowitz and Manski (1995) and the robustness literature, independent random censoring corresponds to "contaminated sample" of x values, whereas bound censoring is a "corrupted sample."

For the next two subsections, we focus on models with a single regressor, where much of the intuition is available.

2.1. Censoring with a Single Regressor

We now assume that the true model has only a single regressor,

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad i = 1, ..., n \tag{5}$$

with the same distributional assumptions, including $E(\varepsilon_i|x_i) = 0$. We are interested in the (asymptotic) bias in the OLS coefficient $\hat{b}$ from estimating the model

$$y_i = a + bx_i^{cen} + u_i \quad i = 1, ..., n. \tag{6}$$

where x_i^{cen} is the censored version of x_i given in (2), namely

$$\hat{b} = \frac{\sum_{i=1}^{n} (x_i^{cen} - \bar{x}^{cen})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i^{cen} - \bar{x}^{cen})^2}. \tag{7}$$

It is useful to express the bias in proportional terms, as

$$\text{plim } \hat{b} = \beta (1 + \Lambda). \tag{8}$$

There is no bias if $\Lambda = 0$. *Attenuation bias* refers to the situation where $-1 < \Lambda < 0$. *Expansion bias* refers to the situation when $\Lambda > 0$.

Proposition 1 gives a general characterization of the bias in simple OLS regression .

Proposition 1. ***OLS Bias: Single Regressor:*** *Provided that* $0 < p < 1$, *we have*

$$\Lambda = p(1-p) \cdot \frac{(E(x|d=1) - \xi)(\xi - E(x|d=0))}{Var(x^{cen})}. \tag{9}$$

Consequently,

1. $\Lambda = 0$ *if and only if* $\xi = E(x|d=1)$ *or* $\xi = E(x|d=0)$,
2. $\Lambda > 0$ *if and only if*

$$E(x|d=0) < \xi < E(x|d=1) \text{ or } E(x|d=1) < \xi < E(x|d=0),$$

so that $\Lambda < 0$ *otherwise.*

The proofs of all results are direct, and given in Appendix A. It is clearly important to consider the censoring process d_i and the censoring value ξ separately. For instance, the sign and extent of the bias are affected by the position of ξ: there is no bias if ξ equals either conditional mean $E(x|d=1)$ or $E(x|d=0)$, there is expansion bias if ξ is between the means, and there is attenuation bias in all other circumstances.

Our primary examples are addressed by two immediate corollaries. First, with independent random censoring, there is attenuation bias in OLS coefficients:

Corollary 2. ***OLS Bias: Uncorrelated Censoring.*** *Suppose* $Cov(x,d)=0$ *and* $Cov(x^2,d)=0$. *Then*

$$\Lambda = p \cdot \frac{-\left(\xi - E\left(x\right)\right)^2}{Var\left(x\right) + p\left(\xi - E\left(x\right)\right)^2}. \tag{10}$$

We have $\Lambda \leq 0$, *with equality holding only if* $\xi = E\left(x\right)$.

For bound censoring, there is expansion bias in OLS coefficients:

Corollary 3. ***OLS Bias: Bound Censoring.*** *Suppose* x_i^{cen} *is*

1. *top-coded at* ξ_1, *with* $d_i = 1\left[x_i > \xi_1\right]$, *then plim* $\hat{b} = \beta\left(1+\Lambda_1\right)$ *where* $\Lambda_1 > 0$.
2. *bottom-coded at* ξ_0, *with* $d_i = 1\left[x_i < \xi_0\right]$, *then plim* $\hat{b} = \beta\left(1+\Lambda_0\right)$ *where* $\Lambda_0 > 0$.
3. *top-coded at* ξ_1 *and bottom-coded at* ξ_0, *with* $\xi_0 < \xi_1$, *then plim* $\hat{b} = \beta\left(1+\Lambda\right)$ *where* $\Lambda > \Lambda_1 + \Lambda_0$ *above.*

Figure 1 illustrates bias with independent random censoring, where $\xi > E\left(x\right)$. The small circles indicate the censored data points, including the block of points with $x_i = \xi$. Clearly the observations with $x_i = \xi$ have center of mass below the regression line, which induces attenuation bias. The case with $\xi < E\left(x\right)$ is similar, and no bias arises only if $\xi = E\left(x\right)$. In contrast, Figure 2 illustrates the bias with top-coding, or censoring to an upper bound ξ. The "pile-up" of observations on the bound induces expansion bias in the coefficient. The same result would arise with a lower bound from bottom-coding, or with both top-coding and bottom-coding.

The specific formula for bias from top-coding and bottom-coding depends on expectations over truncated distributions. We can get a sense of the size of the bias by computing it for a specific distribution. Figure 3 presents the expansion bias from one-sided and two-sided bound censoring under the assumption that x_i is normally distributed. (Detailed bias formulae are available from the authors.) This is computed for different levels of censoring (p), which is equivalent to setting different bound limits (ξ). The solid line displays the bias Λ of from top- and bottom-coding under the assumption of symmetric (two-sided) censoring, with probability p censored in each tail, and it is plotted against the total probability of censoring $2p$. The dashed line is the expansion bias Λ_1 from using top-coded data (one-sided censoring), which is plotted against the total same total censoring probability. For instance, plotted over $2p = .20$ 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.

2.2. Problems with Residuals

At this point it is useful to make some points about using the residuals from a regression with a censored regressor. These points apply in all the situations we consider below, but are easiest seen with simple regression.

Often, regression analysis is used in a first-stage analysis, with residuals of primary interest. That is, the original model

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

captures the notion that ε_i represents y_i after removing the influence of x_i. Here ε_i is mean independent of x_i or at least, there is zero correlation between ε_i and x_i. First-stage regression analysis produces residuals $\hat{\varepsilon}_i$ that are consistent estimators of ε_i, which are used for the subsequent, second-stage analysis. In this setting, the coefficient values (β) are of secondary interest, whereas the residuals are of primary importance. For instance, household consumption is often regressed on some life-cycle controls, with the residuals from that regression used in subsequent

analysis.

Suppose that the first-stage estimates are performed with a censored regressor. This creates serious problems for computed residuals. In addition to using biased coefficients, the residuals also reflect variation in the censored data. That makes the computed residuals very poor proxies for the true ε_i's, which would often invalidate their use in second-stage analysis.

To see these points, suppose that (6) is estimated (with the censored regressor) and the residual is computed as

$$\hat{u}_i \equiv y_i - \hat{a} - \hat{b}x_i^{cen} \quad i = 1, ..., n. \tag{11}$$

We eliminate sampling variation in the coefficients by working with $U_i \equiv \text{plim } \hat{u}_i = y_i - (\text{plim } \hat{a}) - \left(\text{plim } \hat{b}\right) \cdot x_i^{cen}$. Our interest is in how similar U_i is to ε_i.

Using our notation that plim $\hat{b} = \beta(1 + \Lambda)$, it is straightforward to show that

$$\begin{aligned} U_i &= \varepsilon_i - \beta\Lambda \cdot (x_i - E(x)) + \beta(1 + \Lambda) \cdot [d_i(x_i - \xi) - E(d(x - \xi))] \\ &= \varepsilon_i - A - \beta\Lambda \cdot x_i + \beta(1 + \Lambda) \cdot d_i(x_i - \xi) \end{aligned}$$

Therefore U_i is seriously different than ε_i, by terms that vary with x_i. The difference arises because of coefficient bias (if $\Lambda \neq 0$), but more seriously, because U_i contains $d_i(x_i - \xi)$, the part of x_i that is censored. To check similarity to the correlation condition $Cov(\varepsilon, x) = 0$, we have

$$\begin{aligned} Cov(U, x) &= -\beta\Lambda \cdot Var(x) + \beta(1 + \Lambda) \cdot Var(x|d = 1) \\ &\quad + \beta(1 + \Lambda) \cdot p(1 - p)[E(x|d = 1) - \xi][E(x|d = 1) - E(x|d = 0)] \end{aligned}$$

using derivations similar to those in the Appendix. Even when there is no bias, $\Lambda = 0$, there still is nonzero covariance (unless $\beta = 0$). For instance, in view of (9), if the censoring point is set to $\xi = E(x|d = 1)$, then $\Lambda = 0$ but we still have

$$Cov(U, x) = \beta \cdot Var(x|d = 1)$$

This occurs because censored variations in x_i appear in U_i (through y_i).

The residual U_i will typically bear a nonlinear relationship to the true x_i. For instance, if $E(\varepsilon_i|x_i = x) = 0$, we have

$$E(U_i|x_i = x) = -A - \beta\Lambda \cdot x + \beta(1 + \Lambda) \cdot p(x)(x - \xi)$$

where $p(x) = E(d_i|x_i = x)$ is a nonzero function of x. If the censoring is independent of x_i, with $p(x) = p$, then $E(U_i|x_i = x)$ is linear in x.

These dependencies would have serious consequences for using $\hat{u}_i$ in second-stage analysis. That is, effects estimated using $\hat{u}_i$ could be due to variations in x_i, which invalidates the purpose of the first-stage analysis. When censoring is severe, such as with 0-1 censoring as described later, it would be difficult to see how to correct for such substantial error in measurement.

2.3. Endogenous Censored Regressors

We now consider the case where the regressor that is censored is endogenous, and where we have a valid instrument for the uncensored regressor. We will see that censoring induces bias in IV estimators that is quite different than bias in OLS coefficients.

As above, we remain with the single regressor format

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad i = 1, ..., n \tag{12}$$

where x_i is now an endogenous regressor and z_i denotes a valid instrument for x_i. In particular, we assume that $\{(x_i, z_i, \varepsilon_i) \mid i = 1, ..., n\}$ is an i.i.d. random sample from a distribution with finite second moments, with $E(\varepsilon) = 0$, $Cov(z, x) \neq 0$ and $Cov(z, \varepsilon) = 0$. This implies that

$$\frac{Cov(z, y)}{Cov(z, x)} = \beta \tag{13}$$

which is the consistent limit of the IV estimator if there were no censoring.

Instead of x_i, assume we observe

$$x_i^{cen} = (1 - d_i) x_i + d_i \xi. \tag{14}$$

where again d_i represents a general censoring process with $p = \Pr\{d = 1\} \neq 0$. We assume that d_i is such that $Cov(z, x^{cen})$ exists and is nonzero. (Note, for bias analysis, we do not assume an exogenous censoring condition such as $E(\varepsilon_i | d_i, z_i) = 0$)

When we use x_i^{cen} in estimation, we compute the IV estimator $\hat{c}$ that uses z_i to instrument x_i^{cen}. We have

$$\hat{c} = \frac{\sum_{i=1}^{n} (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n} (z_i - \bar{z})(x_i^{cen} - \bar{x}^{cen})} \rightarrow \frac{Cov(z, y)}{Cov(z, x^{cen})} \tag{15}$$

where $Cov(z, x^{cen}) \neq 0$ is assumed, since otherwise $\hat{c}$ has no probability limit. Thus

$$\text{plim } \hat{c} = \frac{Cov(z, x)}{Cov(z, x^{cen})} \cdot \beta. \tag{16}$$

The bias of the IV estimator is expressed in proportional form as

$$\text{plim } \hat{c} = (1 + \Lambda) \cdot \beta, \quad \text{with } \Lambda = \frac{Cov(z, x^o)}{Cov(z, x^{cen})}. \tag{17}$$

where $x_i = x_i^{cen} + x_i^o$, with

$$x_i^o = d_i (x_i - \xi) \tag{18}$$

the part of x_i lost by censoring. This makes it clear why the IV estimator is biased. The instrument z_i is valid in the "true" data; z_i is correlated with x_i and uncorrelated with the disturbance ε_i. With censored data, the difference x_i^o is omitted and correlated with the instrument. That is, z_i is correlated with $\beta x_i^o + \varepsilon_i$, so z_i is not a valid instrument for the equation with x_i^{cen}.

Proposition 4 characterizes IV bias.

Proposition 4. ***IV Bias: Single Endogenous Regressor:*** *The proportional IV bias* Λ *of (17) is*

$$\Lambda = \frac{p}{1 - p} \frac{Cov(z, x | d = 1) + (1 - p) \Delta_z [E(x | d = 1) - \xi]}{Cov(z, x | d = 0) + p \Delta_z [\xi - E(x | d = 0)]} \tag{19}$$

where

$$\Delta_z \equiv E(z | d = 1) - E(z | d = 0).$$

The expression (19) shows that IV bias depends on the distribution of (z, x) for censored and uncensored data through conditional means and covariances, but also on the value ξ that

the data is censored to.

For independent random censoring, we have a striking result:

Corollary 5. ***IV Bias: Uncorrelated Censoring.*** *If* $Cov(z,d) = 0$, $Cov(x,d) = 0$ *and* $Cov(zx,d) = 0$, *then*

$$\Lambda = \frac{p}{1-p} \tag{20}$$

with $\Lambda > 0$ *always.*

The conditions are equivalent to assuming that z, x and zx are mean independent of d, and clearly include statistical independence of (z,x) and d. The IV bias is always positive, varies directly with the amount of censoring p, but is not affected by the distribution of (z,x) or the censoring point ξ. This is in strong contrast to the attenuation bias induced in OLS estimators (or zero bias when $\xi = E(x)$). (While not related to censoring, Black, Berger and Scott (2000) note the same feature, that the bias in OLS coefficients is in the opposite direction of the bias in IV estimates, for a specific measurement error model.)

The role of the distribution and of the censoring point are clarified by simplifying (19) using "partial" lack of correlation. First, if censoring is only uncorrelated with the instrument, $Cov(z,d) = 0$, then

$$\Lambda = \frac{p}{1-p} \frac{Cov(z,x|d=1)}{Cov(z,x|d=0)} \tag{21}$$

so that (20) is modified by the covariances in censored and uncensored data. Expansion bias follows if they are of the same sign. If censoring is only uncorrelated with the endogenous regressor, $Cov(x,d) = 0$, then

$$\Lambda = \frac{p}{1-p} \frac{Cov(z,x|d=1) + (1-p)\varphi}{Cov(z,x|d=0) - p\varphi} \tag{22}$$

where $\varphi \equiv \Delta_z \cdot [E(x) - \xi]$. Here the censoring value ξ is relevant, and $\varphi = 0$ only when $\xi = E(x)$.

For bound censoring, we have the following result, where we have assumed $Cov(z,x) > 0$ without loss of generality:

Corollary 6. ***IV Bias: Bound Censoring.*** *Suppose* $Cov\left(z, x^{cen}\right) > 0$. *If* x *is top-coded at* ξ *and* $Cov\left(z, d\right) > 0$ ***or*** x *is bottom-coded at* ξ *and* $Cov\left(z, d\right) < 0$, *then*

$$\Lambda > 0$$

if and only if

$$Cov\left(z, x|d = 1\right) > -\left(1 - p\right) \cdot \Delta_z \cdot \left[E\left(x|d = 1\right) - \xi\right] \tag{23}$$

This result says that expansion bias arises unless the correlation structure of the censored data is radically different than that of the uncensored data. With either top-coding or bottom-coding, $\Delta_z \cdot \left[E\left(x|d = 1\right) - \xi\right] > 0$, so that right-hand side of (23) is negative. The proportional bias $\Lambda > 0$ unless $Cov\left(z, x|d = 1\right)$ is so negative as to invalidate (23). Clearly, $\Lambda > 0$ if $Cov\left(z, x|d = 1\right) \geq 0$ in either case.

It would be desirable to discover some primitive conditions that are closely associated with IV expansion bias when there is bound censoring. Unfortunately, all the primitive conditions that the authors have discovered are much stronger than the tenets of Corollary 6. Of those, there is one set of conditions worth mentioning, which essentially assures a positive relation between the endogenous regressor, instrument and censoring. This is where z is **mean-monotonic** in x; namely

$$E\left(z|x = x_1\right) \geq E\left(z|x = x_0\right) \text{ for any values } x_1 \geq x_0 \tag{24}$$

This condition guarantees all the conditions of Corollary 6, as summarized in

Corollary 7. ***IV Bias: Bound Censoring with Mean Monotonicity of*** z ***in*** x. *Suppose (24) holds, then* $\Lambda > 0$ *if* x *is top-coded at* ξ ***or*** x *is bottom-coded at* ξ.

Mean-monotonicity gives a uniform structure to the covariance of z and x over ranges of x values. It is not a counterintuitive condition. For instance, if (z, x) is joint normally distributed then z is mean-monotonic in x. But it clearly implies a strong relationship between the endogenous regressor and the instrument.

2.3.1. Special Case: Random Assignment

We can get some intuition for the bias results from the case where the instrument represents random assignment into two groups. Assume z is a binary instrument indicating groups 0 and 1, and denote the probability of being in group 1 as $q = \Pr\{z = 1\} \neq 0$. Here $Cov(z, x) > 0$ implies $E(x \mid z = 1) > E(x \mid z = 0)$, or that the grouping is associated with a shift in the mean of x. Likewise $Cov(z, \varepsilon) = 0$ implies $E(\varepsilon \mid z = 1) = E(\varepsilon \mid z = 0) = 0$, or that there is no shift in the mean of ε associated with the grouping.

The IV estimator with instrument z_i is the group-difference estimator of Wald (1940); without censoring, this is $\hat{\gamma} = (\bar{y}_1 - \bar{y}_0) / (\bar{x}_1 - \bar{x}_0)$ where $\bar{y}_0$, $\bar{x}_0$ are averages over group 0 and $\bar{y}_1$, $\bar{x}_1$, are the averages over group 1. Equation (13) is now

$$\frac{E(y \mid z = 1) - E(y \mid z = 0)}{E(x \mid z = 1) - E(x \mid z = 0)} = \beta. \tag{25}$$

When we use the censored regressor x_i^{cen} instead of x_i, the IV estimator is the group difference estimator $\hat{c} = (\bar{y}_1 - \bar{y}_0) / (\bar{x}_1^{cen} - \bar{x}_0^{cen})$, with

$$\text{plim } \hat{c} = \frac{E(y \mid z = 1) - E(y \mid z = 0)}{E(x^{cen} \mid z = 1) - E(x^{cen} \mid z = 0)} = (1 + \Lambda) \cdot \beta, \tag{26}$$

$$\Lambda = \frac{E(x^o \mid z = 1) - E(x^o \mid z = 0)}{E(x^{cen} \mid z = 1) - E(x^{cen} \mid z = 0)}. \tag{27}$$

and again, $x_i^o = x_i - x_i^{cen} = d_i(x_i - \xi)$. The size and sign of the IV bias Λ is determined by how the censoring operates on the two different assignment groups. There is expansion bias, $\Lambda > 0$, when

$$E(x^{cen} \mid z = 1) - E(x^{cen} \mid z = 0) < E(x \mid z = 1) - E(x \mid z = 0) \tag{28}$$

or equivalently that the mean shifts $E(x^o \mid z = 1) - E(x^o \mid z = 0)$ and $E(x^{cen} \mid z = 1) - E(x^{cen} \mid z = 0)$ are of the same sign. There is attenuation bias, $\Lambda < 0$, only when they are of opposite signs.

Consider independent random censoring. It is easy to show that

$$E(x^{cen} \mid z = 1) - E(x^{cen} \mid z = 0) = (1 - p) \cdot [E(x \mid z = 1) - E(x \mid z = 0)]$$

so (28) always holds, and the proportional IV bias is

$$\Lambda = \frac{p}{1-p}.$$

This matches Corollary 5. The expansion bias arises because the (between group) mean shift for x^{cen} is a simple proportion of that for x, with the relative amount not varying with the distribution of x.

For intuition on bound censoring, consider the implication of top-coding. Since $E(x \mid z=1) > E(x \mid z=0)$, one might expect more large x values in group 1 than 0, with a bigger mean impact of censoring on group 1 than group 0, or

$$E(x \mid z=1) - E(x^{cen} \mid z=1) > E(x \mid z=0) - E(x^{cen} \mid z=0). \tag{29}$$

But this is just condition (28) for expansion bias. The problem is that if group 0 has much bigger variance than group 1, top-coding would have a greater impact on group 0 mean and (29) would fail. This type of situation is eliminated by mean monotonicity.

Figures 4 and 5 illustrate IV bias with independent random censoring, with the following specification. The true model is $y_i = 2 + .5x_i + \varepsilon_i$. The probability of $z_i = 1$ is $q = .6$. We have (x_i, ε_i) joint normal conditional on z_i; with mean $(2, 0)$ for $z_i = 0$ and mean $(6, 0)$ for $z_i = 1$. The covariance of (x_i, ε_i) is the same for each z_i; the variance of x_i is 4 and of ε_i is 1 and the correlation between x_i and ε_i is $-.5$. Figure 4 shows the uncensored data, including the z value grouping, with substantial overlap between the groups. Also illustrated are the group means and the uncensored IV (group difference) estimator.

Figure 5 shows what happens with 30% random censoring. The censored data are shown as small circles, with the censoring value $\xi = 4$. The IV fit is clearly steeper, which illustrates the positive IV bias. Mechanically, the within-group means of x are both shifted toward ξ but the within-group means of y are unchanged, so that the slope (26) is increased. As consistent with the our formulae, slope bias does not depend on the specific censoring value ξ. The same tilting would occur if $\xi = 0$ or $\xi = 6$, for instance.

Figure 6 illustrates how expansion bias arises from top-coding. The same data are used here as in Figures 4 and 5, and now censoring occurs for values greater than $\xi = 6$. Clearly, much

more censoring occurs for observations with $z = 1$ than for those with $z = 0$, so this example illustrates (29). Alternatively, if the $z = 0$ group had much wider dispersion than the $z = 1$ group, then top-coding could involve censoring more of the right tail of the $z = 0$ group than that of the $z = 1$ group, and (29) would fail. In our calculations, we found that this occured if the standard error of x_i for $z_i = 0$ were multiplied by 4.

3. Multivariate Regression

The case when there are several regressors is the most relevant to empirical practice. Bias from censoring one variable can contaminate the estimates of coefficients of other variables, which we refer to as bias transmission. This is important when the censored variable is representing an effect of great interest, but perhaps of more importance when the censored variable is not the primary focus of interest. We may, in fact, care the most about the coefficients of the other variables, including the censored variable as a generalized control. With bias transmission, the estimates of the coefficients of primary interest can be badly wrong because of including an imperfect control. We now focus on this situation for censoring as we have studied above. In Section 4, we focus on the same problem when we use a 0-1 variable in place of a continuous regressor.

3.1. Bias Transmission with Several Regressors

It is often very difficult to obtain specific results on coefficient bias when there are several regressors (a exception is Klepper and Leamer (1984) on multivariate errors-in-variables). When one regressor is censored, we are able to get some broad insight as well as a couple specific results, as presented below. We focus only on OLS estimators, although results of a similar nature should be available for IV estimators.

We return to basic model (1), where we assume a single additional regressor w_i, given as

$$y_i = \alpha + \beta x_i + \phi w_i + \varepsilon_i \quad i = 1, ..., n. \tag{30}$$

The regressor x_i is censored as $x_i^{cen} = (1 - d_i)\, x_i + d_i \xi$, and we are interested in what happens

when we estimate the model

$$y_i = a + bx_i^{cen} + fw_i + u_i \quad i = 1, ..., n., \tag{31}$$

How are the OLS estimates $\hat{b}$ and $\hat{f}$ biased as estimators of β and ϕ?

In broad terms, the issue centers on how well w_i proxies x_i for the censored observations. Write (30) as

$$y_i = \alpha + \beta x_i^{cen} + \phi w_i + \beta x_i^o + \varepsilon_i \quad i = 1, ..., n, \tag{32}$$

with $x_i^o = d_i (x_i - \xi)$ as before. If w_i and x_i^o are only slightly correlated, then $\hat{f}$ will estimate ϕ with little bias, and $\hat{b}$ will estimate β with the bias appropriate for regression with a single regressor. If w_i and x_i^o are highly correlated, then $\hat{f}$ will be very biased as an estimate of ϕ, as w_i acts to proxy for x_i in the censored data.

We can sharpen this logic somewhat. There is no bias transmission when $\hat{f}$ is consistent for ϕ, which occurs if $Cov(x^{cen}, w) = 0$. We can develop this as

$$\begin{aligned} 0 &= Cov(x^{cen}, w) = Cov(x, w) - Cov(x^o, w) \\ &= (1-p)[Cov(x, w|d=0) \\ &\quad -p\{E(w|d=1) - E(w|d=0)\}\{\xi - E(x|d=0)\}] \end{aligned} \tag{33}$$

Notice that it is not sufficient for w_i to be uncorrelated with x_i in the uncensored data. We either must have censoring to the mean, $\xi = E(x|d=0)$, or w_i uncorrelated with the censoring, $E(w|d=1) = E(w|d=0)$.

On the nature of bias when w_i and x_i are correlated, we appeal to our primary examples. Consider independent random censoring, where d_i is statistically independent of x and w. With calculations similar to those applied in Corollary 2, it is easy to show that

$$\begin{aligned} \text{plim} \begin{pmatrix} \hat{b} \\ \hat{f} \end{pmatrix} &= \begin{bmatrix} Var(x^{cen}) & Cov(x^{cen}, w) \\ Cov(x^{cen}, w) & Var(w) \end{bmatrix}^{-1} \begin{bmatrix} Cov(x^{cen}, y) \\ Cov(w, y) \end{bmatrix} \\ &= [G + pH]^{-1} G \begin{pmatrix} \beta \\ \phi \end{pmatrix} + [G + pH]^{-1} (pH) \begin{pmatrix} 0 \\ \phi + \beta \cdot \frac{Cov(x,w)}{Var(w)} \end{pmatrix} \end{aligned} \tag{34}$$

where

$$G = \begin{bmatrix} Var(x) & Cov(x,w) \\ Cov(x,w) & Var(w) \end{bmatrix} \text{ and } H = \begin{bmatrix} (\xi - E(x))^2 & 0 \\ 0 & \frac{1}{(1-p)} Var(w) \end{bmatrix} \tag{35}$$

This gives the asymptotic bias as

$$\text{plim} \begin{pmatrix} \hat{b} \\ \hat{f} \end{pmatrix} - \begin{pmatrix} \beta \\ \phi \end{pmatrix} = \beta p \cdot [G + pH]^{-1} H \begin{pmatrix} -1 \\ \frac{Cov(x,w)}{Var(w)} \end{pmatrix} \tag{36}$$

This reflects attenuation bias (as found before) if there is low correlation between w_i and x_i, as well as the transmission of bias if w_i and x_i are highly correlated. Moreover, recall that Corollary 2 showed that there is zero bias with a single regressor when $\xi = E(x)$. That is *not* true in multivariate regression with a correlated regressor. Equation (36) shows there is nonzero bias if $Cov(x,w) \neq 0$, even when $\xi = E(x)$.

For bound censoring, the exact bias formula are too complicated to admit easy interpretation (and too tedious to derive here). Some interpretation is possible if we expand the exact bias in p and examine the leading terms. In particular, suppose we have top-coding with $d_i = 1[x_i > \xi]$, and we assume $E(x) = 0$, $E(w) = 0$ for simplicity. Then we have

$$plim \begin{pmatrix} \hat{b} \\ \hat{f} \end{pmatrix} - \begin{pmatrix} \beta \\ \phi \end{pmatrix} \cong \frac{\beta \cdot p}{Var(x) Var(w) - Cov(x,w)} \begin{pmatrix} Var(w) \cdot E - Cov(x,w) \cdot C \\ Var(x) \cdot C - Cov(x,w) \cdot E \end{pmatrix} \tag{37}$$

with

$$\begin{aligned} C &= Cov(x, w|d=1) + E(w|d=1) \{E(x|d=1) - \xi\} \\ E &= \xi \cdot \{E(x|d=1) - \xi\} \end{aligned} \tag{38}$$

Suppose that w and x are positively correlated (in censored and uncensored data), and that $\xi > 0$. Because of top-coding, we have $E > 0$ and because of the positive correlation we have $C > 0$. Thus the bias in each coefficient is a difference of positive terms. With low correlation between w and x, the term C is small and the expansion bias term E dominates for $\hat{b}$, together with a negative bias term induced for $\hat{f}$. As the correlation increases, expansion bias in $\hat{b}$ decreases and, in essence, transmits to bias in $\hat{f}$.

Censoring	Bias of:	Correlation -50%	0%	50%	75%	95%
10%	$\widehat{b}$	6.2%	7.4%	6.2%	3.2%	-17.6%
	$\hat{f}$	-2.0%	0.0%	2.0%	5.7%	24.3%
20%	$\widehat{b}$	12.2%	15.5%	12.2%	5.0%	-32.8%
	$\hat{f}$	-5.2%	0.0%	5.2%	11.8%	43.8%
40%	$\widehat{b}$	26.2%	34.3%	26.2%	7.4%	-52.5%
	$\hat{f}$	-12.2%	0.0%	12.0%	26.8%	68.0%
60%	$\widehat{b}$	45.5%	62.8%	45.5%	12.8%	-61.8%
	$\hat{f}$	-20.9%	0.0%	20.9%	41.3%	80.5%

Table 1: Coefficient Biases: One Top-Coded Regressor.

It is possible for the bias to have either sign depending on the correlation between x and w. For extremely large correlation, the positive expansion bias in $\hat{b}$ can be wholly reversed, and a positive bias arises for $\hat{f}$. In that case, it appears that w is doing a better job of proxying for x than the censored x^{cen} is. We now illustrate these cases.

3.2. Bias Transmission with Normal Regressors

To clarify the intuition discussed above, we assumed that the underlying (uncensored) regressors x and w are joint normally distributed, and simulated the biases for different levels of censoring and different correlations between x and w. (Specifically, we set $\alpha = 1$, $\beta = 1$, and $\phi = 1$ and assumed that x and w have the same variance.) The biases resulting from top-coding are presented in Table 1. There are four levels of censoring from mild (10%) to severe (60%). There are five correlation values between x and w, from no correlation (0%) to moderate correlation (50%, 75%) and finally, extreme correlation (95%).

For the coefficient $\hat{b}$ of the censored regressor, the zero correlation case shows the highest expansion bias, increasing with censoring, and there is no bias in $\hat{f}$, the coefficient of the other regressor. As correlation is increased, expansion bias in $\hat{b}$ is reduced and bias emerges in $\hat{f}$. The amount of transmission is pretty substantial with moderate correlation, and we have included the cases of -.5 and .5 correlation to illustrate the symmetry in the structure of these biases. Finally, with extreme correlation, the bias is reversed for $\hat{b}$ and there is large bias in $\dot{f}$. In this case w does a better job of representing the omitted x than the censored version x^{cen}. This is all in line with the intuition discussed above.

Censoring	Bias of:	Correlation -50%	0%	50%	75%	95%
10%	$\widehat{b}$	-3.1%	0.0%	-3.1%	-11.3%	-47.9%
	$\hat{f}$	-6.3%	0.0%	6.3%	15.1%	50.2%
20%	$\widehat{b}$	-6.3%	0.0%	-6.3%	-20.6%	-64.8%
	$\hat{f}$	-12.6%	0.0%	12.6%	27.5%	68.4%
40%	$\widehat{b}$	-11.7%	0.0%	-11.7%	-33.4%	-78.6%
	$\hat{f}$	-23.6%	0.0%	23.6%	45.3%	82.6%
60%	$\widehat{b}$	-16.7%	0.0%	-16.7%	-44.1%	-84.9%
	$\hat{f}$	–33.3%	0.0%	33.3%	58.2%	89.1%

Table 2: Coefficient Biases: One Regressor Independently Censored to Mean.

Table 2 presents bias results for the situation where x is independently censored to its mean $\xi = E(x)$, with an additional regressor w in the equation. With zero correlation, there are no bias, which is consistent with the single regressor result for independent random censoring to the mean. As the correlation is increases, bias emerges in $\hat{f}$, as w is proxying for x for the censored observations, and we have a resultant attenuation bias in $\hat{b}$. This phenomenon increases monotonically in both the censoring level and the correlation value, with moderate correlation and censoring resulting in large bias in $\hat{f}$ and $\hat{b}$. In particular, censoring to the mean eliminates bias only in situations analogous to those with a single regressor – with several correlated regressors, huge bias can result in the case. With extreme correlation, even random censoring of 10% can result in coefficient biases of 50%.

3.3. Illustration with Consumption and Income Data

We now illustrate the impact of various types of censoring of income data in estimating a consumption equation. We use data from Parker (1999), which is a synthetic panel of cohorts constructed from PSID and CEX data. The model for estimation is .

$$\Delta \ln c_i = \alpha + \phi_1 \cdot \Delta \ln W_i + \phi_2 \cdot \Delta \ln YP_i + \beta \cdot \ln Y_i + \varepsilon_i \tag{39}$$

where $\Delta \ln c_i$ is the first difference (between t and $t-1$) of consumption of household i, $\Delta \ln W_i$ is the first-difference of log financial wealth (housing, stock holdings, etc.), $\Delta \ln YP_i$ is the first-difference of a measure of log permanent income (which proxies human capital wealth), and Y_i is current income (period t). See Parker (1999) for details on the construction of this data,

	$\Delta \ln W$		$\Delta \ln YP$		$\ln Y$	
Base Estimates	0.0297	(0.0057)	0.0910	(0.0099)	0.0646	(0.0062)
Random Censoring						
to median	0.0327	(0.0058)	0.1064	(0.0098)	0.0586	(0.0086)
to zero	0.0349	(0.0058)	0.1170	(0.0097)	-0.0021	(0.0010)
Top Coding						
25%	0.0305	(0.0057)	0.0916	(0.0099)	0.0773	(0.0075)
50%	0.0316	(0.0058)	0.0943	(0.0099)	0.0890	(0.0092)

Table 3: OLS Estimates of the Consumption Model

which includes 2656 observations after dropping all original observations that had top-coded income and zero financial wealth. In the model, ϕ_1 and ϕ_2 represent the marginal propensities to consume out of financial wealth and human capital respectively, and β measures the excess sensitivity (or degree of credit constraint) of consumption in current income.

Here we illustrate what happens to the estimates when we artificially censor log income in fairly severe ways. We consider two specifications where we randomly censor log income – first, we censor 50% of the values to the median, and second, we censor 50% of the values to 0. We consider two levels of top-coding – bound censoring the top 25% of the values, and the top 50% of the values.

The OLS estimates are displayed in Table 3 (with standard errors in parentheses). The base estimates display significant excess sensitivity with regard to current income, with the estimate $\hat{\beta} = .0646$. There is relatively little change in the propensities to consume out of wealth across the censoring scenarios, which is consistent with the relatively low correlations of .19 between $\Delta \ln W_i$ and $\ln Y_i$ and of .24 between $\Delta \ln YP_i$ and $\ln Y_i$. By the same token, the estimates of excess sensitivity fall roughly in line with single regressor results. That is, random censoring to the median coincides with relatively little bias, as expected, and random censoring to zero has severe bias, with a tiny estimate of the wrong sign. Top coding gives rise to larger estimates of excess sensitivity than the base, which are in line with expansion bias.

When there are higher correlations between the censored regressor and the other regressors, the bias can manifest in all coefficients. For comparison, in levels the correlation between $\ln W$

	$\ln W$		$\ln YP$		$\ln Y$	
Base Estimates	0.050	(0.0045)	0.2188	(0.0127)	0.1829	(0.0101)
Random Censoring						
to median	0.0557	(0.0046)	0.3429	(0.0094)	0.1162	(0.0102)
to zero	0.0587	(0.0046)	0.3973	(0.0082)	0.0002	(0.0010)
Top Coding						
25%	0.0611	(0.0045)	0.2788	(0.0116)	0.1541	(0.0108)
50%	0.0621	(0.0046)	0.3303	(0.0104)	0.1222	(0.0118)

Table 4: OLS Estimates of the Consumption Model in Levels

and $\ln Y$ is .38 and the correlation between $\ln YP$ and $\ln Y$ is .82 so we would expect different impacts from the censoring. That is exactly what we find if we estimate the log consumption equation in levels (with dependent variable $\ln c_i$), with results displayed in Table 4. Here the base estimates are different than the first-differenced equation, with a smaller financial wealth effect and larger permanent income and excess sensitivity estimates. Censoring of current income now manifests largely in the impact of permanent income. That is, random censoring results to the median results in a smaller current income effect but a larger permanent income effect and a somewhat larger wealth effect. Those differences are increased when random censoring is to zero. Here top-coding results in a smaller current income effect, but again larger permanent income and wealth effects. Here, it is pretty clear that $\ln YP$, and to some extent $\ln W$, is taking the place of $\ln Y$ when the values of the latter are censored.

To illustrate the impact of censoring on IV estimators, we consider the possibility that the change in consumption is determined jointly with current income, making current income endogenous. We computed IV estimates of the first-differenced consumption equation using lagged log- income as instrument to identify the current log income effect. The IV estimates are displayed in Table 5. Here the base estimates display negative excess sensitivity with estimate $\hat{\beta} = -.0324$, and again, there is not much variation in the estimated propensities to consume out of wealth across all the censoring scenarios. Here, each of the estimates with random censoring is larger in absolute value as expected – the single regressor results imply expansion bias and insensitivity to the censoring value, in contrast to the OLS results in Table 3. Top-coding gives rise to larger absolute values that again depend on the extent of the amount of censoring, as

	$\Delta \ln W$		$\Delta \ln YP$		$\ln Y$	
Base Estimates	0.0376	(0.0060)	0.1290	(0.0106)	-0.0324	(0.0097)
Random Censoring						
to median	0.0375	(0.0060)	0.1275	(0.0105)	-0.0660	(0.0199)
to zero	0.0341	(0.0074)	0.1323	(0.0138)	-0.0503	(0.0189)
Top Coding						
25%	0.0374	(0.0060)	0.1304	(0.0108)	-0.0440	(0.0133)
50%	0.0373	(0.0060)	0.1323	(0.0111)	-0.0646	(0.0195)

Table 5: IV Estimates of the Consumption Model

expected.

4. Censoring a Regressor to a 0-1 Variable

Our final topic is to consider the implications of replacing a continuous regressor with a dummy (0-1) variable indicating low and high values. This is a severe form of two-value censoring, where almost all of the information of the original continuous variable is lost. Nevertheless, empirical practice is replete with examples of the use of dummy variables where an underlying continuous variable may be more appropriate. We wish to open this area of inquiry with a few general points in line with the ones we have raised above.

4.1. Single Regressor Case

Begin with the original framework (5). With the true model

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad i = 1, ..., n, \tag{40}$$

we are interested in the results from estimating

$$y_i = a + bD_i + u_i \quad i = 1, ..., n. \tag{41}$$

where D_i is a dummy variable indicating whether x_i exceeds a threshold ξ,

$$D_i = 1\left[x_i > \xi\right]. \tag{42}$$

Obviously, this represents two-value censoring (0, 1), and all information about x_i is lost except whether it above the threshold or not. We denote $\Pr\{D = 1\} \equiv P$.

From the practical point of view it is clear that b and β are not the same. However, in practice, it is common to interpret one coefficient as a measure of the other. In our example about returns to years of schooling, the issue is how the return β is related to the "college effect" measured by estimating (41). For the OLS estimators, we have the expressions

$$\begin{gathered} \text{plim}\ \hat{a} = E\left(y|D=0\right) = \alpha + \beta \cdot E\left(x|D=0\right) \\ \text{plim}\ \hat{b} = E\left(y|D=1\right) - E\left(y|D=0\right) = \beta \cdot \left[E\left(x|D=1\right) - E\left(x|D=0\right)\right] \end{gathered} \tag{43}$$

The OLS slope coefficient $\hat{b}$ measures β up to a positive scale. So, if the only question of interest concerns the sign of β, then estimation with 0-1 censoring allows a consistent answer to that question. Any further interpretation of the value of $\hat{b}$ depends on the between-difference $E\left(x|D=1\right) - E\left(x|D=0\right)$, which is an unknown aspect of the distribution of the uncensored variable x.

The same sort of bias arises for IV estimators. Suppose that z_i is a valid instrument for x_i in (40) and that $\hat{c}$ is the IV estimator using z_i to instrument D_i in (41). It is straightforward to show that

$$\text{plim}\ \hat{c} = \beta \cdot \left\{\left[E\left(x|D=1\right) - E\left(x|D=0\right)\right] + \frac{\left(1-P\right)Cov\left(x,z|D=0\right) + PCov\left(x,z|D=1\right)}{P\left(1-P\right)\left[E\left(z|D=1\right) - E\left(z|D=0\right)\right]}\right\}. \tag{44}$$

There is an additional bias term that depends on the within interaction of the instrument z_i with x_i. If z_i and x_i are positively correlated, then it is natural to expect that the final covariance term is positive (higher x_i broadly associated with higher z_i values), in which case the IV estimator would estimate a term with same sign as β. However, it is easy to construct examples where the covariance term of (44) is negative and outweighs the first term, so that $\hat{c}$ estimates an effect of the wrong sign. In such cases the censoring of x_i to D_i gives a completely wrong depiction of

the relation between y_i and x_i.

4.2. Multivariate Case

Given the extreme nature of 0-1 censoring, one might expect the biggest bias issues to arise when there are additional regressors in the equation. If an additional regressor w_i is correlated with x_i, then the censoring will cause w_i to proxy x_i. The resulting transmission of bias is likely to be more extreme than in the cases studied earlier (since x_i was observed for a positive fraction of the data). This will contaminate the coefficient of w_i, and will likely affect bias in the coefficient of D_i as well.

Now the true model is (30), reproduced as

$$y_i = \alpha + \beta x_i + \phi w_i + \varepsilon_i \quad i = 1, ..., n. \tag{45}$$

and we are interested in the OLS estimates $\hat{a}$, $\hat{b}$, $\hat{f}$ of the coefficients of

$$y_i = a + bD_i + fw_i + u_i \quad i = 1, ..., n. \tag{46}$$

where D_i is observed instead of x_i.

To develop the bias in this case, denote the residual of y_i regressed on D_i as

$$\Delta y_i = y_i - (1 - D_i)\,\bar{y}_0 - D_i\bar{y}_1$$

where $\bar{y}_1 = \sum_{i=1}^{n} D_i y_i / \sum_{i=1}^{n} D_i$ is the average of y_i for $D = 1$ and $\bar{y}_0 = \sum_{i=1}^{n} (1 - D_i)\, y_i / \sum_{i=1}^{n} (1 - D_i)$, for $D = 0$. Now, sweep D_i from both sides of the true model (45) to give

$$\Delta y_i = \beta \Delta x_i + \phi \Delta w_i + \Delta \varepsilon_i \quad i = 1, ..., n. \tag{47}$$

The bias in $\hat{f}$ of (46) is the same as the bias from omitting Δx_i from (47), or estimating $\hat{f}$ from

$$\Delta y_i = f \Delta w_i + v_i \quad i = 1, ..., n. \tag{48}$$

From the standard omitted variable bias formula, we have

$$\text{plim } \hat{f} = \phi + \beta\eta \equiv f \tag{49}$$

where

$$\eta = \frac{Cov\left(\Delta w, \Delta x_i\right)}{Var\left(\Delta w\right)} = \frac{\left(1-P\right)Cov\left(x, w|D=0\right) + PCov\left(x, w|D=1\right)}{\left(1-P\right)Var\left(w|D=0\right) + PVar\left(w|D=1\right)} \tag{50}$$

The parameter η gauges how the within-deviations of x_i are proxied by the within-deviations of the other regressor w_i. If the within-deviations of x are closely proxied by those of w_i, say with $\eta \cong 1$, then w_i's estimated coefficient will reflect both the true effect ϕ as well as β, the effect of x_i. This verifies the intuition about proxying outlined above. There is no bias in $\hat{f}$ only if the within-covariances are zero (or net to zero with $\eta = 0$). If one has no information regarding within-variation of x_i, it is impossible to assess or disentangle the bias.

This bias will affect the other parameters as well. For the OLS intercept of (46), we have

$$\begin{aligned} \text{plim } \hat{a} &= E\left(y|D=0\right) - f \cdot E\left(w|D=0\right) \\ &= \alpha + \beta \cdot E\left(x|D=0\right) + \left(\phi - f\right)E\left(w|D=0\right) \\ &= \alpha + \beta \cdot \left[E\left(x|D=0\right) + \eta E\left(w|D=0\right)\right] \end{aligned} \tag{51}$$

and for the coefficient of D_i, we have.

$$\begin{aligned} \text{plim } \hat{b} &= E\left(y|D=1\right) - E\left(y|D=0\right) - f \cdot \left[E\left(w|D=1\right) - E\left(w|D=0\right)\right] \\ &= \beta \cdot \left[E\left(x|D=1\right) - E\left(x|D=0\right) - \eta \cdot \left\{E\left(w|D=1\right) - E\left(w|D=0\right)\right\}\right] \end{aligned} \tag{52}$$

Focusing on $\hat{b}$, the multivariate bias differs from the single regressor bias by a term that depends on how the additional regressor varies with the censoring. If $\eta \neq 0$, the difference vanishes only if w_i is mean-independent of D_i.

It is possible for $\hat{b}$ to be so severely biased as to be systematically of the wrong sign. η is determined by the covariation of x and w within the $D = 1$ and $D = 0$ groups; and is unaffected by the position of the group means. Therefore, if

$$E\left(w|D=1\right) - E\left(w|D=0\right) > \left(1/\eta\right) \cdot \left[E\left(x|D=1\right) - E\left(x|D=0\right)\right] \tag{53}$$

Bias of:		Correlation				
$\Pr\{D=1\}$		-50%	0%	50%	75%	95%
10%	$\widehat{b}$	60.0%	95.9%	60.0%	5.4%	–72.1%
	$\hat{f}$	-35.9%	0.0%	35.9%	61.0%	90.2%
20%	$\widehat{b}$	49.3%	75.0%	49.3%	5.3%	-70.1%
	$\hat{f}$	-29.2%	0.0%	29.2%	52.8%	87.0%
40%	$\widehat{b}$	42.9%	61.0%	42.9%	8.1%	-65.1%
	$\hat{f}$	-22.4%	0.0%	22.4%	43.9%	82.3%
50%	$\widehat{b}$	42.1%	60.1%	42.1%	9.0%	-63.9%
	$\hat{f}$	-21.7%	0.0%	21.7%	42.4%	81.2%

Table 6: Coefficient Biases: One 0-1 Regressor Censored from Normal.

then $\hat{b}$ consistently measures a value that is the opposite sign of the coefficient β. In any event, knowledge of the joint distribution of x_i and w_i is required to assess or understand the impact of the 0-1 censoring on the full regression.

We conclude this section with some bias calculations. Table 6 presents the results of 0-1 censoring, where x and w are assumed to be joint normal as in Section 3.2, and censoring is defined as in (42). The biases are computed as the difference between the coefficient limits and the true values of the coefficients for uncensored regressors (each value is 1.0). There is a very clear pattern of bias in $\hat{f}$. Namely, there is no bias only with 0 correlation, and increasing bias (transmission) with increasing correlation. The scale of the bias in $\hat{b}$ is not immediately interpretable, but we see the value of its limit is highest with 0 correlation, and then decreases sharply as the bias in $\hat{f}$ increases. It is not clear why this phenomena is slightly less severe for a balanced design ($\Pr\{D=1\}=.5$), but it is still very pronounced in that case.

Table 6 took the normal design from before. In order to illustrate sign changes, in Table 7 we present the results where we have adjusted the second regressor w_i by adding $3D_i$. This shifts the mean, adding 3 to the mean $E(w|D=1)-E(w|D=0)$ without changing the within covariance between x and w. In view of (53), this shift will increase the bias in $\hat{b}$. In fact, for the higher correlation values, $\hat{b}$ estimates a value that is the opposite sign of β. In this case, the bias transmission is so severe that the impact of both regressors is over attributed to w through $\hat{f}$, with $\hat{b}$ is the wrong sign to accommodate this erroneous attribution.

Bias of:		Correlation				
$\Pr\{D=1\}$		-50%	0%	50%	75%	95%
10%	$\widehat{b}$	167.8%	94.9%	-49.5%	-179.5%*	–343.2%*
	$\hat{f}$	-35.9%	0.0%	-35.9%	61.0%	90.2%
20%	$\widehat{b}$	136.1%	75.0%	-38.8%	-152.7%*	-331.6%*
	$\hat{f}$	-29.2%	0.0%	29.2%	52.8%	87.0%
40%	$\widehat{b}$	110.3%	61.0%	-24.1%	-122.7%*	-309.9%*
	$\hat{f}$	-22.4%	0.0%	22.4%	43.9%	82.3%
50%	$\widehat{b}$	107.6%	60.1%	-22.6%	-118.3%*	-305.7*
	$\hat{f}$	-21.7%	0.0%	21.7%	42.4%	81.2%
		* Coefficient Negative (Bias Produces Wrong Sign)				

Table 7: Coefficient Biases: One 0-1 Regressor Censored from Normal, with Mean Shift

5. Conclusion

This paper has shown many results on bias that arises with censored regressors. Our intention was to provide a rich depiction of the kinds of issues that censored regressors can bring to empirical work. While there are certain situations where coefficient estimates are too small (such as attenuation bias in OLS estimates with independent censoring), our view is that the more common situation is that effects are too large, such as the many cases of expansion bias noted above. Part of this view is the intuition that censoring involves eliminating important variation in the regressors, so that estimated effects will overcompensate. But this intuition is clearly too simplistic, as the nature and sign of the bias depends on both the censoring process and the censoring value. Hence, it is important to study each case separately.

Another lesson concerns the transmission of bias with several regressors. There is no surprise in the finding that censoring bias affects correlated regressors, but rather that the impacts can be huge. Using a dummy (0-1) variable as a control in place of a continuous variable, when the true model depends on the continuous variable, can result in biases of 50-100% in the coefficients of the other regressors. Using a top-coded version of a continuous variable can have a similar impact. The common practice of using 0-1 variable generates this problem in a severe form. Almost all of the variation (all of the within variation) is censored away, and will be proxied by any other correlated variables.

This paper has been mostly about the problems caused by censored regressors, not solutions to those problems. In general, flexible methods of estimating the bias in an empirical analysis do

not exist. One can adopt a parametric model of the true data and censored data, but estimates will be based on distributional assumptions made in that model. Without any restrictions, the censored data have zero semiparametric information, so some additional structure must be assumed to make use of the censored data. These issues are discussed in Rigobon and Stoker (2007), who also include a normal parametric model applied to the analysis of household consumption and wealth.

When exogenous censoring holds, consistent estimation is possible using the complete cases alone. This allows the assessment of bias, by comparing estimates using the complete cases with estimates using the full sample (see Rigobon and Stoker (2005) for a test of this type). Comparison of this type would seem to be a prudent empirical step whenever any of the regressors of interest are censored; that is to see whether the censored character of the data has changed the basic results. But even this is not justified with general endogenous and censored regressors. That is, the complete cases likely involve selection of the dependent variable, and instruments available in the uncensored sample will not be valid in the complete cases. The endogenous case is very difficult, for which some preliminary results are presented in Chernozhukov, Rigobon and Stoker (2007). Finally, with 0-1 censoring, the true value of the regressor is (almost) never observed, so such data has no complete cases.

References

[1] Ai, C. (1997), "An Improved Estimator for Models with Randomly Missing Data," *Nonparametric Statistics*, 7, 331-347.

[2] Black, D.A., M.C. Berger and F.A. Scott (2000), "Bounding Parameter Estimates with Nonclassical Measurement Error," *Journal of the American Statistical Association*, 95, 739-748.

[3] Chen, X., H. Hong and E. Tamer (2005), "Measurement Error Models with Auxiliary Data," *Review of Economic Studies*, 72, 343-366.

[4] Chen, X., H. Hong and A. Tarossa (2004), "Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects," Working Paper, November.

[5] Chernozhukov, V., R. Rigobon and T.M. Stoker (2007), "Set Identification with Tobin Regressors," MIT Working Paper, October.

[6] Davidson, R. and J. D. McKinnon (2004), *Econometric Theory and Methods*, Oxford University Press, New York.

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

[8] Heitjan, D. F. and D. B. Rubin (1990), "Inference from Coarse Data Via Multiple Imputation With Application to Age Heaping," *Journal of the American Statistical Association*, 85, 304-314.

[9] Heitjan, D. F. and D. B. Rubin (1991), "Ignorability and Coarse Data," *Annals of Statistics*, 19, 2244-2253.

[10] Hong, H and E. Tamer (2003), "Inference in Censored Models with Endogenous Regressors," *Econometrica*, 71, 905-932.

[11] Horowitz, J. and C. F. Manski (1995), "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63, 281-302.

[12] Horowitz, J. and C. F. Manski (1998), "Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations," *Journal of Econometrics*, 84, 37-58.

[13] Horowitz, J. and C. F. Manski (2000), "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95, 77-84.

[14] Klepper, S. and E.E. Leamer, (1984), "Consistent Sets of Estimates for Regressions with Errors in All Variables," *Econometrica* 52, 163-184.

[15] Lewbel, A. (2000), "Semiparametric Qualitative Response Model Estimation with Unknown Heteroskedasticity or Instrumental Variables," *Journal of Econometrics* 97, 145–177.

[16] Liang, H, S. Wang, J.M. Robins and R.J. Carroll (2004), "Estimation in Partially Linear Models with Missing Covariates," *Journal of the American Statistical Association*, 99, 357-367.

[17] Little, R. J. A. (1992), "Regression with Missing X's: A Review," *Journal of the American Statistical Association*, 87, 1227-1237.

[18] Little, R. J. A. and D. B. Rubin (2002), *Statistical Analysis with Missing Data, 2nd edition*, John Wiley and Sons, Hoboken, New Jersey.

[19] Magnac, T. and E. Maurin, (2004). "Partial Identification in Monotone Binary Models: Discrete and Interval Valued Regressors." Working Paper, CREST, no. 2004-11.

[20] Magnac, T. and E. Maurin, (2007). "Identification and Information in Monotone Binary Models: Discrete and Interval Valued Regressors." *Journal of Econometrics* 139, 76-104.

[21] Mahajan, A. (2006), "Identification and Estimation of Regression Models with Misclassification," *Econometrica*, 74, 667-680.

[22] Manski, C.F. and E. Tamer (2002) "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70, 519-546.

[23] Ridder, G. and R. Moffit (2003), "The Econometrics of Data Combination," chapter for *Handbook of Econometrics, Volume 6*, forthcoming.

[24] Rigobon, R and T. M. Stoker (2005) "Testing for Bias from Censored Regressors", MIT Working Paper, June.

[25] Rigobon, R. and T. M. Stoker (2007), "Estimation with Censored Regressors: Basic Issues," forthcoming *International Economic Review*.

[26] Tripathi, G. (2003), "GMM and Empirical Likelihood with Imcomplete Data," Working Paper, December.

[27] Tripathi, G. (2004), "Moment Based Inference with Incomplete Data," Working Paper, June.

[28] Wald, A. (1940), "The Fitting of Straight Lines if Both Variables are Subject to Error," *Annals of Mathematical Statistics*, 11, 284-300.

A. Appendix: Notes on Proofs

For Proposition 1, the true model (5) is

$$y_i = \alpha + \beta x_i^{cen} + \beta x_i^o + \varepsilon_i \quad i = 1, ..., n \tag{54}$$

where $x_i^o = d_i (x_i - \xi)$ is the omitted term. Since x_i^{cen} is uncorrelated with ε_i, we have

$$\text{plim}\, \hat{b} = \beta \cdot \left(1 + \frac{Cov(x^o, x^{cen})}{Var(x^{cen})}\right) \tag{55}$$

We have:

$$\begin{aligned} Cov(x^o, x^{cen}) &= E(x^o \cdot x^{cen}) - E(x^o) E(x^{cen}) \\ &= E(\xi d(x - \xi)) - E(x^o) E(x^{cen}) \\ &= E(x^o) \cdot [\xi - E(x^{cen})] \\ &= p[E(x|d=1) - \xi] \cdot (1-p)[\xi - E(x|d=0)] \end{aligned} \tag{56}$$

Substituting this into (55) gives Proposition 1. Corollary 2 and Corollary 3, parts 1. and 2.are immediate, and since $Cov(x^2, d) = 0$, it is easy to derive $Var(x^{cen})$ as used in (10) . For Corollary 3, part 3., we note that the numerator of the bias is the sum of the analogous numerator terms for 1. and 2., and that the denominator $Var(x^{cen})$ is smaller than the analogous denominators in 1. and 2., so the strict inequality holds.

Proposition 4 is also a direct calculation that follows from

$$Cov(z, x^{cen}) = (1-p) Cov(z, x|d=0) + p(1-p) \Delta_z [\xi - E(x|d=0)], \tag{57}$$

$$Cov(z, x^o) = pCov(z, x|d=1) + p(1-p) \Delta_z [E(x|d=1) - \xi] \tag{58}$$

The formulae (57) and (58) follow from straightforward arithmetic. For instance, for (57), we have

$$\begin{aligned} Cov(z, x^{cen}) &= Cov(z, (1-d)x) + Cov(z, d) \cdot \xi \\ &= E((1-d)xz) - E(z) \cdot E((1-d)x) + p(1-p) \Delta_z \xi \end{aligned}$$

Write out all of the expectations in terms of expectations conditional on $d = 1$, and simplify to get (57).

Corollary 5 follows from noting that three things. First, $Cov(z, d) = 0$ implies $\Delta_z = 0$.

Second, $Cov(zx, d) = 0$ and $Cov(x, d) = 0$ imply that $E(zx|d = 1) = E(zx|d = 0)$ and $E(x|d = 1) = E(x|d = 0)$, so that $Cov(z, x|d = 1) = Cov(z, x|d = 0) = Cov(z, x)$. Finally, $Cov(z, x|d = 0) \neq 0$ since $Cov(z, x) \neq 0$.

Corollary 6 1.and 2. follow from (19). The final statement is true for top-coding since $\Delta_z > 0$ and $E(x|d = 1) - \xi > 0$, and for bottom coding since $\Delta_z < 0$ and $E(x|d = 1) - \xi < 0$.

Corollary 7 follows from three steps. First, note that $Cov(z, x) \neq 0$ implies that $\zeta(x_0) \equiv E(z|x = x_0)$ is not a constant function. Second,

$$\begin{aligned} Cov(z, d) &= p(1-p)[E(z|d=1) - E(z|d=0)] \\ &= p(1-p)[E(\zeta(x_1)|x_1 > \xi) - E(\zeta(x_0)|x_0 \leq \xi)] > 0 \end{aligned}$$

by mean-monotonicity. Finally,

$$\begin{aligned} Cov(z, x|d=1) &= E(zx|d=1) - E(z|d=1)E(x|d=1) \\ &= E(z \cdot [x - E(x|d=1)] \mid d=1) \\ &= E(\zeta(x) \cdot [x - E(x|x > \xi)] \mid x > \xi) \\ &= E\{[\zeta(x) - \zeta[E(x|x > \xi)]] \cdot [x - E(x|x > \xi)] \mid x > \xi\} \\ &\geq 0 \end{aligned}$$

by mean-monotonicity, since the final expectation is an integral of non-negative terms.

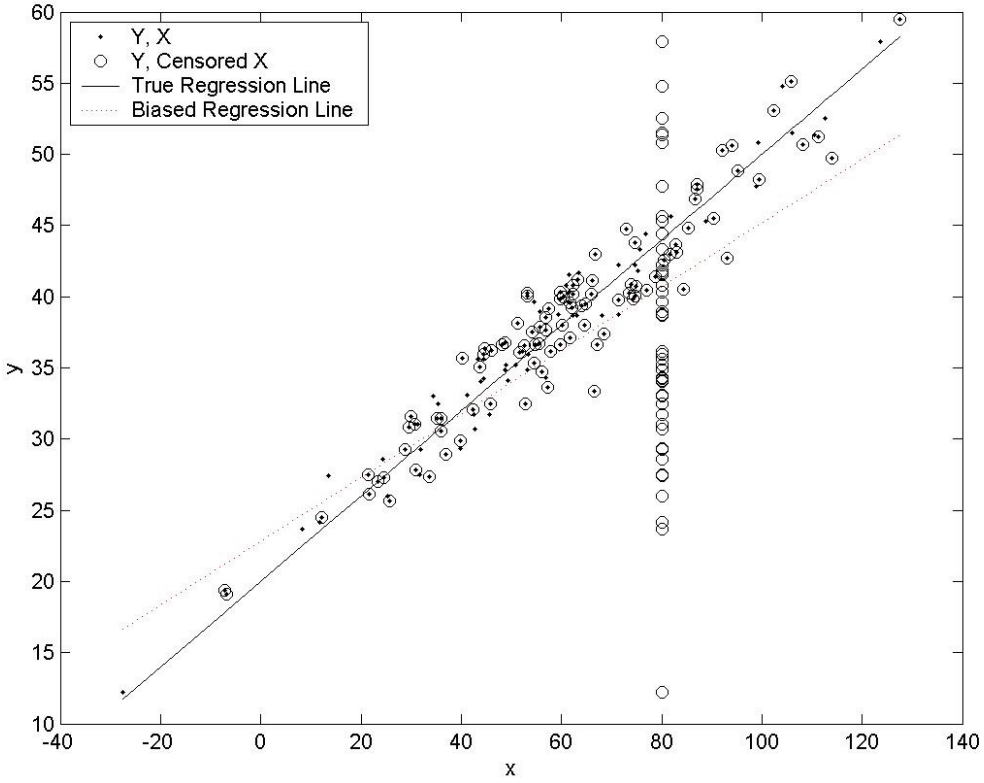

Figure 1: OLS: Independent Censoring and Attenuation Bias

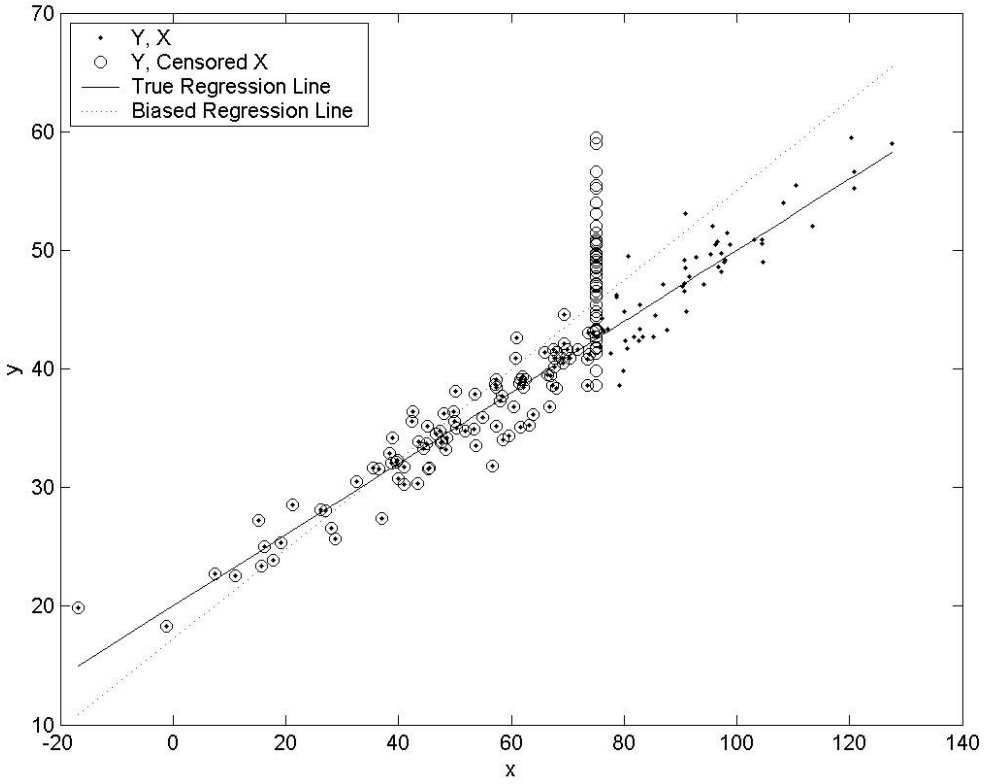

Figure 2: OLS: Top-Coding and Expansion Bias

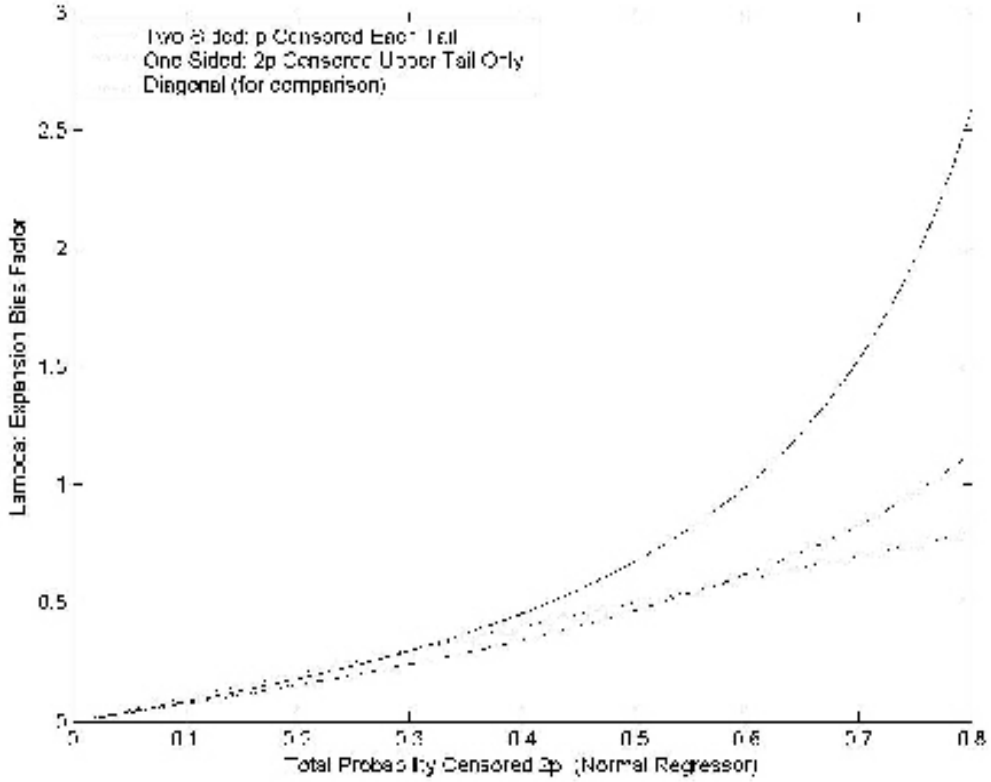

Figure 3: Bias from Bound Censoring with a Normal Regressor

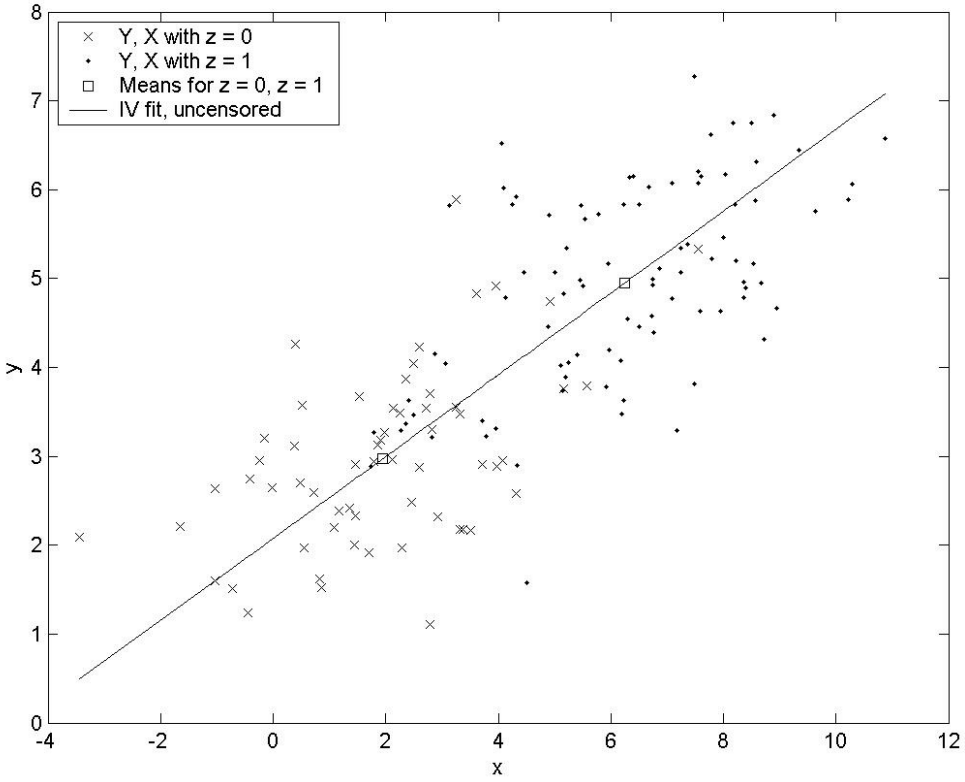

Figure 4: IV: Data with Random Assignment

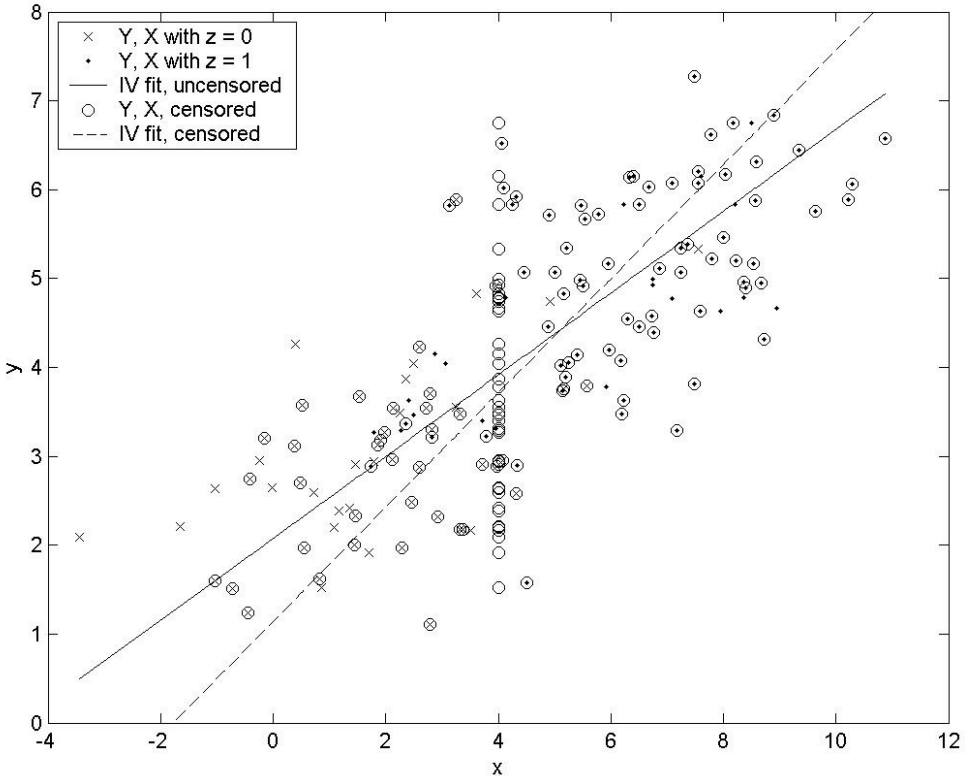

Figure 5: IV: Independent Censoring and Expansion Bias

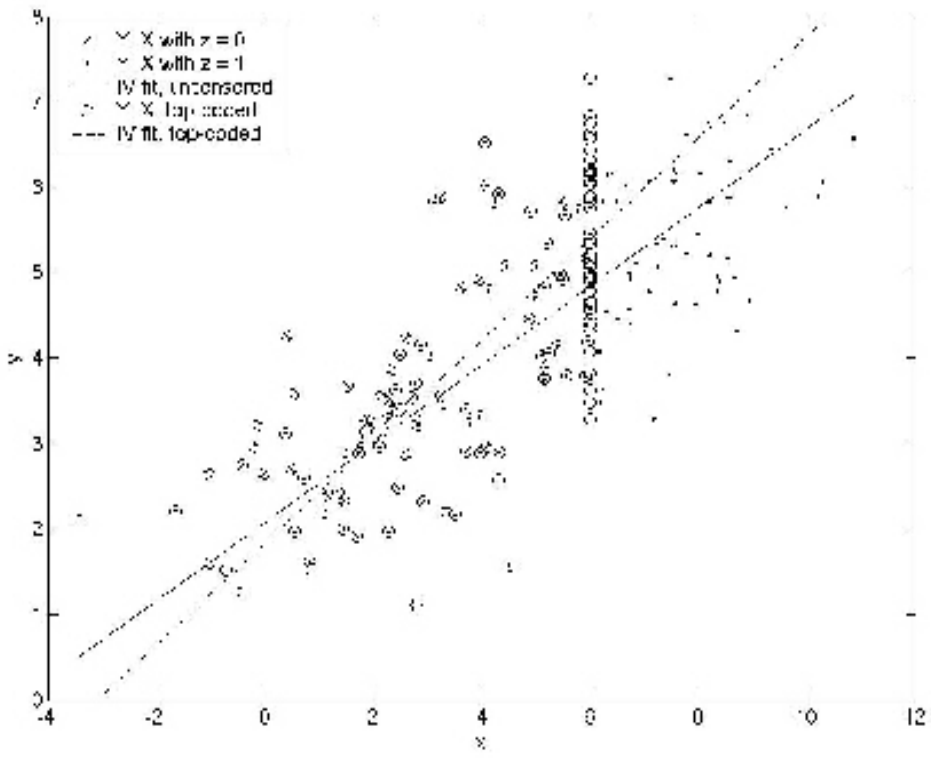

Figure 6: IV: Top-Coding and Expansion Bias