Censored Regressors and Expansion Bias

Roberto Rigobon Thomas M. Stoker*

March 2005

Abstract

In this paper we study the bias that arises from using censored regressors in regression analysis. One focus is on censoring to bounds (top-coding or bottom-coding) which give rise to *expansion bias*, or coefficient estimates that are proportionally too large. We give a simple but general formula for bias from using a censored regressor in bivariate regression, and we derive approximate formulae for bias from top-coding with many regressors. We discuss several aspects of consistent estimation with censored regressors, including showing the necessity of certain restrictions for the censored data to offer any information at all on the regression coefficients. We propose a model of mixed censoring, which captures both censoring to bounds and random censoring, The concepts are illustrated by showing how censored regressors arise in the analysis of wealth effects on consumption, including application to household consumption data.

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 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 bias that arises from censoring of an independent variable, or regressor, in the estimation of a linear

* Sloan School of Management, MIT, 50 Memorial Drive, Cambridge, MA 02142 USA. We thank Jonathan Parker for his generosity in sharing with us not only his data but all the programs used to reproduce the results in his NBER Macro Annual paper. We have received valuable comments from James Heckman, Vincent Hogan, Dale Jorgenson, Richard Blundell, Charles Manski, Orazio Attanasio, James Banks, Costas Meghir, Alberto Abadie, Jerry Hausman, Gary Chamberlain, Peter Bickel, Jin Hahn, Elie Tamer and especially Whitney Newey.

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." In this paper we study the bias that arises from using censored regressors, with a focus on censoring to bounds. For instance, what difference does it make if we estimate a regression with top-coded income data when the correct specification has income (no top coding) as the appropriate regressor?.

We show that using a top-coded regressor results in *expansion bias* in OLS estimates of regression coefficients; namely, estimated effects are too large in absolute value.[1] For instance, if income is top-coded, a positive income effect will be overestimated. Expansion bias also arises with the use of bottom-coded data, or data that is both top-coded and bottom coded.

Figure 1 shows a scatterplot of data with and without top- and bottom-coding 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. This is what we call *expansion bias*, and it arises because of the "pile-up" of observations at each limit. Moreover, it is clear that expansion bias would result if there was only top-coded censoring (or only bottom-coding); and that each censoring limit contributes to the extent of expansion bias.

For bivariate regression, we show a simple but general expression for the bias that arises from using censored regressors, that specializes to expansion bias with top-coding or bottom coding. 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 results for the case when all regressors are normally distributed. We also compare attenuation bias from errors-in-variables to expansion bias, for intuition on the relative size of these problems.

We consider several aspects of model estimation with top-coding and other censoring. For the censoring problems we consider, consistent estimation can be done by dropping all censored observations, namely including only complete cases. We discuss the efficiency aspects of using

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

some of the censored data, beginning with what the potential efficiency gains are. We argue that the common practice of including a dummy variable for censoring is not advisable – there will be no bias only under strong restrictions and otherwise, no information is gained from including the censored observations. We show broadly that if there are no restrictions on the distribution of the censored variable, for estimating the parameters of interest, there is zero semiparametric information available from the censored data. We discuss parametric modeling of censoring, including specifying and estimating a model of mixed censoring.

We discuss how censored regressors arise in the estimation of wealth effects on consumption, including an empirical application. Our discussion is intended to give concrete illustration to the ideas, and we plan to carry out further applications as part of future research.

There is relatively little literature on censored regressors in econometrics. An exception is Manski and Tamer (2000), who study identification and consistent estimation with interval data. The statistical literature on missing data problems covers some situations of censored regressors, with most results applicable to data missing at random. See the survey by Little (1992) and Little and Rubin (2002) for coverage of this large literature, and Ridder and Moffit (2003) for survey of the related literature on data combination.[2] Top-coding and bottom-coding are nonignorable data coarsenings in the sense of Heitjan and Rubin (1990, 1991).[3] For related discussions on information and efficiency, see Horowitz and Manski (1998, 2000), Robins and Rotnitzky (1995) and Rotnitzky, Robins and Scharfstein (1998).

Bias from censored regressors is a straightforward problem, but ignoring that problem can lead to substantial mismeasurement of effects. Expansion bias implies that estimated effects are artificially too large, which can give a misleading picture of what influences are important. 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 to wealth effects on consumption, and Section 5 gives some concluding remarks.

[2] Recent contributions include Chen, Hong and Tamer (2005), Chen, Hong and Tarossa (2004), Liang, Wang, Robins and Carroll (2004), Tripathi (2003,2004), Mahajan (2004) and Ichimura and Martinez-Sanchis (2005).

[3] Some recent work has shown how data heaping in duration data (censoring or rounding due to memory effects) data can be accomodated in estimation of survival models. See Torelli and Trivellato (1993) and Petoussis, Gill and Zeelenberg (2004), among others.

2. Bias with Censored Regressors

2.1. Censoring in Bivariate Regression

We start with bivariate regression and develop a simple bias formula. 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 we have a general censoring process described by a censoring indicator d_i, where d_i is uncorrelated with ε_i conditional on x_i. Denote the probability of censoring as $p = \Pr\{d = 1\}$. Suppose that censored regressor values are set to ξ. That is, x_i^{cen} is the censored version of x_i, where

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

We are interested in what happens when we use x_i^{cen} to estimate β; namely if we estimate the OLS coefficient $\hat{b}$ with the model

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

The answer is given as Proposition 1:

Proposition 1. *Suppose . Provided that* $0 < p < 1$*, we have*

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

where

$$\Lambda^o = p (1 - p) \cdot \frac{(E(x|d = 1) - \xi)(\xi - E(x|d = 0))}{Var(x^{cen})}. \tag{5}$$

Consequently, the bias term

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

2. $\Lambda^o > 0$ *if*

$$E(x|d=1) < \xi < E\left(x|d=0\right) \quad or \quad E(x|d=1) > \xi > E(x|d=0),$$

otherwise $\Lambda^o < 0$.

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

Three principle examples are covered in the following Corollaries:

Corollary 2. ***Random Censoring.*** *Suppose* d_i *is distributed independently of* x_i. *Then*

$$plim\ \hat{b} = \beta\left(1 + \Lambda^{ran}\right)$$

where $\Lambda^{ran} \leq 0$, *with equality holding only if* $\xi = E\left(x\right)$.

Corollary 3. ***Expansion Bias.*** *Suppose*

1. *Suppose the regressor is top-coded at* ξ^+, *with* $d_i = 1\left[x_i > \xi^+\right]$ *and*

$$x_i^{cen} = x_i \cdot 1\left[x_i \leq \xi^+\right] + \xi^+ \cdot 1\left[\xi^+ < x_i\right] \tag{6}$$

then

$$plim\ \hat{b} = \beta\left(1 + \Lambda^+\right) \tag{7}$$

where $\Lambda^+ > 0$.

2. *Suppose the regressor is bottom-coded at* ξ^-, *with* $d_i = 1\left[x_i < \xi^-\right]$ *and*

$$x_i^{cen} = x_i \cdot 1\left[x_i \geq \xi^-\right] + \xi^- \cdot 1\left[x_i < \xi^-\right] \tag{8}$$

then

$$plim\ \hat{b} = \beta\left(1 + \Lambda^-\right) \tag{9}$$

where $\Lambda^- > 0$.

3. Suppose the regressor is top-coded at ξ^+ and bottom-coded at ξ^-, with $\xi^- < \xi^+$, and

$$x_i^{cen} = x_i \cdot 1\left[\xi^- \leq x_i \leq \xi^+\right] + \xi^- \cdot 1\left[x_i < \xi^-\right] + \xi^+ \cdot 1\left[\xi^+ < x_i\right]. \tag{10}$$

Then

$$plim\ \hat{b} = \beta\left(1 + \Lambda\right) \tag{11}$$

where $\Lambda > \Lambda^+ + \Lambda^-$ above.

Proposition 1 and the Corollaries highlight the importance of considering the censoring indicator d_i and the censoring value ξ separately. The bias in each case depends on the position of the censoring value ξ relative to the means of the censored and the uncensored data.

In the parlance of the missing data literature (c.f. Little and Rubin (2002)), our notion of random censoring is analogous to "missing completely at random," or MCAR, and bias arises if the censoring value ξ is not the mean of the regressor. Top-coding and bottom-coding involve censoring determined by the value of the regressor, so that they are analogous to "not missing at random" processes, or NMAR, where in addition, the censoring threshold is given by the censoring value ξ.

For the remainder of this section, we focus on censored regressors that are top-coded or bottom coded. After the proofs, we illustrate the size of the expansion bias when the uncensored regressor is normally distributed.

2.1.1. Proof of Proposition 1 and the Corollaries

For Proposition 1, the true model (1) is

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

where

$$x_i^o = d_i\left(x_i - \xi\right) \tag{13}$$

is the omitted term. Since x_i^{cen} is uncorrelated with ε_i, the omitted variable bias formula gives

$$\text{plim}\,\hat{b} = \beta \cdot \left(1 + \frac{Cov\left(x^o, x^{cen}\right)}{Var\left(x^{cen}\right)}\right) \tag{14}$$

We have:

$$\begin{aligned} Cov\left(x^o, x^{cen}\right) &= E\left(x^o \cdot x^{cen}\right) - E\left(x^o\right) E\left(x^{cen}\right) \\ &= E\left(\xi d\left(x - \xi\right)\right) - E\left(x^o\right) E\left(x^{cen}\right) \\ &= E\left(x^o\right) \cdot \left[\xi - E\left(x^{cen}\right)\right] \\ &= p\left[E(x|d=1) - \xi\right] \cdot (1-p)\left[\xi - E\left(x|d=0\right)\right] \end{aligned} \tag{15}$$

Substituting this into (14) gives Proposition 1. For Corollary 2, we have that $E(x|d=1) = E(x|d=0) = E\left(x\right)$, so (5) is negative unless $\xi = E\left(x\right)$. Corollary 3 follows from some straightforward observations: For 1., we have $E(x|d=1) > \xi^+ > E(x|d=0)$, so the bias is positive. For 2., we have $E(x|d=1) < \xi < E\left(x|d=0\right)$, so again the bias is positive. For 3., we note that the numerator of the bias is the sum of the analogous numerator terms in 1. and 2., and that the denominator $Var\left(x^{cen}\right)$ is smaller than the analogous denominators in 1. and 2., so the strict inequality holds.

2.1.2. Expansion Bias with a Normal Regressor

The formula for bias from top-coding and bottom-coding 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.

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 bias Λ of (11) from top- and bottom-coding under the assumption of symmetric (two-sided) censoring, with the same probability $p^- = p^+ \equiv p$ of censoring in the high and low region, and it is plotted against the total probability of censoring $2p$. The dashed line is the expansion bias Λ^+ of (7) from using top-coded data (one-sided censoring), which is plotted against the total 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 coefficient.

2.1.3. Interpretation via Errors-in-Variables

We close this section by comparing expansion bias with errors-in-variables bias in a certain way. On one level, one can ask whether the attenuation bias from, say, 20% measurement error is comparable to the expansion bias from 20% censoring. Alternatively, since those biases are in different directions, one could ask what level of censoring could be done on a variable to undo the attenuation bias from 20% measurement error. In any case, since errors-in-variables bias is very familiar in econometrics, there may be some interpretive value in a comparison to expansion bias.

Consider the bivariate model

$$y_i = \alpha + \beta w_i + \varepsilon_i, \quad i = 1, ..., n \tag{16}$$

where $x_i = w_i + \nu_i$, and ε_i, ν_i and w_i are mutually uncorrelated. If we estimate

$$y_i = a + b^* x_i + \eta_i \quad i = 1, ..., n, \tag{17}$$

then plim $\hat{b}^* = \beta(1 - \lambda)$, where $\lambda = Var(\nu)/Var(x)$. If we further censor x as above, and estimate

$$y_i = a + b x_i^{cen} + u_i \quad i = 1, ..., n, \tag{18}$$

then we have plim $\hat{b} = \beta(1 - \lambda)(1 + \Lambda)$. So, the bias balances if $(1 - \lambda)(1 + \Lambda) = 1$, or

$$\Lambda = \frac{\lambda}{1 - \lambda}. \tag{19}$$

Figure 3 compares the censoring level with error variance level when the regressor is normally distributed, showing the level of censoring required for the balance condition (19). 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. 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.[4]

2.2. Expansion Bias and Several Regressors

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

$$y_i = \alpha + \beta x_i + \gamma' z_i + \varepsilon_i \quad i = 1, ..., n \tag{20}$$

where we assume that the distribution of $\left(x_i, z_i', \varepsilon_i\right)$ is nonsingular, has finite second moments and obeys $E(\varepsilon_i | x_i, z_i) = 0$. We focus on the impact of top-coding, when x_i^{cen} of (6) is used

[4] While far afield from our topic, there is the possibility that some censoring could be advisable for variables that are known to be measured with error, especially in the tails – e.g. hours spent in traffic, gallons of beer consumed, etc.

instead of x_i. If we estimate the model

$$y_i = a + bx_i^{cen} + c^{'} z_i + u_i \quad i = 1, ..., n, \tag{21}$$

then how are the OLS coefficients $\hat{a},\ \hat{b},\ \hat{c}$ biased as estimators of $\alpha,\ \beta,\ \gamma$?

It is not difficult to develop the following expression of the bias:

$$B^+ \equiv \text{plim} \begin{pmatrix} \hat{a} \\ \hat{b} \\ \hat{c} \end{pmatrix} - \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \beta \cdot \Omega^{-1} \begin{pmatrix} p^+ \left(\mu_x^+ - \xi^+\right) \\ p^+ \xi^+ \left(\mu_x^+ - \xi^+\right) \\ p^+ \left(M_{xz}^+ - \mu_z^+ \xi^+\right) \end{pmatrix} \tag{22}$$

where '$^+$ 'indicates expectation over the censored region; namely $\mu_x^+ = E\left(x \mid x > \xi^+\right)$, $\mu_z^+ = E\left(z \mid x > \xi^+\right)$, $M_{xx}^+ = E\left(x^2 \mid x > \xi^+\right)$, $M_{xz}^+ = E\left(xz^{'} \mid x > \xi^+\right)$; and

$$\Omega \equiv E\left[\left(1, x^{cen}, z^{'}\right)^{'} \left(1, x^{cen}, z^{'}\right)\right]. \tag{23}$$

The bias (22) is easily interpreted in the 'no correlation' case. Suppose $\mu_z = 0$ and that z is *mean independent* of x. Then $\mu_z^+ = M_{xz}^+ = 0$, and Ω is block diagonal (partitioned according to $(1, x^{cen})$ and 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. It is straightforward to verify that

$$\text{plim}\ \hat{b} - \beta = \beta \cdot \frac{p^+ \left[\xi^+ - E\left(x^{cen}\right)\right] \left(\mu_x^+ - \xi^+\right)}{Var\left(x^{cen}\right)} \tag{24}$$

which simplifies to (4), (5) of Proposition 1.[5]

For the case with correlation, the exact bias formula (22) is too complex to admit clear interpretation. However, we can learn from expanding the bias in terms of the censoring probabilities, as follows. 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

[5] Exactly the same analysis applies to the case of bottom-coding and to the case of (two-sided) top- and bottom-coding. In terms of the intercept, top-coding involves a positive intercept bias whereas bottom-coding involves a negative intercept bias.

$$B^{+} \cong \beta \cdot \frac{p^{+}}{M_{zz}M_{xx} - (M_{xz})^{2}} \begin{pmatrix} \left(M_{zz}M_{xx} - (M_{xz})^{2}\right)\left(\mu_{x}^{+} - \xi^{+}\right) \\ M_{zz}\xi^{+}\left(\mu_{x}^{+} - \xi^{+}\right) - M_{xz}\left(M_{xz}^{+} - \xi^{+}\mu_{z}^{+}\right) \\ M_{xx}\left(M_{xz}^{+} - \xi^{+}\mu_{z}^{+}\right) - M_{xz}\xi^{+}\left(\mu_{x}^{+} - \xi^{+}\right) \end{pmatrix} \tag{25}$$

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 $\beta p^{+}\left(\mu_{x}^{+} - \xi^{+}\right)$, and so will be positive for top-coding. For $\hat{b}$ and $\hat{c}$, the biases are:

$$\text{plim } \hat{b} - \beta \;\propto\; \beta p^{+}\left\{\xi^{+}\left(\mu_{x}^{+} - \xi^{+}\right) - \rho_{xz}\frac{\sigma_{x}}{\sigma_{z}}\left[M_{xz}^{+} - \xi^{+}\mu_{z}^{+}\right]\right\} \tag{26}$$

and

$$\text{plim } \hat{c} - \gamma \;\propto\; \beta p^{+}\left\{\frac{\sigma_{x}}{\sigma_{z}}\left[M_{xz}^{+} - \xi^{+}\mu_{z}^{+}\right] - \rho_{xz}\xi^{+}\left(\mu_{x}^{+} - \xi^{+}\right)\right\} \tag{27}$$

where the positive proportionality constants are from (25).

To interpret these terms, suppose that x and z are positively correlated overall ($\rho_{xz} > 0$), positively correlated within the censored region (with 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 (26) consists of the positive expansion bias term of (24),[6] less a positive term due to z's role as a proxy for the censored values of x. Similar terms arise for $\hat{c}$ in (27) 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

[6] Since we drop higher order terms in p, we have $p\left(\xi^{+} - E\left(x^{cen}\right)\right) \simeq p\left(\xi^{+} - \mu_{x}\right) = p\xi^{+}$.

Truncation	Bias of:	Correlation -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.

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,[7] 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[8].

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 above. 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 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

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

[8] In line with (21), $\hat{a}$ 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.

above. In any case, we feel these results suggest that the biases introduced by the censoring of a regressor can be large and economically significant.

3. Estimation and Efficiency

Having established the bias with censoring, we now consider how to correct the bias, or how to estimate the parameters consistently. We consider three broad approaches: First is complete case (CC) analysis, where the censored data is dropped. Second is the use of partial assumptions, such as the empirical practice of including a dummy variable that indicates the censoring, as well as the question of how much information is contained in the censored data. Third is the use parametric methods, where the process of censoring is specified, as we apply in the next section. We make some general points, but leave the analysis of fully semiparametric estimation methods to future research.[9]

We begin with the model with a single additional regressor

$$y_i = \alpha + \beta x_i + \gamma z_i + \varepsilon_i \quad i = 1, ..., n \tag{28}$$

applying to the full (complete) sample, where the distribution of $(x_i, z_i, \varepsilon_i)$ is nonsingular with finite second moments. We have $E(\varepsilon_i | x_i, z_i) = 0$ and for simplicity, we assume homoskedasticity of ε_i,

$$Var(\varepsilon | x, z) = \sigma^2. \tag{29}$$

We will specialize to the cases of many additional regressors (vector z_i) or no additional regressor (no z_i) as necessary. We continue to focus on top-coding with where $d_i = 1\left[x_i > \xi^+\right]$, though most of our remarks apply to general censoring with d_i uncorrelated with ε_i conditional on x_i.

[9] It is interesting to realize that even though there is a vast set of tools to deal with censoring of the dependent variable, those methodologies cannot be used in the context of censored regressors, even when we consider the reverse regression. While the reverse regression has a censored dependent variable (x_i^{cen}), it is not well specified, because the regressor (y_i) is correlated with the error term, and part of that correlation is due to the censoring of x_i that we are studying here.

3.1. Complete Case Analysis and Efficiency

Suppose, just for notation, that the sample is ordered with the $n_0 = \sum_{i=1}^{n}(1-d_i)$ complete (uncensored) observations first, $i = 1, ..., n_0$, followed by the $n_1 = \sum_{i=1}^{n} d_i$ censored observations, $i = n_0 + 1, ...n_0 + n_1$. Because top-coding is censoring on the value of x_i, we can always get consistent estimates of the regression parameters by dropping all the censored data. Complete case (CC) analysis estimates

$$y_i = \alpha + \beta x_i + \gamma z_i + \varepsilon_i \quad i = 1, ..., n_0 \tag{30}$$

by OLS, obtaining $\hat{\alpha}_0$, $\hat{\beta}_0$, $\hat{\gamma}_0$ and $\hat{\sigma}^2$, estimates that are consistent for α, β, γ and σ^2, respectively.

Consider now the censored observations. The true model is

$$y_i = \alpha + \beta g_1(z_i) + \gamma z_i + u_i \quad i = n_0 + 1, ..., n_0 + n_1 \tag{31}$$

where

$$g_1(z_i) = E(x_i | z_i, d_i = 1) \tag{32}$$

The disturbance

$$u_i = \beta(x_i - g_1(z_i)) + \varepsilon_i \tag{33}$$

has $E(u_i|z_i) = 0$ and $\sigma_u^2(z_i) = Var(u_i|z_i) = \beta^2 Var(x_i|z_i) + \sigma^2$. In essence, since we don't observe x_i for the censored data, the best possible situation is where you know the value $g_1(z_i)$ of the conditional expectation for each i, and $\sigma_u^2(z_i)$ for each i. Then one could do an efficient pooled estimation of (30)-(31), estimating with the whole sample. For the i^{th} observation of the censored data, this amounts to using $g_1(z_i)$ in place of x_i, and weighting by $1 \Big/ \sqrt{\sigma_u^2(z_i)}$.

It is interesting to ask what efficiency loss these procedures entail relative to estimating with the full complete sample. For instance, if there were 20% censoring and censoring was random, then the efficiency loss of CC analysis would be 20%, or the relative efficiency of the coefficient estimates would be 80%. But top-coding is in no way random; it involves censoring some of the most influential observations.[10] CC analysis with 20% top-coding will involve a lower efficiency

[10] "Influential" is used here in the same sense as in the literature on regression diagnostics or experimental design: see Belsley, Kuh and Welsch (1980) among many others.

		Bivariate	One Additional Regressor			
	Procedure		Correlation .5		Correlation .9	
Truncation		Eff $\hat{\beta}$	Eff $\hat{\beta}$	Eff $\hat{\gamma}$	Eff $\hat{\beta}$	Eff $\hat{\gamma}$
20%	CC	47 %	52 %	80 %	71 %	80 %
	Known Mean	88 %	62 %	86 %	76 %	83 %
40%	CC	25 %	30 %	60 %	48 %	60 %
	Known Mean	78 %	46 %	71 %	58 %	65 %
60%	CC	13 %	15 %	40 %	28 %	40 %
	Known Mean	66 %	36 %	56 %	42 %	45 %

Table 2: Efficiency Relative to Complete Sample

than 80%, but how much lower? Moreover, how valuable is it to know the conditional means of the censored data? Is most of the efficiency loss of CC analysis eliminated?

Table 2 presents efficiencies for normally distributed regressors for different amounts of censoring.[11] The multivariate model is as above; with $g_1(z_i)$ and $\sigma_u^2(z_i)$ computed for the normal distribution (formulae presented in the Appendix). The bivariate model has no z_i, so that the conditional mean is $g_1 = E(x_i|d_i = 1)$ and $\sigma_u^2 = \beta^2 Var(x_i|d_i = 1) + \sigma^2$.

We see that for the bivariate model, there relative efficiency of CC analysis is much lower than it would be with random sampling: 47% with 20% top-coding, 25% with 40% top-coding, etc. A great deal of the efficiency loss is eliminated if the mean of the top-coded data is known. When there is an additional regressor, the efficiency loss in estimating β is less that in the bivariate case, and improves with higher correlation between x and z. When the conditional mean of the top-coded data is known, the efficiency improves for each coefficient, but not to the same extent as with the bivariate model. Finally, we notice that the improvements in efficiency for β and γ (from knowing the mean) are more balanced with higher correlation.[12]

3.2. Ineffectiveness of Dummy Variable Methods

A fairly common practice in empirical work is to include the censoring dummy d_i as a regressor, to account for top-coding. Here we make a couple simple remarks about this practice, which will lead us into our discussion of semiparametric methods.

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

[12] These calculations are done with optimal (GLS) weighting, but we did not find that the results were very sensitive to whether weighting was done or not.

The true model with censored regressor (for all data) is

$$y_i = \alpha + \beta x_i^{cen} + \gamma z_i + \beta \left(g_1 \left(z_i \right) - \xi^+ \right) \cdot d_i + u_i \quad i = 1, ..., n \tag{34}$$

where $u_i = \beta \left(x_i - g_1 \left(z_i \right) \right) d_i + \varepsilon_i$. If one regresses y_i on an intercept, x_i^{cen}, z_i and d_i, the coefficients will clearly be biased unless the conditional expectation is constant, say $g_1 \left(z_i \right) = g_1$. So, this procedure is not advisable unless the assumption of a constant conditional expectation is reasonable.[13]

Consider where the constancy assumption is valid by construction, namely when there is no additional variable z_i. Now the true model is

$$y_i = \alpha + \beta x_i^{cen} + \beta \left(g_{1_i} - \xi^+ \right) \cdot d_i + u_i \quad i = 1, ..., n \tag{35}$$

where $g_1 = E \left(x | d = 1 \right)$. This model is a well specified regression including the intercept, x_i^{cen} and censoring indicator d_i. However, there is another issue. For the complete cases ($i = 1, ..., n_0$), the model is linear with intercept α and slope β. For the censored data, the model is a constant, with value $\alpha + \beta g_1$. If g_1 is not known, then there is no parameter restriction between the complete cases and the censored data.[14] Therefore, the estimate of β from model (35) is exactly the same as the estimate from CC analysis, or estimating with complete cases only, and it has the same variance. This follows from our logic but also can easily be shown with matrix algebra. Therefore, there is no gain from including the censored observations together with the censoring indicator.

The same remark applies to the related procedure of using interactions. That is, from (34), one might consider approximating $g_1 \left(z_i \right)$ by a linear function in z_i. For this, one would regress y_i on an intercept, x_i^{cen}, z_i, d_i and $d_i z_i$. It is easy to see that if $g_1 \left(z_i \right)$ were linear, then this would be a well specified model. But, the parametrization has the same effect as discussed for (35); namely there is no parameter restriction between the complete cases and the censored data. As before, with the mean $g_1 \left(z_i \right)$ unknown, this procedure yields no gain over CC analysis.

[13] If x were income and z a demographic variable, then the assumption is the mean of top-coded income is the same for any demographic groups.

[14] This includes the variances as well.

3.3. Semiparametric Information

To consider approaches to estimation, we first consider the information available about the parameters of interest – α, β, γ and σ^2 – in the censored data[15]. For this, consider the following example:

Example 4. *For the model (31)-(33) for censored data, assume*

$$\varepsilon \sim \mathcal{N}\left(0, \sigma^2\right) \quad \text{and} \quad \beta\left(x_i - g_1\left(z_i\right)\right) \sim \mathcal{N}\left(0, \sigma_{\beta x}^2\right).$$

Suppose that η is a vector of nuisance parameters, parameterizing the conditional expectation $g_\eta(z) = E(x|z, d = 1, \eta)$. Under these assumptions, the density of y for the censored data is

$$\ln f\left(y|z, \alpha, \beta, \gamma, \eta\right) = -\ln\sqrt{2\pi} - (1/2)\ln\left(\sigma_{\beta x}^2 + \sigma^2\right) - (1/2)\frac{\left(y - \alpha - \beta g_\eta(z) - \gamma z\right)^2}{\left(\sigma_{\beta x}^2 + \sigma^2\right)} \tag{36}$$

Denoting $\varepsilon = y - \alpha - \beta g_\eta(z) - \gamma z$, the scores of the parameters of interest are

$$\ell_\alpha = \frac{\partial \ln f}{\partial \alpha} = \frac{\varepsilon}{\left(\sigma_{\beta x}^2 + \sigma^2\right)}, \tag{37}$$

$$\ell_\beta = \frac{\partial \ln f}{\partial \beta} = \frac{\varepsilon}{\left(\sigma_{\beta x}^2 + \sigma^2\right)} \cdot g_1(z), \tag{38}$$

$$\ell_\gamma = \frac{\partial \ln f}{\partial \gamma} = \frac{\varepsilon}{\left(\sigma_{\beta x}^2 + \sigma^2\right)} \cdot z \tag{39}$$

and

$$\ell_{\sigma^2} = \frac{\partial \ln f}{\partial \sigma^2} = -\frac{1}{2\left(\sigma_{\beta x}^2 + \sigma^2\right)} + (1/2)\frac{\varepsilon^2}{\left(\sigma_{\beta x}^2 + \sigma^2\right)^2}. \tag{40}$$

The scores of the nuisance parameters are

$$\ell_\eta = \frac{\partial \ln f}{\partial \eta} = \frac{\varepsilon}{\left(\sigma_{\beta x}^2 + \sigma^2\right)} \cdot \frac{\partial g_\eta(z)}{\partial \eta} \tag{41}$$

[15] See Newey (1990) for the definition of semiparametric information and the semiparametric variance bound.

$$\ell_{\sigma^2_{\beta x}} = \frac{\partial \ln f}{\partial \sigma^2_{\beta x}} = -\frac{1}{2\left(\sigma^2_{\beta x} + \sigma^2\right)} + (1/2)\frac{\varepsilon^2}{\left(\sigma^2_{\beta x} + \sigma^2\right)^2} \tag{42}$$

The semiparametric information on α, β, γ and σ^2 is the variance of their scores, after projection onto subspace orthogonal to that spanned by the scores of the nuisance parameters. But if $g_1(z)$ is unrestricted, then a sufficiently rich parameterization $g_\eta(z)$ can be found such that linear combinations of $\left\{\partial g_1(z)/\partial \eta_j\right\}$ will approximate a constant, z and $g_1(z)$ arbitrarily well. Therefore, the projection of ℓ_α, ℓ_β, ℓ_γ, ℓ_{σ^2} onto the subspace orthogonal to the span of ℓ_η, $\ell_{\sigma^2_{\beta x}}$ will be arbitrarily small. Consequently, the semiparametric information on α, β, γ and σ^2 is zero.

It is clear that for more general settings – in particular, general densities of ε and of x given z – we have the same conclusion[16]

Proposition 5. *If there are no restrictions on the conditional expectation $g_1(z) = E(x|z, d = 1)$, then the semiparametric information on α, β, γ and σ^2 from the censored data, is zero. The semiparametric variance bound for the estimation of α, β, γ and σ^2 using complete cases only is the same as the semiparametric variance bound using the complete cases together with the censored data.*

Thus, the phenomena discussed with regard to dummy variable methods above applies more generally. There is no gain in estimation from using the censored data, unless restrictions can be applied to the conditional expectation $g_1(z)$.[17] We now discuss estimation with this in mind.

3.4. Estimation using the Full Data Sample

We return to our general notation, where z is a vector of variables as in (20), and the censoring process is general. The appropriate proxy for x_i in the censored data is $g_1(z_i)$, where $g_1(\cdot)$ denotes the conditional expectation $g_1(z) = E(x|z, d = 1)$. We will be able to use the censored data to increase efficiency if sufficient structure is assumed to permit identification of $g_1(z)$. We do not carry out a full analysis of identification here, but just discuss some aspects to motivate our estimation methods.

[16] The semiparametric variance bound is the inverse of the semiparametric information.

[17] Similar structure is discussed in Horowitz and Manski (1998,2000). See also Robins and Rotnitzky (1995).

Suppose we consider the regression $G(z) = E(x|z)$ of x on z in the full data population, which we express in equation form as

$$x = G(z) + v \tag{43}$$

where $E(v|z) = 0$. In broad terms, if the proxy $G(z)$ can be estimated, then we have a method of estimating $g_1(z)$. This could start with using regression analysis with the complete cases to estimate the regression

$$g_0(z) = E(x|z, d = 0) \tag{44}$$

With the full data sample, we can estimate the probability of censoring

$$p(z) = E(d|z). \tag{45}$$

Therefore, we could characterize $g_1(z)$ from the identity

$$g_1(z) = \frac{G(z) - (1 - p(z)) g_0(z)}{p(z)} \tag{46}$$

The estimation of $G(z)$ depends crucially on the nature of the censoring. If censoring is entirely at random, then we have

$$G(z) = g_0(z) = g_1(z). \tag{47}$$

so that $G(z)$ could be characterized with the complete cases, giving the regression $g_1(z)$.[18]

With top-coding of the form $d = 1\left[x > \xi^+\right]$, the model for estimating $G(z)$ is a censored regression model. Using the complete cases, we have the truncated regression

$$g_0(z) = G(z) + E\left(v \mid v \leq \xi^+ - G(z)\right) \tag{48}$$

so that estimating $G(z)$ would requires either parameteric modeling or semiparametric treatment of the density of v. A similar statement applies with bottom-coding and with related (monotonic)

[18] For instance, Arellano and Meghir (1992) propose using the best linear predictor of x on z as a proxy, which can be estimated using the complete cases only when the censoring is random, or doesn't introduce bias. Much recent methodological work relies on the "censoring at random" or "missing at random" structure.

forms of censoring.

3.5. A Normal Mixed-Censoring Model

In our application below, we apply parametric models to correct for censoring, leaving more general treatments to future research..Our application involves bottom-coding of observed wealth, so we switch our focus from upper bounds to lower bounds in the formulae.

We begin by assuming that the proxy $E(x|z)$ of x is linear

$$G(z) = \delta_0 + \delta_1' z. \tag{49}$$

and we assume that v of (43) is normally distributed and homoskedastic

$$v \sim \mathcal{N}\left(0, \sigma_v^2\right) \tag{50}$$

This assumption facilitates modeling bottom-coding with formulae familiar from censored normal regression models.[19] In particular, with $d = 1\left[x < \xi^-\right]$, we proxy x in the censored data using the regression $g_1(z) = E(x|z, d = 1)$, with

$$g_1(z) = \delta_0 + \delta_1' z - \sigma_v \cdot \lambda_1\left(\frac{\xi^- - \left(\delta_0 + \delta_1' z\right)}{\sigma_v}\right) \tag{51}$$

where $\lambda_1(\cdot) \equiv \phi(\cdot)/\Phi(\cdot)$, with ϕ and Φ the normal density function and the normal c.d.f. respectively.

In our application to wealth effects on consumption, we impliment a slightly more complicated censoring model, that is a mixture of random censoring and bottom-coding. The approach is to model bottom-coding together with random censoring of probability $r(z)$ for observations that are not bottom coded. In particular, let

$$d_1 = 1\left[v < \xi^- - \left(\delta_0 + \delta_1' z\right)\right] \tag{52}$$

[19] See, for instance, Green (2003) or Davidson and McKinnon (2004). Analogous formulae are available for top-coding.

and

$$d_2 = 1\left[s < \eta_0 + \eta_1^{'} z\right] \tag{53}$$

represent the two sources of censoring. We assume $v \sim \mathcal{N}(0, \sigma_v^2)$ and $s \sim \mathcal{N}(0,1)$, and that v and s are conditionally independent given z. The overall censoring indicator d is defined as

$$d = d_1 + d_2 - d_1 d_2 \tag{54}$$

This reflects bottom coding, plus a probability of

$$r_\eta(z) = \Phi\left(\eta_0 + \eta_1^{'} z\right) \tag{55}$$

of (non-bottom-coded) observations being randomly censored to the same value ξ^-. To simplify the formulae that follow, denote the probability of bottom-coding as

$$P_{\delta,\sigma_v}(z) \equiv \Phi\left(\frac{\xi^- - \left(\delta_0 + \delta_1^{'} z\right)}{\sigma_v}\right) \tag{56}$$

To compute the required regression formulae, note first that $d = 0$ if and only if $d_1 = 0$ and $d_2 = 0$. Therefore, by conditional independence,

$$\Pr\{d = 0|z\} = [1 - P_{\delta,\sigma_v}(z)][1 - r_\eta(z)] \tag{57}$$

so that

$$p(z) = \Pr\{d = 1|z\} = P_{\delta,\sigma_v}(z) + r_\eta(z) - P_{\delta,\sigma_v}(z) \cdot r_\eta(z) \tag{58}$$

With mixed censoring, for the regression of x on z in the complete cases, we have

$$\begin{aligned} g_0(z) &= E(x \mid z, d = 0) \\ &= E(x \mid z,\ d_1 = 0 \text{ and } d_2 = 0) \\ &= \delta_0 + \delta_1^{'} z + E\left(v \mid z,\ v < \xi^- - \left(\delta_0 + \delta_1^{'} z\right) \text{ and } s < \eta_0 + \eta_1^{'} z\right) \\ &= \delta_0 + \delta_1^{'} z + E\left(v \mid z,\ v < \xi^- - \left(\delta_0 + \delta_1^{'} z\right)\right) \end{aligned} \tag{59}$$

where the last equality follows from the conditional independence of v and s given z. Therefore, $g_0(\cdot)$ is given by the following formula (which is also appropriate for bottom-coding.only)

$$g_0(z) = \delta_0 + \delta_1' z + \sigma_v \cdot \lambda_0 \left(\frac{\xi^- - \left(\delta_0 + \delta_1' z\right)}{\sigma_v} \right) \tag{60}$$

where $\equiv \phi(\cdot) / [1 - \Phi(\cdot)]$, With mixed censoring, the regression of x on z for the censored data is found by applying (46) using (58) and (59). The result it

$$g_1(z) = \delta_0 + \delta_1' z - \Psi_{\delta,\sigma_v,\eta}(z) \cdot \sigma_v \cdot \lambda_1 \left(\frac{\xi^- - \left(\delta_0 + \delta_1' z\right)}{\sigma_v} \right) \tag{61}$$

where

$$\Psi_{\delta,\sigma_v,\eta}(z) = \frac{P_{\delta,\sigma_v}(z)[1 - r_\eta(z)]}{r_\eta(z) + P_{\delta,\sigma_v}(z)[1 - r_\eta(z)]} \tag{62}$$

The correction term Ψ is easily seen to be $\Psi_{\delta,\sigma_v,\eta}(z) = (p(z) - r_\eta(z))/p(z)$, the relative probability of bottom-coding in the mixed censoring. This completes the formulae that we need for our analysis of the effects of wealth on consumption.

4. The Effects of Wealth on Consumption

4.1. General Discussion

In recent years, many developed countries have witnessed tremendous expansions in consumption expenditures at the same time as substantial increases in household wealth levels.[20] 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

[20] 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.

wealth, say with a stylized model:

$$C_i = \alpha + \beta_1 \cdot INC_i + \beta_2 \cdot W_i + \varepsilon_i \quad i = 1, ..., n \tag{63}$$

where C_i is consumption, INC_i is income and W_i is a measure of wealth (such as the value of stock and bond holdings). It is very common that income is top-coded, producing the kind of censored regressor bias that we have discussed. But also, measures of 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 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.[21] With individual data, estimates tend to be larger,[22] such as 8%. In any case, we plan to investigate whether the censored character of income and wealth can help account for the magnitude of these estimates.[23]

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 important for the design of effective monetary policy.[24]

[21] 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).

[22] See Parker (1999), Juster, Lupton, Smith and Stafford (1999) and Starr-McCluer (1999).

[23] Similarly, particularly large effects of housing wealth on consumption are estimated by Aoki, et. al. (2002a, 2002b) and Attanasio, et. al. (1994), among others. Somewhat smaller estimates are given in Engelhardt (1996) and Skinner (1996).

[24] 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),

	Percentage Censored	Total Observations	Not Censored
Total Wealth	26.6%	11,903	8,735
Housing	43.4%	11,903	6,737
Stock Market	76.5%	11,903	2,797
One or More Censored	80.9%	11,903	2,272

Table 3: Proportion of Censoring in Total Wealth, Housing, and Stock Market Wealth

4.2. Application to Consumption Data

In this section we study the importance of the expansion bias due to censored regressors in an application to consumption and forms of wealth.[25] The data includes consumption, current income, a computed permanent component of consumption that depends on the cohort in which the household belongs, characteristics of the household (such as retirement status, family size, etc.), and financial information. The way the data was constructed, the income variables are *not* censored – the observations with top-coded income variables of the original survey have been dropped – so that our data is already a set of complete cases in terms of income. The only censored variables are the financial variables. There are three sorts of financial variables that interest us: total wealth, housing wealth, and stock market wealth. In this section we refer to all three forms of wealth as financial wealth.

In Table 3 we show the proportion of the variables that are at the censoring bound in the data. As can be seen there is a moderate proportion of total wealth observations that are censored (27 percent) but this increases to 43 percent for housing, to 76 percent for stock market wealth, and to 81% when one or more wealth variables is censored.[26] It is important to highlight that the censoring in the data may not occur as cleanly as our examples have developed. For example, it is reasonable to expect that the degree of censoring increases for smaller values of housing, but we should expect that there is some noise in the reporting. Therefore, some house values may are censored even though their value is larger than some of the ones that are reported and not censored.

Cecchetti et. al. (2000) and Rigobon and Sack (2003).

[25] We thank Jonathan Parker for his tremendous help and support in providing us not only with the data but with valuable suggestions.

[26] For consistency among the components, total wealth is censored when it is less than \$5,000, housing wealth when it is less than \$4,000 and stock market wealth when it is less than \$1,000. This gives slightly higher censoring than when all levels are censored at zero, but facilitates taking logarithms.

	Log Consumption – No Additional Regressors			
	All Data	CC	All Data	CC
Sample Size	11,903	8,735	11,903	2,272
Total Wealth	0.181	0.140	0.149	0.055
	(.0029)	(.0038)	(.0054)	(.0135)
Housing Wealth			0.020	0.069
			(.0053)	(.0132)
Stock Market Wealth			0.033	0.012
			(.0032)	(.0069)

Table 4: Log Consumption Results, Simple Models

To see the most coarse impact of censoring, Table 4 gives estimates of regressing log consumption on wealth and the wealth components, without any additional regressors.[27] If only total wealth is included, there are 8,735 complete (uncensored) cases, and when all three wealth variables are included, there are 2,272 complete cases. With the bivariate regression of log consumption on log wealth, there is an expansion bias of 29%, namely (.181/.140) - 1. With the components included, using all data gives a total wealth elasticity of .202, whereas the complete cases give a total elasticity of .136, which reflects a 48% expansion bias. There are some relative shifts; in particular a much larger housing wealth effect in the complete case data.

For modeling consumption with wealth censoring, we focus on the total wealth effect, using a log-form regression equation similar to that estimated by Parker (1999).

$$\begin{aligned} \ln C_{it} &= \alpha + \beta \cdot \ln W_{it} + \gamma_1 \ln PINC_{it} + \gamma_2 \ln INC_{it} + \\ &\quad \gamma_3 Retirement_{it} + \gamma_4 Family_{it} + \gamma_5 Kids_{it} + \gamma_6 Cohorts_{it} + \gamma_7 Time + \varepsilon_{it} \end{aligned} \tag{64}$$

where $C_{i,t}$ is consumption of household i at time t. W_{it} is total wealth, which is censored.[28] There are two income variables; $PINC_{it}$ is a constructed permanent component of income and

[27] In all tables with empirical results, heteroskedasticity consistent (White) standard errors are presented in parentheses.

[28] To include housing and stock market wealth, we would need to model the joint censoring process of all wealth components. We focus on total wealth only just to keep things simple here.

$INC_{i,t}$ is the current income, which are uncensored regressors in our data. These are uncensored regressors in our data. $Retirement_{it}$, $Family_{it}$, $Kids_{it}$, $Cohorts_{it}$ and $Time$ are other control variables for retirement status, family size, cohorts, etc. For a detailed description of the data and the definition of the variables, see Parker (1999).

Table 5 presents estimates of the model with only the income variables as additional regressors (setting $\gamma_3 = 0, \gamma_4 = 0$, ... in (64)). There is evidence for a mild expansion bias of 11% in the wealth elasticity, with the overall elasticity values much smaller (and much more reasonable) than the estimates without income terms. We estimate the parametric censoring model of Section 3.5 in two forms: with constant probability of random censoring ($r_\eta(z) = r_0$) and the full varying probability model.[29] We see that the income effects are sensitive to the form of censoring, and are dampened when random censoring is assumed constant.

The results with all regressors (income and all controls) are presented in Table 6. With all controls, there is little evidence of variation in the wealth effect across the different specifications. All income elasticities are slightly smaller, and there is again evidence of dampening with a constant probability of random censoring. We view these results as informative and basically in line with our expectations on the impact of censored regressors - in the sense that the original OLS estimates contain biases, and those biases depend on the correlation structure of all the regressors.

One benefit of modeling the censoring process is that we can see how important bottom-coding is relative to random censoring in wealth. Figure 4 shows boxplots of the distribution of estimated probabilities of bottom coding $P_{\hat{\delta},\hat{\sigma}_v}(z_i)$ and of random censoring $r_{\hat{\eta}}(z_i)$. Our estimates suggest that random censoring is much more prevalent than censoring via bottom coding. In particular, the probability of bottom-coding $P_{\hat{\delta},\hat{\sigma}_v}(z_i)$ has a mean of .05 and a median of .01, whereas the probability of random censoring has a mean of .26 and a median of .22. Figure 5 shows a scatterplot of the estimated probabilities, that shows how they vary in similar ways across the observed data. In particular, higher income consumers have a lower chance of bottom-coding, and a lower chance of being randomly censored. In any case, we find these to be particularly interesting findings, worthy of further investigation.

[29] The model is estimated using GMM with moment restrictions implied from the basic model and the censoring process. All results and estimation details are available from the authors.

Regressions with Log Income Terms

			Censoring	
	All Data	CC	Const. Rand.	Varying Rand.
Sample Size	11,903	8,735	11,903	11,903
Total Wealth	0.061	0.054	0.052	0.058
	(.0037)	(.0044)	(.0031)	(.0032)
Current Income	0.198	0.182	0.186	0.201
	(.0126)	(.0151)	(.0048)	(.0049)
Permanent Income	0.204	0.204	0.175	.212
	(.0157)	(.0193)	(.0065)	(.0064)

Table 5: Log Consumption Results, Income Variables included

Regressions with Log Income Terms and All Controls

			Censoring	
	All Data	CC	Const. Rand.	Varying Rand.
Sample Size	11,903	8,735	11,903	11,903
Total Wealth	0.052	0.054	0.054	0.054
	(.0045)	(.0062)	(.0035)	(.0036)
Current Income	0.180	0.165	0.172	0.180
	(.0117)	(.0137)	(.0047)	(.0047)
Permanent Income	0.175	0.177	0.151	.177
	(.0160)	(.0208)	(.0067)	(.0066)

Table 6: Log Consumption Results, Income Variables included

5. Conclusion

The fact that top-coding or bottom-coding of regressors generates expansion bias was a 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 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. We characterized the bias generally, highlighting the importance of the censoring process as well as the value that censored data are set to. With top-coding, the 'many regressor' formulae are complicated, but we were able to see how the censoring biases 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 how this bias compares with familiar bias from errors-in-variables problems.

The estimation issues posed by top-coding, bottom-coding and other kinds of censoring are not trivial. With top-coding, the distribution on the censored tail of x is not observed, which makes it difficult to obtaining efficiency gains from using the censored data. Nevertheless, we were able to study censoring in data on consumption and wealth using a parametric model, and we are very optimistic that semiparametric procedures can be developed in future research.

A similar type of bias arises in instrumental variables estimators when there are censored regressors, although there are some important differences with the case of OLS regression.[30] We present some results of this kind in Rigobon and Stoker (2005a). We have developed specification tests for the presence of censoring bias in Rigobon and Stoker (2005b), applied in the context of empirical work in corporate finance. Beyond studying the substantive empirical questions on consumption and wealth raised above, we plan to study the further efficiency gains available from using variance restrictions available in the censored data.

[30] For instance, IV estimators with random censoring of an endogenous variable display expansion bias, whereas as we have shown, OLS estimators display attenuation bias.

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] Arellano, M. and C. Meghir (1992), "Female Labor Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets," *Review of Economic Studies*, 59, 537-559.

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

[5] 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.

[6] Belsley, D.A., E. Kuh and R. E. Welsch (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Multicollinearity*, Wiley, New York.

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

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

[9] Borjas, G. (1994) "Long-Run Convergence of Ethnic Skill Differentials" NBER 4641.

[10] 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.

[11] 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

[12] 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).

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

[14] 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.

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

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

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

[18] 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.

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

[20] 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.

[21] 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.

[22] Ichimura, H. and E. Martinez-Sanchis (2005), "Identification and Estimation of GMM Models by Combining Two Data Sets," CEMMAP Working Paper, IFS, London, March.

[23] 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..

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

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

[26] 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.

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

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

[29] 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.

[30] Mahajan, A. (2004), "Identification and Estimation of Single Index Models with Misclassified Regressors," Working Paper, Stanford University, July.

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

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

[33] Newey, W. K. (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99-135.

[34] Newey, W.K. and T. M. Stoker (1993), "Efficiency of Weighted Average Derivative Estimators and Index Models," *Econometrica*, 61, 1199-1223.

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

[36] 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.

[37] Petoussis, K., R. D. Gill and C. Zeelenberg (2004), "Statistical Analysis of Heaped Duration Data," draft, Department of Psychology, Vrije Universiteit Amsterdam, February.

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

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

[40] Rigobon, R and T. M. Stoker (2005a) "Instrumental Variables Bias with Censored Regressors,", MIT Working Paper, March.

[41] Rigobon, R and T. M. Stoker (2005b) "Corporate Finance, Financial Statements, and Censoring Problems", MIT Working Paper, April.

[42] Robins, J. M. and A. Rotnitzky (1995), "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90, 122-129.

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

[44] Rotnitzky, A., J. M. Robins and D. O. Scharfstein (1998), "Semiparametric Regression for Repeated Outcomes With Nonignorable Response," *Journal of the American Statistical Association*, 93, 1321-1339.

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

[46] 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.

[47] Starr-McCluer, M. (1999). "Stock Market Wealth and Consumer Spending.' Mimeo, Federal Reserve Board of Governors.

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

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

[50] Torelli, N. and U. Trivellato (1993). "Modeling Inaccuracies in Job-Search Duration Data," *Journal of Econometrics*, 59, 185-211.

[51] 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 $x \sim \mathcal{N}(\mu_x, \sigma_x^2)$. 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 $\xi^-,\ \xi^+$ or the probabilities $p^- = \Pr\left\{x < \xi^-\right\}$, $p^+ = \Pr\left\{x > \xi^+\right\}$; as

$$p^- = \Phi\left(\frac{\xi^- - \mu_x}{\sigma_x}\right);\ 1 - p^+ = \Phi\left(\frac{\xi^+ - \mu_x}{\sigma_x}\right) \tag{65}$$

are fully invertible to

$$\xi^- = \sigma_x \Phi^{-1}\left(p^-\right) + \mu_x;\quad \xi^+ = \sigma_x \Phi^{-1}\left(1 - p^+\right) + \mu_x \tag{66}$$

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

Recall that

$$x_i^{cen} = x_i \cdot 1\left[\xi^- \leq x_i \leq \xi^+\right] + \xi^- \cdot 1\left[x_i < \xi^-\right] + \xi^+ \cdot 1\left[\xi^+ < x_i\right], \tag{67}$$

$$x_i^- = \left(x_i - \xi^-\right) \cdot 1\left[x_i < \xi^-\right] \quad \text{and} \quad x_i^+ = \left(x_i - \xi^+\right) \cdot 1\left[\xi^+ < x_i\right]. \tag{68}$$

We some initial results

$$E\left(x^-\right) = -\sigma_x \phi\left[\Phi^{-1}\left(p^-\right)\right] - p^- \sigma_x \Phi^{-1}\left(p^-\right) \tag{69}$$

$$E\left(x^+\right) = \sigma_x \phi\left[\Phi^{-1}\left(1 - p^+\right)\right] - p^+ \sigma_x \Phi^{-1}\left(1 - p^+\right) \tag{70}$$

$$\begin{aligned} E\left(x^-\right)^2 \ = \ & \left\{\left(1 - \frac{\phi\left[\Phi^{-1}\left(p^-\right)\right]^2}{\left(p^-\right)^2} - \frac{\phi\left[\Phi^{-1}\left(p^-\right)\right]\Phi^{-1}\left(p^-\right)}{p^-}\right)\right. \\ & \left.\left(-\frac{\phi\left[\Phi^{-1}\left(p^-\right)\right]}{p^-} - \Phi^{-1}\left(p^-\right)\right)^2\right\} \cdot \sigma_x^2 p^- \end{aligned} \tag{71}$$

$$E\left(x^{+}\right)^{2} = \left\{\left(1-\frac{\phi\left[\Phi^{-1}\left(1-p^{+}\right)\right]^{2}}{\left(p^{+}\right)^{2}}+\frac{\phi\left[\Phi^{-1}\left(1-p^{+}\right)\right]\Phi^{-1}\left(1-p^{+}\right)}{p^{+}}\right)\right. \tag{72}$$
$$\left.\left(\frac{\phi\left[\Phi^{-1}\left(1-p^{+}\right)\right]}{p^{+}}-\Phi^{-1}\left(1-p^{+}\right)\right)^{2}\right\}\cdot\sigma_{x}^{2}p^{+}$$

$$E\left(x^{cen}x^{-}\right)=\left(\sigma_{x}\Phi^{-1}\left(p^{-}\right)+\mu_{x}\right)\left\{-\sigma_{x}\phi\left[\Phi^{-1}\left(p^{-}\right)\right]-p^{-}\sigma_{x}\Phi^{-1}\left(p^{-}\right)\right\} \tag{73}$$

$$E\left(x^{cen}x^{+}\right)=\left(\sigma_{x}\Phi^{-1}\left(1-p^{+}\right)+\mu_{x}\right)\left\{\sigma_{x}\phi\left[\Phi^{-1}\left(1-p^{+}\right)\right]-p^{+}\sigma_{x}\Phi^{-1}\left(1-p^{+}\right)\right\} \tag{74}$$

The bias is computed by substituting (69)-(74) in the following:

$$E\left(x^{cen}\right)=\mu-E\left(x^{-}\right)-E\left(x^{+}\right) \tag{75}$$

$$Cov\left(x^{-},x^{cen}\right)=E\left(x^{cen}x^{-}\right)-E\left(x^{cen}\right)E\left(x^{-}\right) \tag{76}$$

$$Cov\left(x^{+},x^{cen}\right)=E\left(x^{cen}x^{+}\right)-E\left(x^{cen}\right)E\left(x^{+}\right) \tag{77}$$

$$\begin{aligned} Var\left(x^{cen}\right) &= \mu_{x}^{2}+\sigma_{x}^{2}-E\left(x^{-}\right)^{2}-E\left(x^{+}\right)^{2} \\ &\quad -2E\left(x^{cen}x^{-}\right)-2E\left(x^{cen}x^{+}\right)-\left[E\left(x^{cen}\right)\right]^{2} \end{aligned} \tag{78}$$

$$\Lambda^{-}=\frac{Cov\left(x^{-},x^{cen}\right)}{Var\left(x^{cen}\right)}\ \text{ and }\ \Lambda^{+}=\frac{Cov\left(x^{+},x^{cen}\right)}{Var\left(x^{cen}\right)}. \tag{79}$$

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

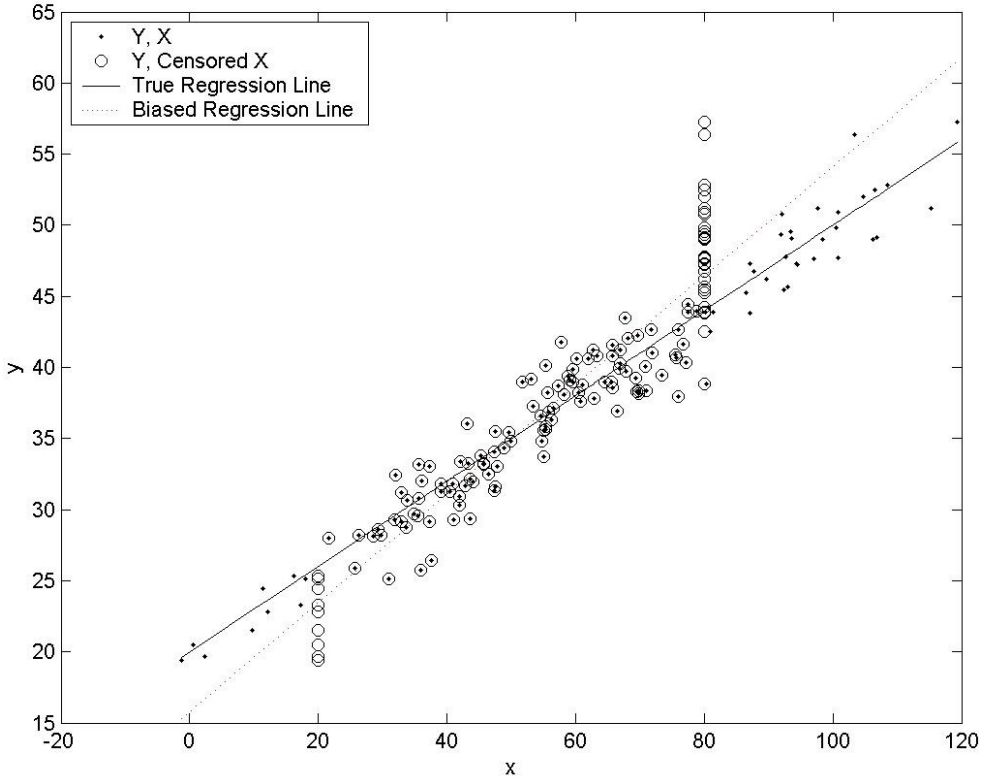

Figure 1: Expansion Bias

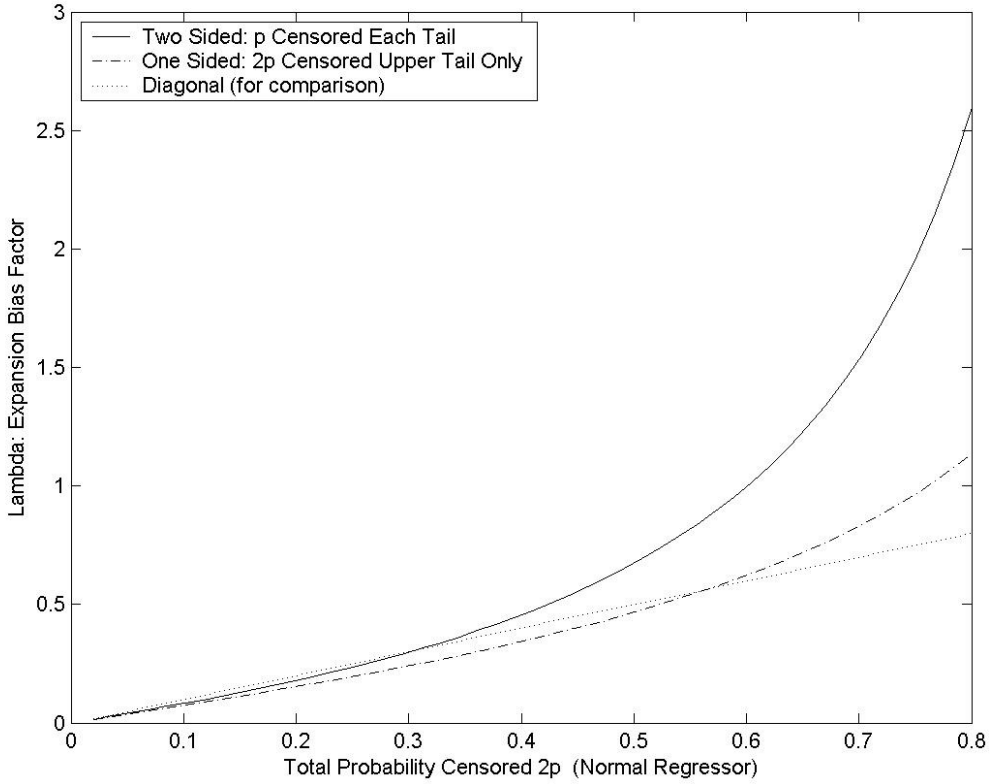

Figure 2: Expansion Bias by Probability Censored

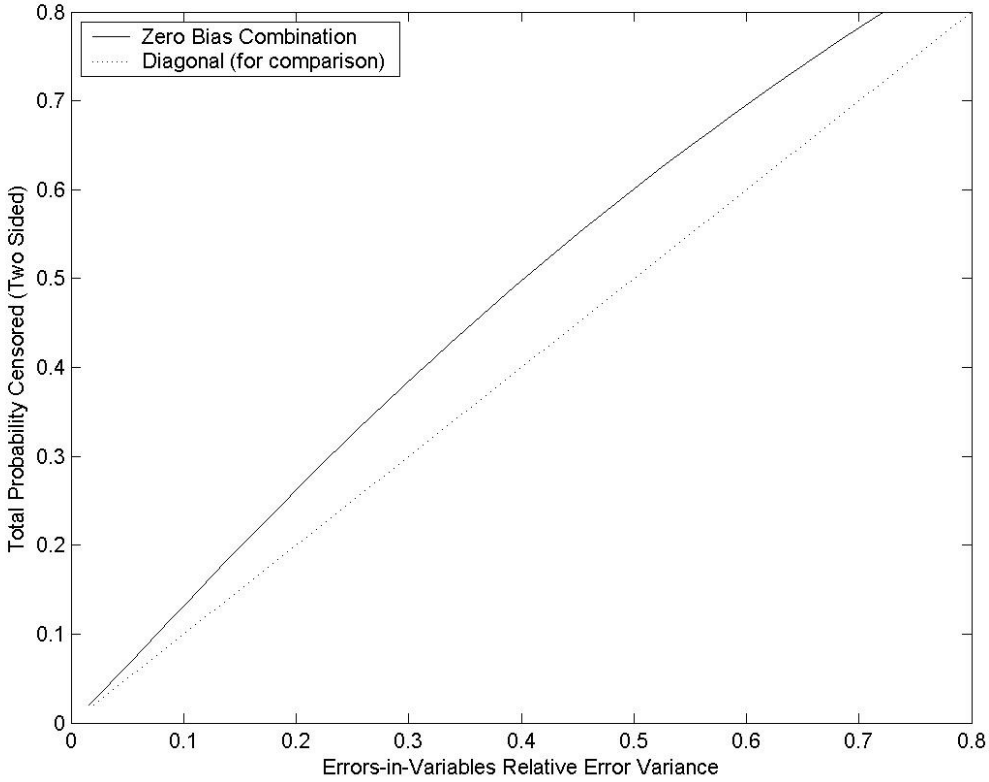

Figure 3: Errors-in-Variables and Expansion Bias: Normal Regressor

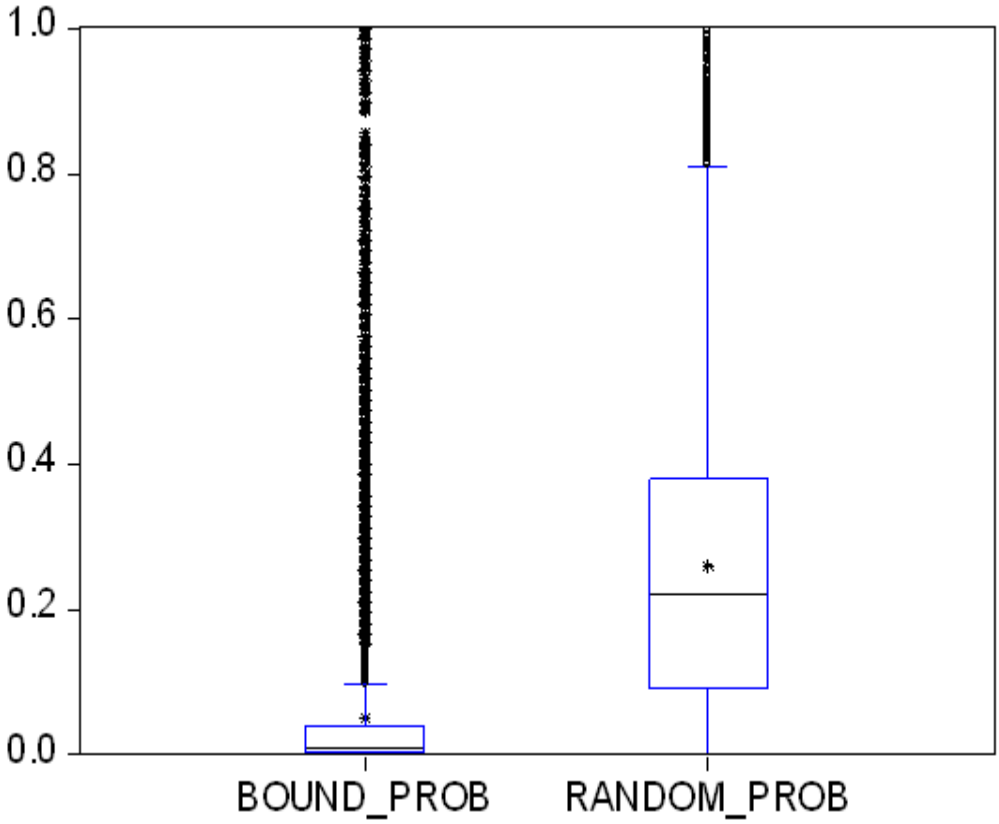

Figure 4: Bound Censoring and Random Censoring: Distribution of Estimated Probabilities

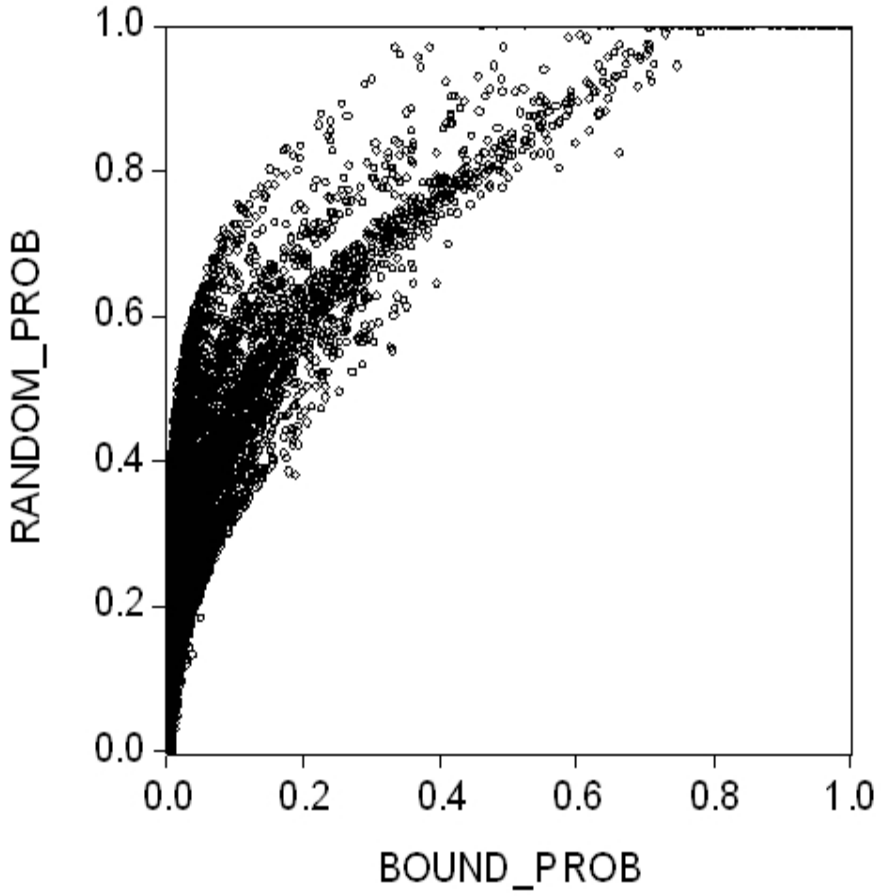

Figure 5: Bound Censoring and Random Censoring: Scatter of Estimated Probabilities