Journal of Economic Perspectives—Volume 15, Number 4—Fall 2001—Pages 29–42

Semiparametric Censored Regression Models

Kenneth Y. Chay and James L. Powell

A regression model is *censored* when the recorded data on the dependent variable cuts off outside a certain range with multiple observations at the endpoints of that range. When the data are censored, variation in the observed dependent variable will understate the effect of the regressors on the "true" dependent variable. As a result, standard ordinary least squares regression using censored data will typically result in coefficient estimates that are biased toward zero.

Traditional statistical analysis uses maximum likelihood or related procedures to deal with the problem of censored data. However, the validity of such methods requires correct specification of the error distribution, which can be problematic in practice. In the past two decades, a number of semiparametric alternatives for dealing with censored data have been proposed. In a semiparametric approach, part of the functional form of the model—usually the regression function—is parametrically specified by the researcher based upon plausible assumptions, while the rest of the model is not parameterized.[1] While the theoretical literature has produced several semiparametric estimators for the censored data model, published applications of these estimators to empirical problems in economics have lagged far behind.

This paper reviews the intuition and computation of a handful of semiparametric estimators proposed for the censored regression model. The various estimators are used to examine changes in black-white earnings inequality during the

[1] In this issue, the paper by DiNardo and Tobias offers further discussion of nonparametric and semiparametric analysis.

■ *Kenneth Y. Chay is Assistant Professor of Economics and James L. Powell is Professor of Economics, University of California, Berkeley, California.*

1960s, around the time of the passage of the Civil Rights Act of 1964, based on longitudinal Social Security Administration (SSA) earnings records. These earnings records are censored at the taxable maximum; that is, anyone earning more than the maximum that was taxable under Social Security is recorded as having earned at the maximum. Thus, above the maximum, the data on earnings do not accurately reflect actual earnings. Ordinary least squares analysis of these data implies little convergence in the earnings of black and white workers during the 1960s. On the other hand, the estimates from the semiparametric models that account for censoring suggest that significant black-white earnings convergence did occur after 1964. Comparisons of the results from parametric and semiparametric procedures help pinpoint sources of misspecification in the parametric approach.

Censored Regression Models and Estimators

The Social Security Administration data set that we analyze suffers from the simplest form of data censoring, *interval censoring*, for which the values of the "true" dependent variable, y^*, are observed only if they fall within some known, often one-sided, interval $[a, b]$. Otherwise, the closest endpoint of the interval is observed instead of y^*. Tobin (1958) used this model to analyze consumer expenditures on automobiles, with $a = 0$ and $b = \infty$, and economists generally refer to regression models with nonnegativity constraints as *Tobit models*. Other typical applications of these censored regression models are to *right-censored data*, where $a = -\infty$ and b represents a maximum recordable value for the dependent variable. Such models arise for *top-coded data*, where sufficiently large values of the true variable y^* are recorded as "at least equal to b." In our primary empirical application, the dependent variable, the logarithm of annual earnings, is "top-coded," or censored from above, with b equal to the logarithm of the maximum annual earnings subject to Social Security taxes in a given year.

Algebraically, the model for the observed dependent variable y under interval censoring is

$$y = \begin{cases} a & \text{if } x'\beta + \varepsilon < a, \\ b & \text{if } x'\beta + \varepsilon > b, \\ x'\beta + \varepsilon & \text{otherwise,} \end{cases}$$

where y is the observed value of the dependent variable, x is a vector of observed explanatory variables, β is a vector of unknown regression coefficients to be estimated, ε is an unobserved error term, and a and b are the censoring interval endpoints. While the true dependent variable y^* satisfies a standard linear regression model, the observed variable y clearly does not when y^* lies outside $[a, b]$. Because y does not vary with the regressors x when it is censored (unlike the true variable y^*), standard least squares regression will underestimate the magnitude of the regression slope coefficients.

If the distribution of the error terms ε given the regressors has a known parametric form—for example, normally distributed and homoskedastic errors—it is straightforward to derive and maximize the likelihood function. This provides a consistent and approximately normal estimator of the regression coefficients β (see, for example, Amemiya, 1985, chapter 10). However, in many empirical problems, the distribution of the errors is not known or is subject to heteroskedasticity of unknown form. In such cases, the maximum likelihood estimator will not provide a consistent estimate (Goldberger, 1983; Arabmazar and Schmidt, 1981, 1982). Also, for censored panel data with fixed effects—that is, censored data with repeated observations on individuals over time and intercept terms that are allowed to vary freely across individuals—maximum likelihood estimation methods will generally be inconsistent even when the parametric form of the conditional error distribution is correctly specified (Honoré, 1992).

Thus, it is important to develop estimation methods that provide consistent estimates for censored data even when the error distribution is nonnormal or heteroskedastic. Here, we focus on describing three particular semiparametric estimators for the censored regression model, with acronyms CLAD, SCLS and ICLAD. All three estimators can be computed by alternating between a "recensoring" step, in which the data are "trimmed" (using the current parameter estimates) to compensate for the censoring problem, and a "regression" step using the trimmed data to obtain coefficient estimates. More complete algebraic derivations and discussions of the various alternatives are available in Powell (1994, section 5.3). Further details on large-sample properties and standard error formulae can be found in the cited references.

The *censored least absolute deviations* (CLAD) estimation method was proposed by Powell (1984). For the linear model, the method of least absolute deviations obtains regression coefficient estimates by minimizing the sum of absolute residuals. It is a generalization of the sample *median* to the regression context just as least squares is a generalization of the sample mean to the linear model. If the true dependent variable y^* were observed, then its median would be the regression function $x'\beta$ under the condition that the errors have a zero median. Least absolute deviations could then be used to estimate the unknown coefficients.

When the dependent variable y is censored, its median is unaffected by the censoring if the regression function $x'\beta$ is in the uncensored region (that is, if $x'\beta$ is in the interval $[a, b]$). However, if the regression function $x'\beta$ is below the lower threshold a (or above the upper threshold b), then more than 50 percent of the distribution will "pile up" at a (or b). In this case, the median of y is that interval endpoint, which does not depend on $x'\beta$. Thus, computation of the CLAD estimator alternates between deleting observations with estimates of the regression function $x'\beta$ that are outside the uncensored region $[a, b]$ (the "recensoring" step) and estimating the regression coefficients by applying least absolute deviations to the remaining observations (the "regression" step), as described by Buchinsky (1994).[2]

The *symmetrically censored least squares* (SCLS) estimation method, proposed by Powell (1986b), is based on a "symmetric trimming" idea. For simplicity, suppose that $a = -\infty$, so that the data are "top-coded" at b (as in our empirical application), and assume that the true dependent variable y^* is symmetrically distributed around the regression function $x'\beta$. Due to the censoring, the observed dependent variable y has an asymmetric distribution, since its upper tail is "piled up" at the censoring point b. This situation is illustrated in Figure 1. However, symmetry can be restored by "symmetrically censoring" the dependent variable y from below at the point $2x'\beta - b$. Now the regression function is equidistant from both censoring points. Since this new "recensored" dependent variable is symmetrically distributed around the regression function, the regression coefficients can be estimated by least squares. Iterating between this "symmetric censoring" of the dependent variable using the current estimates (which drops observations with values of the regression function above b) and least squares estimation of the regression coefficients using the "symmetrically trimmed" data yields the SCLS estimator.

Finally, the *identically censored least absolute deviations* (ICLAD) and *identically censored least squares* (ICLS) estimation methods were proposed by Honoré and Powell (1994). The motivation for these estimators is similar to the "symmetric trimming" idea used to derive the SCLS estimator, but involves recensoring the dependent variable for pairs of observations so that their density functions have the same shape. Suppose that the dependent variables for two observations, y_1 and y_2, are censored from above at b, as depicted in Figure 2. While the shape of the densities for these two observations would be the same in the absence of censoring, the censored densities have different shapes. Also, the distances from the regression functions, $x_1'\beta$ and $x_2'\beta$, to the censoring point b are different.

However, the second observation y_2, which has the smaller regression function $x_2'\beta$, can be artificially "recensored" at the point $x_2'\beta - x_1'\beta + b \equiv \Delta x'\beta + b$. The resulting "identically censored" density for y_2 will have the same shape as the density for y_1. Further, the difference between the two identically censored variables will be symmetrically distributed around the difference in their regression functions, $\Delta x'\beta$. As a result, the regression coefficients can be estimated by finding the value of β that minimizes the sum of absolute (ICLAD) or squared (ICLS) differences of the "identically censored" residuals across all distinct pairs of observations. As with the estimators discussed above, the ICLS and ICLAD estimators can

[2] Quantile regression, as discussed by Koenker and Hallock in this issue, is based on a weighted version of the least absolute deviations approach (Koenker and Bassett, 1978). Indeed, it is straightforward to extend the CLAD approach to analyze the case of censored quantile regression (CRQ) estimation, as proposed by Powell (1986a).

Figure 1
Density of y and "Symmetrically Censored" Density

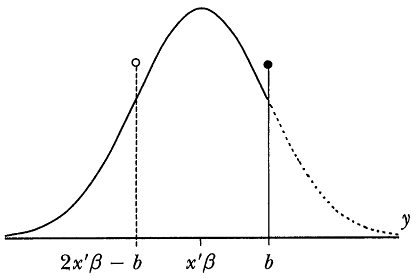
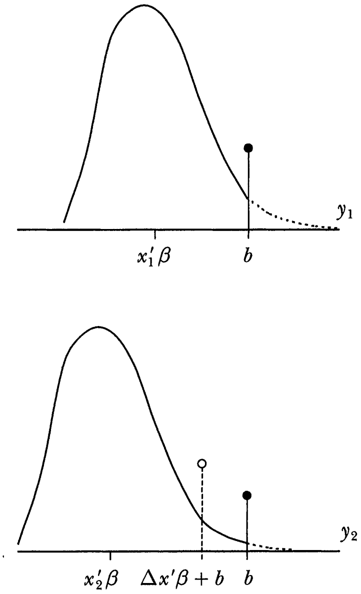


Figure 2
Densities of y_1 and y_2 and "Identically Censored" Densities



be calculated by repeated application of linear least squares or least absolute deviations regression programs.[3]

Honoré (1992) originally proposed the concept behind the ICLAD and ICLS estimators for censored panel data with individual-specific intercepts (also known as "fixed effects"). Instead of identically censoring observations for pairs of individuals, the approach can be applied to pairs of observations across time periods for

[3] In the empirical application, we focus on the least absolute deviations version of the estimator—ICLAD—rather than the least squares version (ICLS), since simulation evidence suggests that it performs better in small samples (Honoré and Powell, 1994).

each individual. Differencing the identically censored observations for a fixed individual eliminates the fixed effect, just as time differences eliminate fixed effects in the standard linear panel data model. In fact, this approach yields consistent estimates even when correctly specified maximum likelihood would not.

Each of these estimation procedures imposes a particular assumption on the underlying error distribution. The SCLS estimator is based on the assumption that the error terms are symmetrically distributed around zero, which implies that their median (and mean) is zero. While compatible with the traditional assumption of normally distributed and homoskedastic errors, the symmetry assumption is less restrictive and provides consistent estimates when these parametric conditions fail to hold. However, it is stronger than the "zero median" restriction exploited by the CLAD estimator, which permits nonnormal, heteroskedastic and asymmetric errors. The ICLAD and ICLS estimators assume that the error terms are identically (but not necessarily symmetrically) distributed, ruling out heteroskedasticity, but permitting asymmetry of the error distribution.[4]

As with the choice of regressors, it is ultimately up to the empirical researcher to determine which assumption is most plausible for the particular application. In practice, though, computation of several parametric and semiparametric estimators provides a useful guide to the sensitivity of the results to the identifying assumptions, either through casual comparison of the coefficient estimates and standard errors or more formal specification tests of the kind described by Newey (1987). In other words, the different estimation approaches give the researcher additional ways to "cut the data" to see which results are robust to alternative specifications.

The interval censored regression model is a special case of the more general *censored selection* model, in which the dependent variable y is generated as

$$y = d \times (x'\beta + \varepsilon),$$

where d is an observable binary ("dummy") variable indicating whether the true dependent variable $y^* = x'\beta + \varepsilon$ is observed $(d = 1)$ or "censored" $(d = 0)$. For the special case of interval censoring considered above, d is an indicator for whether the true variable y^* is in the uncensored region $[a, b]$. More generally, though, d will depend on regressors and error terms that are related to, but distinct from, those in the equation for y^*. For example, market wages may only be observed for individuals with positive labor supply, which is a different form of censoring than top coding in wages.

In such selection models, parametric estimation methods for the coefficients β are typically based on maximum likelihood (Gronau, 1973) or the "two-step" strategy proposed by Heckman (1976, 1979). When the error distribution is not parametrically specified, however, semiparametric estimation of the regression

[4] Much of the theoretical literature on semiparametric estimation of censored regression models has focused on this assumption. However, some simulation evidence suggests that heteroskedasticity causes greater bias in standard maximum likelihood estimation than nonnormality (Powell, 1986b).

coefficients generally involves explicit nonparametric estimation of density or regression functions, unlike the simpler methods for interval censored data described above.[5] Also, semiparametric identification of the censored selection model generally requires an "exclusion restriction"—that is, a regressor that is included in the set of regressors for the binary variable d must be excluded from the list of regressors x in the equation of interest. This exclusion restriction (or instrumental variable), which is not required for the interval-censoring model, may not be plausible in many empirical applications. Due to these difficulties, empirical applications of semiparametric selection models are even less common than applications of the semiparametric censored regression models described above.[6]

An Empirical Application: Relative Earnings of Black Men in the South During the 1960s

Title VII of the Civil Rights Act of 1964, which went into effect on July 2, 1965, outlawed discrimination against black and female workers and established the Equal Employment Opportunity Commission to monitor compliance with Title VII and to enforce its statutes. Executive Order 11246, signed by President Johnson on September 24, 1965, prohibited discrimination by federal contractors and created its enforcement arm, the Office of Federal Contract Compliance, to monitor contractors. Many states had also adopted their own fair employment practice laws before 1964 forbidding discrimination among employers located within the state. These laws were similar to Title VII and established state-level commissions to hear individual discrimination claims. However, none of the 21 states with enforceable state laws before the passage of the 1964 Civil Rights Act were in the South.

We use longitudinal data on earnings to estimate the impact of these civil rights policies. In a joint project of the Census Bureau and the Social Security Administration (SSA), respondents to the 1973 and 1978 March Current Population Surveys were matched by their Social Security numbers to their Social Security earnings histories. The resulting files contain survey responses on race, gender, education, age and region of residence as of the survey year for persons in the March surveys linked to any earnings for which they paid Social Security taxes.

We examine the pooled data containing earnings information from 1958 to 1974, with a particular focus on the years 1963, 1964, 1970 and 1971. We use a sample of black and white men living in southern states who were born in the period 1910–1939 (the youngest man in the sample was 24 in 1963, while the oldest was 61 in 1971). We focus on men in the South because none of the states in the South had fair employment practice laws before 1965, and Title VII enforcement

[5] An exception is given by Honoré, Kyriazidou and Udry (1997).

[6] One example of an early implementation is provided by Newey, Powell and Walker (1990), which applied semiparametric estimation methods to the Mroz (1987) data on the labor supply of married women.

activity was primarily directed at racial discrimination in the South. A within-cohort analysis is used to control for the changing composition of workers over time. The final sample consists of 10,105 men, and our analysis uses all men with nonzero earnings in a given year.[7]

A significant shortcoming of the earnings data is that many records are censored at the Social Security maximum taxable earnings level. In addition, the real value of the tax ceiling changed substantially during the time period of interest, rising from a little over $15,000 (in 1982–1984 dollars) in 1963–1964 to about $20,000 in 1970–1971. At least 32 percent of the sample is classified as earning the top-coded amount from 1958 to 1974, and this share fluctuates considerably during the key periods, reaching a peak of 54 percent in 1965. Consequently, any estimates of the impact of Title VII on the black-white earnings gap that do not explicitly account for censoring at the tax ceiling and changes in it could be severely biased.

We use several approaches to estimate the interval censoring model. The dependent variable in each case is the natural logarithm of annual taxable earnings, and the explanatory variables are race, level of education, age and age-squared. Table 1 presents the estimation results for the race and education coefficients based on the various estimators. The first column, headed OLS1, contains the ordinary least squares estimates based on all of the data. The second column, headed OLS2, presents the least squares results using only the observations that are not censored. The third column contains the Tobit maximum likelihood estimates under the assumption that the errors are normally distributed and homoskedastic. The remaining columns present the results for the three semiparametric estimators: CLAD, SCLS and ICLAD. The Tobit, CLAD and SCLS estimators were implemented using the Stata software package, while the ICLAD estimator was calculated using the Gauss package. For each estimator, we have created Stata "ado" files that are available at ⟨http://elsa.berkeley.edu/~kenchay⟩.[8]

It is clear from Table 1 that the least squares and maximum likelihood estimates of the black-white log-earnings gap and the returns to education are extremely biased when compared to the semiparametric estimators. We think of the CLAD estimator as the natural benchmark, since it is consistent under the normality of errors assumption justifying the maximum likelihood estimator, under the independence of errors assumption justifying the ICLAD estimator, and under the conditional symmetry of errors assumption justifying the SCLS estimator. When compared to the CLAD benchmark, the least squares estimator based on all of the data (OLS1) actually does better than the maximum likelihood estimator. For this

[7] Over 84 percent of the men in the sample have positive earnings in 1963–1964. This figure is 83 percent in 1970–1971.

[8] The standard errors for OLS, MLE and ICLAD were calculated using standard approximations. The standard errors for CLAD and SCLS were calculated using the bootstrap techniques discussed by Brownstone and Valletta in this issue. It is much more efficient to calculate the ICLAD estimator using Gauss instead of Stata.

Table 1

Estimated Effects of Race and Education on Log-Earnings

(estimated standard errors in parentheses)

	OLS1	*OLS2*	*MLE*	*CLAD*	*SCLS*	*ICLAD*
Black-White Gap						
1963	−0.355	−0.183	−0.629	−0.416	−0.444	−0.474
	(0.033)	(0.038)	(0.044)	(0.027)	(0.031)	(0.032)
1964	−0.349	−0.154	−0.674	−0.428	−0.444	−0.473
	(0.032)	(0.038)	(0.044)	(0.033)	(0.036)	(0.031)
1970	−0.262	−0.115	−0.508	−0.278	−0.302	−0.338
	(0.032)	(0.037)	(0.044)	(0.020)	(0.031)	(0.029)
1971	−0.242	−0.111	−0.486	−0.244	−0.287	−0.312
	(0.031)	(0.038)	(0.044)	(0.022)	(0.032)	(0.031)
Returns to Education						
1963	0.041	0.012	0.102	0.051	0.068	0.073
	(0.003)	(0.004)	(0.004)	(0.004)	(0.007)	(0.003)
1964	0.040	0.013	0.103	0.064	0.079	0.075
	(0.003)	(0.005)	(0.004)	(0.006)	(0.007)	(0.003)
1970	0.037	0.003	0.101	0.055	0.066	0.071
	(0.003)	(0.005)	(0.004)	(0.003)	(0.006)	(0.003)
1971	0.035	0.002	0.100	0.054	0.065	0.070
	(0.002)	(0.004)	(0.004)	(0.003)	(0.005)	(0.003)

Notes: The dependent variable is the natural logarithm of annual taxable earnings. Regressions also include a constant and age and age-squared as explanatory variables. Observations with nonpositive earnings are dropped from the analysis. The sample sizes for 1963, 1964, 1970 and 1971 are 8525, 8529, 8391 and 8275, respectively. The OLS2 specification also drops top-coded observations, leading to sample sizes of 4632, 4267, 4485 and 4163. MLE is Tobit maximum likelihood; CLAD is censored least absolute deviations; SCLS is symmetrically censored least squares; ICLAD is identically censored least absolute deviations.

application, it appears that misspecifying the errors as being normally distributed and using maximum likelihood estimation results in more biased estimates than ignoring the censoring problem entirely and using least squares estimation. A more formal test of the normality assumption also suggests that it is violated for the log-earnings model.[9]

There are sizeable differences in the estimated effects of education on earnings across the three semiparametric estimators. While the ICLAD and SCLS estimators of the education premium are similar, they are always greater than the CLAD estimator. These differences are significant given the precision of the estimates and range from 17 percent to 43 percent for the ICLAD estimator and

[9] Chay and Honoré (1998) calculate the test statistics for nonnormality and for heteroskedasticity in censored regression models discussed by Chesher and Irish (1987). The test statistic for detecting nonnormality ranges from 900.47 to 1200.68. Under the null, this statistic has an (asymptotic) $\chi^2(2)$ distribution with a 1 percent critical value of 9.21. Therefore, we easily reject the hypothesis that the errors are normally distributed. The test for heteroskedasticity yields statistics between 84.71 and 90.85. Under the null, these have (asymptotic) $\chi^2(12)$ distributions with a 1 percent critical value of 26.22. We therefore also reject the null of no heteroskedasticity.

20 percent to 33 percent for the SCLS estimator. The differences in the estimates of the black-white earnings gap are smaller, with the CLAD estimator about 11 percent to 28 percent and 4 percent to 18 percent smaller in magnitude than the ICLAD and SCLS estimators, respectively. Strikingly, the semiparametric approaches all result in more precise estimates of the race coefficient than maximum likelihood estimation. For example, the standard errors of the CLAD estimator are 25 percent to 55 percent smaller than the standard errors of the Tobit estimator.[10]

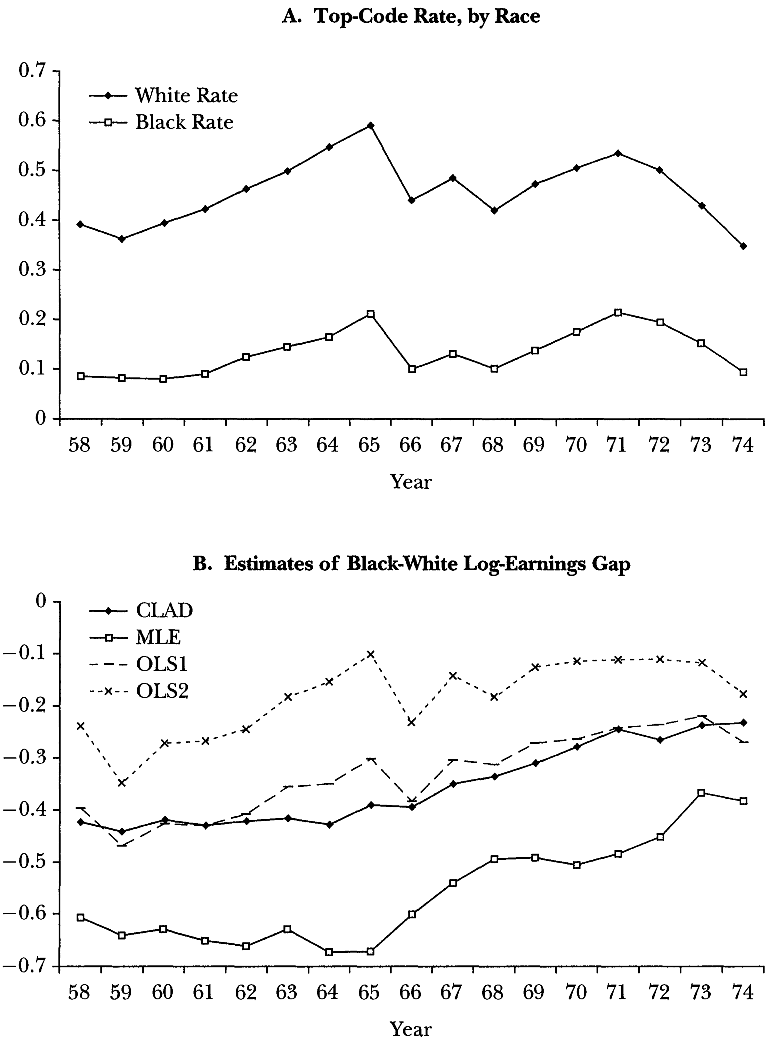
The differences in the coefficient estimates across the various estimators can be used as a sort of specification check, similar in spirit to the Newey (1987) specification analysis mentioned earlier. For the education coefficient, the large differences between the maximum likelihood and semiparametric estimates suggest that nonnormal errors are an important source of bias in the Tobit estimator. Further, the significant differences among the semiparametric estimates imply that heteroskedasticity and asymmetry of the errors are also sources of misspecification in the maximum likelihood estimator of the education premium. Conversely, for the black-white earnings gap, the smaller differences among the semiparametric estimates suggest that nonnormality is the biggest source of bias in the Tobit estimator, with heteroskedasticity and asymmetry playing smaller roles.

To examine the question of specification in more detail, we estimated the distribution of the error terms derived from the CLAD estimates, using the Kaplan and Meier (1958) estimator. The resulting estimated error distribution for log-earnings has fatter tails than does a normal distribution. The maximum likelihood estimator is sensitive to values in the tails, while the least absolute deviations estimator, which focuses on the median value, is unaffected by extreme observations. Since black men are more likely to be in the left-hand tail of the distribution, fat tails can explain the consistently larger (in magnitude) maximum likelihood estimates of the race coefficient. They can also explain the larger sampling errors of the Tobit estimator relative to the semiparametric estimators. Thus, abnormally long tails in the log-earnings distribution may be the major source of misspecification in the maximum likelihood estimates of the black-white earnings gap (Chay, 1995; Chay and Honoré, 1998).

Based on the series of cross-sectional estimators for the four years shown in Table 1, the maximum likelihood and semiparametric approaches yield very similar estimates of changes in black-white relative earnings during the late 1960s. The maximum likelihood and semiparametric estimates all imply that the black-white earnings gap narrowed about 0.15 log points from 1963–1964 to 1970–1971. We conclude from this that while there is bias in the Tobit estimator of the race coefficient, this bias is fixed over time. Thus, it is "differenced out" when one examines changes in the estimated race coefficient. However, the two ordinary least

[10] The standard errors for the CLAD and SCLS estimators were calculated using 500 bootstrap replications. Applying the bootstrap to the Tobit maximum likelihood estimator results in standard errors that are nearly identical to those presented in Table 1, which are based on the asymptotic approximation. Thus, the bootstrap technique is not the source of the differences in the estimated standard errors.

Figure 3
Top-Code Rate and Estimates of Black-White Log-Earnings Gap, 1958–1974



A. Top-Code Rate, by Race

B. Estimates of Black-White Log-Earnings Gap

squares estimators imply that relative earnings only converged between 0.06 (OLS2) and 0.10 (OLS1) log points during the period of interest. Not accounting for the severe censoring in the earnings data results in downwardly biased estimates of the impact of Title VII. Also, although the CLAD estimator imposes the weakest stochastic restrictions on the error terms, it results in the most precise estimates of the policy effects.

Figures 3A and 3B provide a more detailed picture of the various estimators. The top panel shows the percentage of workers in the sample recorded as earning at the taxable maximum (the top-code rate) from 1958 to 1974, separately by race. The bottom panel plots the estimated black-white log-earnings gaps from the OLS1,

Table 2
Fixed-Effects ICLAD Estimates of Effects of Race and Education
(estimated standard errors in parentheses)

	1963–64	*1963–70*	*1963–71*	*1964–70*	*1964–71*	*1970–71*
Change in						
Black-White Gap	0.011	0.102	0.136	0.095	0.108	0.015
	(0.007)	(0.017)	(0.021)	(0.020)	(0.019)	(0.007)
Returns to Education	0.002	0.001	0.000	0.000	−0.003	0.000
	(0.001)	(0.002)	(0.003)	(0.003)	(0.002)	(0.001)

Notes: See notes to Table 1. The sample is the 7,435 men with positive earnings in all four years. For each pair of years, the absolute error loss function was used to estimate the identically censored panel data model with fixed effects. The estimates represent the change in the coefficients between the two years.

OLS2, MLE and CLAD estimators for the series of cross-sections from 1958–1974. There is a striking correspondence between changes in the top-code rate in the top panel and changes in the ordinary least squares estimates in the bottom panel. Indeed, it seems that most of the changes in the ordinary least squares estimates over time are being driven by changes in the amount of censoring varying by race. This results in a severe understatement of the racial earnings convergence in the late 1960s. The time series of the CLAD and maximum likelihood estimates, on the other hand, have no association with the top-code rates, although, as noted earlier, the maximum likelihood estimates systematically overstate the size of the black-white earnings gap when compared to the CLAD estimates. The maximum likelihood and CLAD estimates imply substantial black economic progress in the South after 1964; a result that is masked by the ordinary least squares estimates.

Finally, Table 2 presents the fixed-effects estimation results based on the ICLAD estimator for each of the six possible pairs of time periods in our panel data. The table entries give the estimated change in the race and education coefficients between the two specified periods. The analysis includes age and age-squared as explanatory variables. To compare the parameter estimates across the six columns, the sample is restricted to the 7,435 men with positive earnings in all four years. This reduces the sample size by about 10 percent to 13 percent relative to the cross-sectional samples.

The black-white earnings gap narrowed substantially during the period of interest, even after accounting for individual-specific fixed effects. The relative earnings of black men increased about 0.12–0.14 log-points from 1963 to 1971. These estimates are similar to those implied by the series of maximum likelihood and semiparametric estimates of the cross-sectional censored regression model in Table 1. A key assumption underlying the fixed-effects ICLAD estimator is that the distribution of the unobservables is the same in all time periods for a given individual. A formal specification test did not reject the restrictions implied by this assumption at conventional levels of significance (Chay and Honoré, 1998).

Conclusion

When data are censored, ordinary least squares regression can provide misleading estimates. The results from the semiparametric models show that there was significant earnings convergence among black and white men in the American South after the passage of the 1964 Civil Rights Act, a result that was masked by least squares analysis. The semiparametric methods can also provide information on the sources of misspecification in parametric estimation approaches. In the log-earnings model, it appears that abnormally long tails are the major source of bias in the Tobit maximum likelihood estimates.

■ *We thank Bo Honoré, David Lee, and especially Alan Krueger, Timothy Taylor and Michael Waldman for their helpful comments. Marcelo Moreira provided outstanding research assistance. Support from the Alfred P. Sloan Foundation and the Center for Advanced Study in the Behavioral Sciences is gratefully acknowledged.*

References

Amemiya, Takeshi. 1985. *Advanced Econometrics.* Cambridge: Harvard University Press.

Arabmazar, Abbas and Peter Schmidt. 1981. "Further Evidence on the Robustness of the Tobit Estimator to Heteroskedasticity." *Journal of Econometrics.* November, 17:2, pp. 253–58.

Arabmazar, Abbas and Peter Schmidt. 1982. "An Investigation of the Robustness of the Tobit Estimator to Non-Normality." *Econometrica.* July, 50:4, pp. 1055–63.

Buchinsky, Moshe. 1994. "Changes in the U.S. Wage Structure 1963–1987: Application of Quantile Regression." *Econometrica.* March, 62:2, pp. 405–58.

Chay, Kenneth Y. 1995. "Evaluating the Impact of the 1964 Civil Rights Act on the Economic Status of Black Men using Censored Longitudinal Earnings Data." Unpublished manuscript, Department of Economics, Princeton University.

Chay, Kenneth Y. and Bo E. Honoré. 1998. "Estimation of Semiparametric Censored Regression Models: An Application to Changes in Black-White Earnings Inequality During the 1960s." *Journal of Human Resources.* Winter, 33:1, pp. 4–38.

Chesher, Andrew and Margaret Irish. 1987. "Residual Analysis in the Grouped and Censored Normal Linear Model." *Journal of Econometrics.* January/February, 34:1-2, pp. 33–61.

Goldberger, Arthur S. 1983. "Abnormal Selection Bias," in *Studies in Econometrics, Time Series, and Multivariate Statistics.* S. Karlin et al., eds. New York: Academic Press, pp. 67–84.

Gronau, Reuben. 1973. "The Effect of Children on the Housewife's Value of Time." *Journal of Political Economy.* March/April, 81:2, pp. S168–S199.

Heckman, James J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economics and Social Measurement.* Fall, 5:4, pp. 475–92.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica.* January, 47:1, pp. 153–61.

Honoré, Bo E. 1992. "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects." *Econometrica.* May, 60:3, pp. 533–65.

Honoré, Bo E. and James L. Powell. 1994. "Pairwise Difference Estimators for Censored and Truncated Regression Models." *Journal of*

Econometrics. September/October, 64:1-2, pp. 241–78.

Honoré, Bo E., Ekaterini Kyriazidou and Christopher Udry. 1997. "Estimation of Type 3 Tobit Models Using Symmetric Trimming and Pairwise Comparisons." *Journal of Econometrics.* January/February, 76:1-2, pp. 107–28.

Kaplan, E. L. and P. Meier. 1958. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association.* 53, pp. 457–81.

Koenker, Roger and Gilbert S. Bassett, Jr. 1978. "Regression Quantiles." *Econometrica.* January, 46:1, pp. 33–50.

Mroz, Thomas A. 1987. "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions." *Econometrica.* July, 55:4, pp. 765–99.

Newey, Whitney K. 1987. "Specification Tests for Distributional Assumptions in the Tobit Model." *Journal of Econometrics.* January/February, 34:1-2, pp. 125–45.

Newey, Whitney K., James L. Powell and James M. Walker. 1990. "Semiparametric Estimation of Selection Models: Some Empirical Results." *American Economic Review.* May, 80:2, pp. 324–28.

Powell, James L. 1984. "Least Absolute Deviations Estimation for the Censored Regression Model." *Journal of Econometrics.* July, 25:3, pp. 303–25.

Powell, James L. 1986a. "Censored Regression Quantiles." *Journal of Econometrics.* June, 32:1, pp. 143–55.

Powell, James L. 1986b. "Symmetrically Trimmed Least Squares Estimation for Tobit Models." *Econometrica.* November, 54:6, pp. 1435–60.

Powell, James L. 1994. "Estimation of Semiparametric Models," in *Handbook of Econometrics, Volume IV.* Robert F. Engle and Daniel L. McFadden, eds. Amsterdam: North Holland, pp. 2443–521.

Tobin, James. 1958. "Estimation of Relationships for Limited Dependent Variables." *Econometrica.* January, 26, pp. 24–36.