CENSORED DATA

Econometric Analysis of Cross Section and Panel Data, 2e
MIT Press
Jeffrey M. Wooldridge

1. Introduction
2. Binary Censory
3. Interval Coding
4. Censoring from Above and Below

1. INTRODUCTION

- We now consider estimation of standard (population) models where the response variable has been censored in some way. This could be severe censoring (we know only whether the outcome is above or below a threshold) or less severe (we observe an outcome until it is above a certain value).
- We obtain a random sample of units from the population, but we do not observe the entire range of the dependent variable.

- Later, we will consider the case where we observe nothing about one or more variables for part of the population.
- Some estimation methods are similar to those we have covered. However, here they are used to correct missing data problems and not to obtain better functional forms. Therefore, we now refer to “data censoring,” as distinct from, say, “corner solution” responses. (We could have both.)
- While data censoring produces a pile up at certain values, it is not for behavioral reasons. It is just the rule about how the data are collected.

2. BINARY CENSORING

• Binary censoring is an extreme form of data censoring. Start with a standard linear model for the population:

$$y = \mathbf{x}\boldsymbol{\beta} + u$$
$$E(u|\mathbf{x}) = 0$$

where $\mathbf{x}$ is $1 \times K$ with $x_1 = 1$, as usual.

• The variable y might be willingness to pay (WTP) for a proposed public project, $y = wtp$. When we draw family (say) i from the population, we would like to observe $(\mathbf{x}_i, wtp_i)$; if we did for all i, we would estimate $\boldsymbol{\beta}$ by OLS.

- WTP can be difficult to elicit, and reported amounts might be noisy.

Instead, suppose that each family is presented with a cost of the project, r_i. The household either says it is in favor of the project or not.

- Along with $\mathbf{x}_i$ and r_i, we observe the binary response

$$w_i = 1[y_i > r_i].$$

For now, assume that the chance that y_i equals r_i is zero.

- Importantly, $\boldsymbol{\beta}$ contains the partial effects of interest. We are interested in $E(wtp|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$, but $\boldsymbol{\beta}$ cannot be estimated by OLS because *wtp* is not observed.
- If we impose some strong assumptions on the underlying population and the nature of r_i, then we can proceed with maximum likelihood. Assume

$$u_i|\mathbf{x}_i, r_i \sim Normal(0, \sigma^2).$$

- So the underlying population model satisfies the classical linear model (CLM) assumptions.

• We also require that, for a random draw, r_i is independent of y_i conditional on $\mathbf{x}_i$, that is,

$$D(y_i|\mathbf{x}_i, r_i) = D(y_i|\mathbf{x}_i).$$

• This conditional independence assumption is satisfied if r_i is randomized – set independently of $(\mathbf{x}_i, y_i)$ – or if r_i is chosen as a function of $\mathbf{x}_i$. Or, conditional on $\mathbf{x}_i$, r_i is randomized.

• We can derive the expression for $P(w_i = 1|\mathbf{x}_i, r_i)$:

$$\begin{aligned} P(w_i = 1|\mathbf{x}_i, r_i) &= P(y_i > r_i|\mathbf{x}_i, r_i) = P[u_i/\sigma > (r_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma|\mathbf{x}_i, r_i] \\ &= 1 - \Phi[(r_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma] = \Phi[(\mathbf{x}_i\boldsymbol{\beta} - r_i)/\sigma] \\ &= \Phi[(\beta_1/\sigma) + (\beta_2/\sigma)x_{i2} + \ldots + (\beta_K/\sigma)x_{iK} + (-1/\sigma)r_i]. \end{aligned}$$

• Because we observe $(w_i, \mathbf{x}_i, r_i)$ for random draws from the population, probit of w_i on $\mathbf{x}_i, r_i$ consistently estimates $\boldsymbol{\beta}/\sigma$ as the vector of coefficients on $\mathbf{x}_i$ and $-1/\sigma$ as the coefficient on r_i. (In almost all applications $\mathbf{x}_i$ would include an intercept, and we allow that here.)

• Let $\boldsymbol{\gamma} = \boldsymbol{\beta}/\sigma$ and $\alpha = -1/\sigma$. After probit, $\hat{\beta}_j = -\hat{\gamma}_j/\hat{\alpha}$. Standard errors can be obtained via the delta method or bootstrapping.

- Costs of binary censoring can be severe. If we could observe y_i, specifying $E(y_i|\mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta}$ would suffice for consistent estimation of $\boldsymbol{\beta}$; in fact, we could just specify a linear projection and use OLS. More important to test homoskedasticity and normality assumptions with censoring compared with modeling a binary response (because the latter is a functional form issue).
- Can estimate parameters up to scale without placing strong restrictions on $D(u_i|\mathbf{x}_i, r_i)$, but then could only get relative effects.

• What if the linear model for WTP is unrealistic? If we could observe actual willingness to pay, we probably would observe $wtp \geq 0$ with $wtp = 0$ for some fraction of the population.

• What if the population model is Tobit?

$$y = wtp = \max(0, \mathbf{x}\boldsymbol{\beta} + u)$$
$$u|\mathbf{x}, r \sim Normal(0, \sigma^2)$$

• If we have binary censoring with $r_i > 0$ for all i, the estimation procedure is identical to that outlined for a linear model for wtp. Because we do not observe y_i, we cannot distinguish between a linear model and Tobit when all $r_i > 0$.

• Nevertheless, if we believe that y is zero for a nontrivial fraction of the population, any calculations should reflect that belief by using the type I Tobit formulas for estimating partial effects.

• If $y > 0$ always, a better model is

$$y = \exp(\mathbf{x}\boldsymbol{\beta} + u)$$

so $\log(y) = \mathbf{x}\boldsymbol{\beta} + u$. Now, can apply previous analysis with $\log(r_i)$ replacing r_i (assuming $r_i > 0$):

$$P(w_i = 1|\mathbf{x}_i, r_i) = \Phi[(\mathbf{x}_i\boldsymbol{\beta} - \log(r_i))/\sigma].$$

• If the correct population model is $y = \mathbf{x}\boldsymbol{\beta} + u$ (with u independent of $\mathbf{x}$ and normally distributed) then $P(w_i = 1|\mathbf{x}_i, r_i) = \Phi[(\mathbf{x}_i\boldsymbol{\beta} - r_i)/\sigma]$. If the correct population model is $\log(y) = \mathbf{x}\boldsymbol{\beta} + u$ then $P(w_i = 1|\mathbf{x}_i, r_i) = \Phi[(\mathbf{x}_i\boldsymbol{\beta} - \log(r_i))/\sigma]$. In principle, we can choose between these two models by comparing log likelihoods. (Does using r_i or $\log(r_i)$ produce the largest value of the log-likelihood function?)

• Might even use the Vuong model selection statistic.

3. INTERVAL CODING

- Now consider the standard linear model where the response variable is recorded as falling into certain intervals. The underlying variable, y, is continuous.
- We say we have **interval-coded data** (or **interval-censored data**). We are still interested in the population regression $E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$.
- Let $r_1 < r_2 < \ldots < r_J$ denote the *known* interval limits; these are specified as part of the survey design. For example, rather than asking individuals to report actual annual income, they report the interval that their income falls into.

• Under the CLM assumptions for y, we can estimate β and σ^2 by MLE. The structure of the problem is similar to the ordered probit model.

• Define, in the population, an ordered variable w:

$$
\begin{aligned}
w &= 0 \quad \text{if } y \leq r_1 \\
w &= 1 \quad \text{if } r_1 < y \leq r_2 \\
&\vdots \\
w &= J \quad \text{if } y > r_J
\end{aligned}
$$

• The probabilities $P(w = j|\mathbf{x})$ for $j = 0, 1, \ldots, J$ have the same form as ordered probit. The log likelihood for a random draw i is

$$\begin{aligned}\ell_i(\boldsymbol{\beta}, \sigma) &= 1[w_i = 0] \log\{\Phi[(r_1 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\} \\ &+ 1[w_i = 1] \log\{\Phi[(r_2 - \mathbf{x}_i\boldsymbol{\beta})/\sigma - \Phi[(r_1 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\} \\ &+ \ldots + 1[w_i = J] \log\{1 - \Phi[(r_J - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\}.\end{aligned}$$

• The maximum likelihood estimators, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$, are often called **interval regression** estimators, with the understanding that the underlying population distribution is homoskedastic normal.

- There are important differences from ordered probit. First, in ordered probit, we are interested in the discrete response variable, which is something like a credit rating. Here, we are interested in the underlying continuous variable *y*, which has quantitative meaning.
- In OP, the cut points are parameters to estimate, and the parameters β do not completely measure partial effects. With interval regression, the interval endpoints are given (or are themselves data), and β contains the partial effects of interest.

- As in the case of binary censoring, when we obtain the interval regression estimates, we interpret the $\hat{\boldsymbol{\beta}}$ *as if* we had been able to run the regression y_i on $\mathbf{x}_i$, $i = 1, \ldots, N$.
- Imposing the assumptions of the classical linear model allows us to estimate the parameters in the distribution $D(y|\mathbf{x})$, even though they data are interval-censored.
- Sometimes one sets the censored variable, w, to some value within the interval that contains y. For example, might set w to the midpoint of the interval that y falls into. (We need some other rule if $y < r_1$ or $y > r_J$.) If the definition of w determines the proper interval, the maximum likelihood estimators of $\boldsymbol{\beta}$ and σ will be the same.

• When w is defined to have the same units as y, it is tempting to ignore the grouping of the data and just to run an OLS regression of w_i on $\mathbf{x}_i$, $i = 1, \ldots, N$. Naturally, such a procedure is generally inconsistent for $\boldsymbol{\beta}$, but there are conditions under which the *ratios* of coefficients are consistent.

• Sometimes the interval limits change across i, which causes no problems if we assume the limits are exogenous in the following sense:

$$D(y_i|\mathbf{x}_i, r_{i1}, \ldots, r_{iJ}) = D(y_i|\mathbf{x}_i).$$

This includes the special case of binary censoring.

• In Stata, for each observation we specify variables *lower* and *upper*, and these determine the interval that y_i falls into. If y_i is below the smallest interval value, r_{i1}, $lower_i$ is set to missing. If y_i is above the largest interval value, r_{iJ}, $upper_i$ is set to missing.

```
intreg lower upper x1 x2 ... xK
```

EXAMPLE: Interval Coding for Net Financial Wealth

```
. use 401ksubs_intcode
. sum nettfa

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      nettfa |       975    14.87889    57.24609     -59.97   1134.098

. list nettfa lower upper in 1/10

     +-------------------------+
     | nettfa   lower   upper |
     |-------------------------|
  1. |  4.575       0       5 |
  2. |    154      25       . |
  3. |  18.45      10      25 |
  4. |   29.6      25       . |
  5. |      0       .       0 |
     |-------------------------|
  6. |  9.687       5      10 |
  7. |    .13       0       5 |
  8. | -21.02       .       0 |
  9. | 24.999      10      25 |
 10. |  2.999       0       5 |
     +-------------------------+

. * The intervals are y <= 0, 0 < y <= 5, 5 < y <= 10, 10 < y <= 25, y > 25.
```

```
. tab lower

      lower |
   interval |
  limit for |
     nettfa |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        264       41.31       41.31
          5 |         90       14.08       55.40
         10 |        133       20.81       76.21
         25 |        152       23.79      100.00
------------+-----------------------------------
      Total |        639      100.00

. * Note that 336 observations have nettfa <= 0.
```

```
. intreg lower upper inc incsq age agesq male e401k

Interval regression                               Number of obs   =        975
                                                  LR chi2(6)      =     274.90
Log likelihood = -1446.7593                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   .6263831   .0885805     7.07   0.000     .4527684    .7999977
       incsq |  -.0028399   .0008818    -3.22   0.001    -.0045683   -.0011116
         age |    -.13666   .3724458    -0.37   0.714    -.8666404    .5933204
       agesq |   .0056804   .0043551     1.30   0.192    -.0028554    .0142163
        male |  -.6346241    1.00675    -0.63   0.528    -2.607818     1.33857
       e401k |   6.577516   1.044148     6.30   0.000     4.531023     8.62401
       _cons |  -16.37731   7.627359    -2.15   0.032    -31.32666   -1.427963
-------------+----------------------------------------------------------------
    /lnsigma |   2.626524   .0368932    71.19   0.000     2.554215    2.698834
-------------+----------------------------------------------------------------
       sigma |   13.82563   .5100724                      12.8612    14.86239
------------------------------------------------------------------------------

  Observation summary:       336  left-censored observations
                               0     uncensored observations
                             152 right-censored observations
                             487       interval observations
```

```
. * How does this compare if we use the uncensored data in standard OLS
. * regression?

. reg nettfa inc incsq age agesq male e401k

      Source |       SS       df       MS              Number of obs =     975
-------------+------------------------------           F(  6,   968) =   15.74
       Model |  283681.821     6  47280.3036           Prob > F      =  0.0000
    Residual |  2908228.37   968  3004.36815           R-squared     =  0.0889
-------------+------------------------------           Adj R-squared =  0.0832
       Total |  3191910.19   974  3277.11518           Root MSE      =  54.812

------------------------------------------------------------------------------
      nettfa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   1.109313   .3183536     3.48   0.001     .4845707    1.734056
       incsq |  -.0041516   .0031799    -1.31   0.192    -.0103919    .0020888
         age |  -1.946907   1.337097    -1.46   0.146    -4.570849    .6770356
       agesq |   .0335103   .0156526     2.14   0.033     .0027935    .0642272
        male |   2.754853   3.618632     0.76   0.447    -4.346415    9.856121
       e401k |    7.51211   3.778518     1.99   0.047     .0970803    14.92714
       _cons |    3.15536   27.31521     0.12   0.908    -50.44849    56.75921
------------------------------------------------------------------------------
```

```
. * The OLS estimates are not especially close to the interval regression
. * estimates, quite likely because nettfa is
. * neither homoskedastic nor conditionally normally distributed. Of course,
. * the conditional mean may be misspecified, too. And it is just one sample of
. * data. But the interval regression estimates are less sensitive to extreme
. * values of nettfa. But then we are admitting the underlying distribution
. * cannot be homoskedastic normal.
```

Violations of the Assumptions

- Unlike with ordered probit or logit, allowing for a nonnormal distribution or heteroskedasticity is no longer just allowing for more flexible functional forms for the observed response. In OP and OL, we might still get decent estimates of partial effects even if we do not have the correct model.
- Now, we are worried that violations of the underlying CLM result in inconsistent estimation of $\boldsymbol{\beta}$, which is what we want to estimate. It makes to test for heteroskedasticity and even extend the estimation to allow for a flexible form, say $Var(y|\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\gamma})$.

- Nonnormality could be allowed, in principle, by using something like the Pearson family of distributions.
- Simple way to check robustness of the results: with many intervals, can combine some intervals and reestimate the parameters using interval regression. If the underlying population model holds, the estimates should differ only by sampling error.
- Using a “robust” option with the MLE estimation is an admission that the underlying population model is incorrect. Inference is robust, but it is inference on the wrong parameters.

```
. intreg lower upper inc incsq age agesq male e401k, robust

Interval regression                                   Number of obs   =        975
                                                      Wald chi2(6)    =     238.01
Log pseudolikelihood = -1446.7593                     Prob > chi2     =     0.0000

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   .6263831   .0927412     6.75   0.000     .4446136    .8081525
       incsq |  -.0028399   .0009492    -2.99   0.003    -.0047002   -.0009796
         age |    -.13666   .3872199    -0.35   0.724    -.8955971    .6222771
       agesq |   .0056804   .0045934     1.24   0.216    -.0033225    .0146834
        male |  -.6346241   1.018647    -0.62   0.533    -2.631136    1.361888
       e401k |   6.577516    1.06679     6.17   0.000     4.486647    8.668386
       _cons |  -16.37731   7.820532    -2.09   0.036    -31.70527   -1.049351
-------------+----------------------------------------------------------------
    /lnsigma |   2.626524   .0405579    64.76   0.000     2.547032    2.706016
-------------+----------------------------------------------------------------
       sigma |   13.82563   .5607383                      12.76915    14.96952
------------------------------------------------------------------------------

  Observation summary:       336  left-censored observations
                               0     uncensored observations
                             152 right-censored observations
                             487       interval observations
```

```
. reg nettfa inc incsq age agesq male e401k, robust

Linear regression                                      Number of obs =     975
                                                       F(  6,   968) =   16.00
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0889
                                                       Root MSE      =  54.812

------------------------------------------------------------------------------
             |               Robust
      nettfa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   1.109313   .3614363     3.07   0.002     .4000244    1.818602
       incsq |  -.0041516   .0037804    -1.10   0.272    -.0115704    .0032672
         age |  -1.946907   1.934885    -1.01   0.315    -5.743959    1.850146
       agesq |   .0335103   .0251003     1.34   0.182    -.0157469    .0827676
        male |   2.754853   3.777354     0.73   0.466    -4.657894     10.1676
       e401k |    7.51211   3.837661     1.96   0.051    -.0189842     15.0432
       _cons |    3.15536   32.73283     0.10   0.923    -61.08013    67.39084
------------------------------------------------------------------------------
```

Endogenous Explanatory Variables

- Because of the underlying normality assumption, we can use the Rivers-Vuong (1988) and Smith-Blundell (1986) control function approach to test and correct for endogeneity of continuous explanatory variables.
- The underlying model is the standard linear model

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1,$$

and we observe the censored variable, w_1.

• Given the linear reduced form $y_2 = \mathbf{z}\boldsymbol{\delta}_2 + v_2$, we proceed as before: just add the first-stage residuals, $\hat{v}_2$, to the interval regression model, along with $(\mathbf{z}_1, y_2)$. Of course, we are interested in α_1 and $\boldsymbol{\delta}_1$, along with the coefficient on $\hat{v}_2$ to determine whether y_2 is in fact endogenous.

• Unfortunately, such an approach only works when y_2 is not censored. It is very difficult to account for interval censoring of y_2 along with that for y_1.

• Allowing binary y_2 is possible but requires full MLE. No simple two-step methods.

Panel Data

• Modifying interval regression for linear, unobserved effects panel data models is straightforward, provided we are willing to rely on the Chamberlain-Mundlak device. We would write

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, \quad t = 1, \ldots, T$$
$$c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i$$

where all unobservables have normal distributions and the interval limits, $\{r_{itj} : j = 1, 2, \ldots, J\}$, can vary by i and t.

- Estimation can be carried out using the equation

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \psi_t + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i + u_{it},\ t = 1, \ldots, T$$

under serial independence in $\{u_{it} : t = 1, \ldots, T\}$ (a random effects structure). Easier to pooled interval regression.

- In Stata,

```
intreg lower upper x1 x2 ... xK x1bar ... xKbar
d2 ... dT, cluster(id)
```

- Ideally we would always observe y_{it} and then just use fixed effects, which would require neither the Mundlak assumption nor homoskedasticity and normality of $a_i + u_{it}$.

4. CENSORING FROM ABOVE AND BELOW

- Consider the case now where the underlying variable, y, follows a classical linear model, but it is censored from above, or *right censored*. For a random draw i,

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + u_i$$
$$w_i = \min(y_i, r_i)$$

where r_i is the right censoring variable that may change with i (but sometimes it does not).

- **Top coding** is a common example of right censoring, where, say, income or wealth is recorded up to a certain amount, and then we just know whether the value is above the amount.
- Censoring of duration data is also common.
- Left censoring or censoring from below occurs with minimum wages when we are interested in the value of marginal product.

- We are interested in features of the population distribution $D(y|\mathbf{x})$. In the current case, under $E(u|\mathbf{x}) = 0$, that means we are primarily interested in the population regression $E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$, which is completely characterized by $\boldsymbol{\beta}$.
- As we will see, we can estimate $\boldsymbol{\beta}$ under fewer assumptions if instead we assume $Med(u|\mathbf{x}) = 0$, in which case $Med(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$ (and this may or may not equal $E(y|\mathbf{x})$).
- Traditional parametric approaches specify $D(y|\mathbf{x})$ up to a (finite) set of unknown parameters, and then uses maximum likelihood estimation.

- Semiparametric approaches attempt to estimate $\boldsymbol{\beta}$ by placing few restrictions on $D(y|\mathbf{x})$.
- In both parametric and semiparametric approaches, some sort of conditional independence is assumed between u_i and r_i. The most restrictive form is $D(u_i|\mathbf{x}_i, r_i) = D(u_i|\mathbf{x}_i)$, but this can be relaxed in some cases.
- For semiparametric estimation, we can sometimes get by with $E(u_i|\mathbf{x}_i, r_i) = E(u_i|\mathbf{x}_i) = 0$ or $Med(u_i|\mathbf{x}_i, r_i) = Med(u_i|\mathbf{x}_i) = 0$.

• With parametric approaches, we model the entire population distribution, $D(y|\mathbf{x})$. Focus here on case where distribution of y_i is continuous. Let $f(y|\mathbf{x};\boldsymbol{\theta})$ denote the conditional density.

• Under $D(y_i|\mathbf{x}_i, r_i) = D(y_i|\mathbf{x}_i)$, can easily obtain the density of w_i conditional on $(\mathbf{x}_i, r_i)$ because, for $w < r_i$,

$P(w_i \leq w|\mathbf{x}_i, r_i) = P(y_i \leq w|\mathbf{x}_i) = F(w|\mathbf{x}_i;\boldsymbol{\theta})$, where $F(\cdot|\mathbf{x}_i;\boldsymbol{\theta})$ is the cdf of y_i conditional on $\mathbf{x}_i$. Therefore, the probability density of w_i given $(\mathbf{x}_i, r_i)$ is simply $f(w|\mathbf{x}_i;\boldsymbol{\theta})$ for $w < r_i$, that is, for values strictly less than the censoring point.

- Further, $P(w_i = r_i|\mathbf{x}_i, r_i) = P(y_i \geq r_i|\mathbf{x}_i, r_i) = 1 - F(r_i|\mathbf{x}_i; \boldsymbol{\theta})$.
- The probability density of w_i given $(\mathbf{x}_i, r_i)$ is

$$g(w|\mathbf{x}_i, r_i; \boldsymbol{\theta}) = [f(w|\mathbf{x}_i; \boldsymbol{\theta})]^{1[w<r_i]}[1 - F(r_i|\mathbf{x}_i; \boldsymbol{\theta})]^{1[w=r_i]}.$$

- The log likelihood function for a random draw i (where we do not bother to distinguish between the "true" value of theta and a generic value) is

$$\begin{aligned}\log[g(w_i|\mathbf{x}_i, r_i; \boldsymbol{\theta})] &= 1[w_i < r_i] \log[f(w_i|\mathbf{x}_i; \boldsymbol{\theta})] \\ &\quad + 1[w_i = r_i] \log[1 - F(r_i|\mathbf{x}_i; \boldsymbol{\theta})],\end{aligned}$$

and we sum this expression across all i to obtain the log likelihood for the entire sample.

- In the vast majority of cases, the conditions sufficient for MLE to be well behaved (consistent, $\sqrt{N}$-asymptotically normal) hold for censored estimation because the model $f(y|\mathbf{x};\boldsymbol{\theta})$ is smooth in $\boldsymbol{\theta}$.
- Interesting feature of the log likelihood: we only need to observe the censoring point, r_i, for censored observations. (We also need to know which observations are censored and which are not.) Useful for duration applications when the censoring value is reported only for observations that are actually censored.

• In the leading case, y follows a classical linear model in the population of interest, that is,

$$D(y|\mathbf{x}) = Normal(\mathbf{x}\boldsymbol{\beta}, \sigma^2),$$

which gives use the **censored normal regression model**.

• Easy to confuse censored normal regression with Type I Tobit, but they serve different purposes.

• The log likelihood for the censored normal regression model is

$$\begin{aligned} l_i(\boldsymbol{\theta}) &= 1[w_i < r_i] \log\{\sigma^{-1}\phi[(w_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\} \\ &+ 1[w_i = r_i] \log\{1 - \Phi[(w_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\}. \end{aligned}$$

• Easy to compute. Often a transformation, such as taking the natural log, is needed to make normality and homoskedasticity reasonable.

• Or, can apply the general censored density log likelihood directly.

• In Stata, the variable that is used is *w*, and an indicator is needed for whether the data are censored or not.

```
cnreg w x1 x2 ... xK, cen(cens)
```

where “cens” is the dummy variable equal to one of the observation is censored.

• For left censoring, *cens* is –1 for censored, zero for uncensored.

Allows both right and left censoring within the same data set.

EXAMPLE: Top Coding of Net Financial Wealth

```
. use 401ksubs_topcode
. * Censoring is at nettfa >= 50.

. sum nettfa nettfac

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      nettfa |       975    35.55443      81.435       -409   1003.126
     nettfac |       975    17.80794    26.73529       -409         50

. tab cens

      =1 if |
     nettfa |
   censored |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        751       77.03       77.03
          1 |        224       22.97      100.00
------------+-----------------------------------
      Total |        975      100.00

. * So about 23% of the observations are right censored.
```

```
. * First, linear regression using the actual data on nettfa:

. reg nettfa inc incsq age agesq male e401k, robust

Linear regression                               Number of obs =     975
                                                F(  6,   968) =   38.66
                                                Prob > F      =  0.0000
                                                R-squared     =  0.2588
                                                Root MSE      =  70.326

------------------------------------------------------------------------------
             |               Robust
      nettfa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |  -.4726794   .6361378    -0.74   0.458    -1.721048    .7756887
       incsq |   .0116774   .0054104     2.16   0.031     .0010599    .0222948
         age |  -1.527668   1.838973    -0.83   0.406      -5.1365    2.081165
       agesq |   .0354728   .0221191     1.60   0.109    -.0079341    .0788797
        male |  -9.332761   4.586691    -2.03   0.042    -18.33377   -.3317568
       e401k |   10.70226   4.673405     2.29   0.022     1.531087    19.87343
       _cons |   8.342726   33.27078     0.25   0.802    -56.94843    73.63389
------------------------------------------------------------------------------
```

```
. cnreg nettfac inc incsq age agesq male e401k, cen(cens)

Censored-normal regression                      Number of obs   =        975
                                                LR chi2(6)      =     301.64
                                                Prob > chi2     =     0.0000
Log likelihood = -3774.6932                     Pseudo R2       =     0.0384

------------------------------------------------------------------------------
     nettfac |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   .7225285   .1192285     6.06   0.000     .4885527    .9565043
       incsq |  -.0018362   .0008255    -2.22   0.026    -.0034562   -.0002162
         age |  -.1480192   .7230439    -0.20   0.838    -1.566932    1.270893
       agesq |   .0122743   .0081677     1.50   0.133    -.0037542    .0283028
        male |  -2.032747   3.123538    -0.65   0.515    -8.162425    4.096931
       e401k |   7.496106    2.00374     3.74   0.000     3.563936    11.42828
       _cons |  -31.34548   15.02683    -2.09   0.037    -60.83437   -1.856601
-------------+----------------------------------------------------------------
      /sigma |   28.67045   .7756753                      27.14825    30.19264
------------------------------------------------------------------------------
  Observation summary:         0  left-censored observations
                             751     uncensored observations
                             224 right-censored observations
```

- It is possible that the estimates from the censored regression are "better" (that is, closer to the population values). The uncensored estimates are likely very sensitive to extremely high values of wealth. Some researchers intentionally right censor variables such as wealth to avoid outliers, and then use censored normal regression. Unfortunately, important differences tell us the underlying CLM assumptions are false. We cannot know which estimates are "better."
- Remember, in most applications of censored regression, we do not have the luxury of running a regression with the actual y_i.

• As in the interval censoring case, violations of normality or homoskedasticity can be very costly, and it makes sense to test these assumptions (probably via the score principle). More flexible population distributions might be warranted. for example, allow asymmetry in $D(y|\mathbf{x})$ if, say, y is wealth, and also allow $Var(y|\mathbf{x})$ to be nonconstant, say $\exp(\mathbf{x}\boldsymbol{\gamma})$.

• For right censoring, can always choose to censor at a smaller value and check robustness of restimates.

• Example of how corner solution and censoring are different. Suppose a survey records annual family charitable contributions up to $10,000, but the amount is top coded. The reported variable, $w = \min(y, 10000)$ will clearly have a pile up at 10,000 due to the top coding. It is proper to say that charitable contributions is “censored from above at $10,000.”

- But we would also likely see a pile up at zero because some fraction of the population will have zero charitable contributions. An appropriate course of action is to treat charitable contributions in the population as a corner solution response, with a corner at zero. Unlike the censoring from above at $10,000, it makes no sense to say charitable contributions is also “censored from below at zero.” There is a corner at zero, but it is not due to data censoring.

- If charitable contributions follows a Tobit model in the population, but is also being top coded at $10,000, the estimation method is identical to two-limit Tobit, with limits zero and $10,000$. But partial effects are computed based on the standard Type I Tobit model because we are interested in $D(y|\mathbf{x})$, not $D(w|\mathbf{x})$.
- More subtle example: suppose by law individuals may contribute no more than 15% of their income to retirement plans. In the population, some individuals will contribute zero, some will contribute at the 15 percent upper limit, and many will contribute a percentage strictly between zero and 15. Might use a two-limit Tobit.

- If we are interested in the effect of explanatory variables on expected percentage contribution under the *current* legal regime, we would use the formulas for the two-limit Tobit model.
- However, one might want to know the effects of covariates on the contribution percentage in the absense of institutional constraints. Then, we would be back to the previous situation: the corner at zero is a corner that arises from utility maximization, but the corner at 15 is externally imposed.

Endogenous Explanatory Variables

- Easy to handle continuous endogenous explanatory variables using a control function approach.
- Suppose the population model is

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1$$
$$y_2 = \mathbf{z}\boldsymbol{\delta}_1 + v_2$$

where (u_1, v_2) is independent of $\mathbf{z}$ and bivariate normal. The variable y_1 is right censored.

- Can apply the two-step Smith-Blundell (1986) control function approach to account for the right censoring of y_1. The first step is OLS of y_2 on $\mathbf{z}$ using a random sample. The residuals, $\hat{v}_2$, are added to the censored normal regression in the second stage. Of course, because the underlying population model is linear, we are interested in α_1 and $\boldsymbol{\delta}_1$. Joint MLE is possible, too, and would be more efficient and avoid the problem of inference after two-step estimation.
- Bootstrap can be applied. We simply include the first-step estimation and censored normal estimation within each bootstrap iteration.

- As with all CF approaches, can allow a general functional form in $(\mathbf{z}_1, y_2)$, provided y_2 has a linear reduced form with the normality assumption given above.
- Allowing y_2 to be censored is more difficult, but there are some simple solutions after we cover general sample selection.

• With enough normality, the Chamberlain-Mundlak device can be used in the context of right and left censoring. The population model is the usual one:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}$$

• Right censoring is at r_{it}, which can change across i and t. Strict exogeneity assumptions, along with convenient distributional assumptions, are

$$D(u_{it}|\mathbf{x}_i, r_{i1}, \ldots, r_{iT}, c_i) = D(u_{it}) = Normal(0, \sigma_u^2), \quad t = 1, \ldots, T$$
$$D(c_i|\mathbf{x}_i, r_{i1}, \ldots, r_{iT}) = D(c_i|\mathbf{x}_i) = Normal(\psi + \bar{\mathbf{x}}_i\boldsymbol{\xi}, \sigma_a^2).$$

- As usual, these assumptions mean we can write

$$y_{it} = \psi + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + v_{it}, \; D(v_{it}|\mathbf{x}_i, \mathbf{r}_i) = Normal(0, \sigma_a^2 + \sigma_u^2),$$

where $\mathbf{r}_i = (r_{i1}, \ldots, r_{iT})$ is the vector of censoring values for unit i.

- Apply pooled censored normal regression, with censoring points r_{it}, and consistently estimate ψ, $\boldsymbol{\beta}$, $\boldsymbol{\xi}$, and $\sigma_v^2 = \sigma_a^2 + \sigma_u^2$. Because this is a partial likelihood method, we need to make inference robust to serial correlation.

• Generally, we cannot separately identify σ_a^2 and σ_u^2 unless we make a further assumption, such as $\{u_{it} : t = 1, \ldots, T\}$ is serially independent. Then can use a correlated random effects likelihood approach similar in structure to the CRE Tobit model.

• With the underlying population model linear, we are mainly interested in $\boldsymbol{\beta}$ and appropriate inference concerning $\boldsymbol{\beta}$.

• If the censoring points $\{r_{it} : t = 1, \ldots, T\}$ actually vary over time, could add the time average, $\bar{r}_i$, as a regressor to check for correlation with c_i. If $D(c_i|\mathbf{x}_i, r_i) = D(c_i|\bar{\mathbf{x}}_i, \bar{r}_i)$, adding $\bar{r}_i$ along with $\bar{\mathbf{x}}_i$ can actually solve the problem of the censoring value being related to c_i, on average.

• Something to think about. What if the model is $y_{it} = \mathbf{x}_{it}\mathbf{b}_i + c_i + u_{it}$ and we are interested in estimating $\boldsymbol{\beta} = E(\mathbf{b}_i)$. Even if we assume $\mathbf{b}_i$ is independent of $\mathbf{x}_i$, censoring has serious consequences. What approach might we take?

Censored Least Absolute Deviations

- How can we relax distributional assumptions? Focus on the cross section case.
- A very useful estimator is Powell's (1984) censored least absolute deviations (CLAD) estimator. Now we start with a linear model for the conditional median (which may or may not be the conditional mean).

$$Med(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}.$$

- The random sample consists of $(\mathbf{x}_i, r_i, w_i)$, $w_i = \min(y_i, r_i)$.
- Assume censoring is exogenous in the sense that

$$Med(y_i|\mathbf{x}_i, r_i) = Med(y_i|\mathbf{x}_i).$$

• LAD can be applied to the censoring case because the median passes throug the min function:

$$\begin{aligned} Med(w_i|\mathbf{x}_i, r_i) &= Med[\min(y_i, r_i)|\mathbf{x}_i, r_i] = \min[Med(y_i|\mathbf{x}_i, r_i), r_i] \\ &= \min[Med(y_i|\mathbf{x}_i), r_i] = \min(\mathbf{x}_i\boldsymbol{\beta}, r_i). \end{aligned}$$

• Now apply LAD to the expression for $Med(w_i|\mathbf{x}_i, r_i)$:

$$\min_{\mathbf{b}\in\mathbb{R}^K} \sum_{i=1}^{N} |w_i - \min(\mathbf{x}_i\mathbf{b}, r_i)|.$$

• As in the corner solution case, CLAD is generally consistent and $\sqrt{N}$-asymptotically normal.

- Subtle point concerning the CLAD estimator when applied to censored data is that it requires the censoring value, r_i, to be available even when the observation is not right censored. Not much of an issue in top-coding cases, especially when the same value is used. (For example, if wealth is top coded at $500,000, that information is known, and $r_i = 500,000$ for all i.)
- In some duration problems (later) only w_i is observed. That is, along with a censoring indicator, we observe either y_i or r_i, but not both. The previous MLE approach can be applied in situations where r_i is not always observed.

- A CLAD routine has been written for Stata. Need to install it. (Use "findit clad" in Stata.) Produces bootstrapped standard errors. Does not appear to allow censoring points to change with *i*, so better suited to corner solutions and fixed top or bottom coding. The command

```
clad y x1 ... xK, ul(10000) reps(500)
```

has top coding at $10,000$ and uses 500 bootstrap replications.

- Can estimate other quantiles, too, by specifying the quantile option `qu(.)`.

EXAMPLE: CLAD Estimation in Wealth Example

```
. use 401ksubs_topcode

. * First use LAD on uncensored data.

. qreg nettfa inc incsq age agesq male e401k

Median regression                                   Number of obs =       975
  Raw sum of deviations 36071.98 (about 11.347)
  Min sum of deviations 30122.46                    Pseudo R2     =    0.1649

------------------------------------------------------------------------------
      nettfa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |  -.2663949   .0697594    -3.82   0.000     -.403292   -.1294978
       incsq |   .0084499   .0004627    18.26   0.000     .0075418     .009358
         age |  -1.139927   .4452454    -2.56   0.011    -2.013685   -.2661698
       agesq |   .0204523   .0050035     4.09   0.000     .0106333    .0302713
        male |  -3.041986   1.919948    -1.58   0.113    -6.809725    .7257536
       e401k |   4.426652   1.225932     3.61   0.000     2.020861    6.832443
       _cons |   12.98194   9.290818     1.40   0.163    -5.250526    31.21441
------------------------------------------------------------------------------
```

```
. * Now use the top-coded variable, nettfac.

. clad nettfac inc incsq age agesq male e401k, ul(50) reps(500)

Initial sample size = 975
Final sample size = 895
Pseudo R2 = .20014184

Bootstrap statistics

Variable |   Reps   Observed       Bias   Std. Err.   [95% Conf. Interval]
-------+-------------------------------------------------------------------
   inc |    500  -.0972045   .1449356   .3886133   -.8607244  .6663155  (N)
       |                                           -.4645128  .8432961  (P)
       |                                           -.4985709  .8208665 (BC)
-------+-------------------------------------------------------------------
 incsq |    500   .0064427  -.0013644   .0039651   -.0013476   .014233  (N)
       |                                            -.002531  .0107853  (P)
       |                                           -.0023081   .011053 (BC)
-------+-------------------------------------------------------------------
   age |    500  -.9933933   .1564738   .5992066   -2.170672  .1838855  (N)
       |                                           -1.842628   .583323  (P)
       |                                           -2.147302  .1800248 (BC)
-------+-------------------------------------------------------------------
 agesq |    500   .0186491  -.0017135   .0069791     .004937  .0323612  (N)
       |                                           .0009719  .0290646  (P)
       |                                           .0044846  .0325657 (BC)
```

male	500	-3.716023	.2681802	1.950435	-7.5481	.116053	(N)
					-7.132865	.021205	(P)
					-7.732491	-.4130837	(BC)
e401k	500	4.762451	-.4150445	1.539838	1.737086	7.787817	(N)
					1.273426	7.443191	(P)
					1.920862	8.004879	(BC)
const	500	7.161489	-6.526609	17.97376	-28.15208	42.47506	(N)
					-43.34412	27.76345	(P)
					-33.87548	30.85902	(BC)

N = normal, P = percentile, BC = bias-corrected

• Something to think about. Suppose the model and data censoring mechanisms are

$$\begin{aligned} y_i &= a_i + \mathbf{x}_i \mathbf{b}_i \\ w_i &= \min(y_i, r_i) \\ D(a_i, \mathbf{b}_i | \mathbf{x}_i, r_i) &= D(a_i, \mathbf{b}_i). \end{aligned}$$

The population parameter are $\alpha = E(a_i)$ and $\boldsymbol{\beta} = E(\mathbf{b}_i)$. When does CLAD applied to

$$y_i = \alpha + \mathbf{x}_i \boldsymbol{\beta} + u_i$$

consistently estimate $\boldsymbol{\beta}$?