Multivariate Censored Regression

Karl G Jöreskog

20 January 2004

A censored variable has a large fraction of observations at the minimum or maximum. Because the censored variable is not observed over its entire range ordinary estimates of the mean and variance of a censored variable will be biased. Ordinary least squares (OLS) estimates of its regression on a set of explanatory variables will also be biased. These estimates are not consistent, *i.e.*, the bias does not get smaller when the sample size increase.

In a previous note *Censored Variables and Censored Regression* in this corner[1], I showed how one can estimate the mean and variance of a censored variable and the regression of a censored variable on a set of explanatory variables and avoid bias. The focus in this note was on *univariate* censored regression. I refer to the previous note as **UCR** (Univariate Censored Regression), for short.

There are two reasons why I return to the topic of censored variables and censored regression:

1. In **UCR** I used an example called **Affairs** where the dependent variable was ordinal rather than continuous. My reason for using this data set was that it had been used previously to illustrate censored regression by Fair (1978) and it was the only data I could get access to at that time (December 2002). In this note I provide an example (called **Maintenance**) of censored regression where the dependent variable is continuous as it should be.

2. I have now developed the procedure further to handle also **bivariate** and **multivariate** censored regression where several dependent censored variables are regressed on a set of explanatory variables. A special case of this is to estimate the mean vector and covariance matrix of several censored variables.

I begin with the **Maintenance** example in the next section. Section 2 describes the model for bivariate and multivariate censored regression and Section 4 gives an example of this. In Section 4 I also illustrate how one can estimate the mean vector and covariance matrix of a set of censored variables and I show how these can be used in a structural equation model (in this case a factor analysis model) in LISREL. In all the illustrative examples I use syntax files but one can also do these examples interactively, see Du Toit & Du Toit (2002).

1 Example: Maintenance

As part of the Household Income Project (HUS-projektet), Statistics Sweden surveyed Swedish people about the time they spent on repair and maintenance of their home. The data for this example was collected in 1984 at a time when tax rates were very high in Sweden. Some people

[1] available at *www.ssicentral.com/lisrel/column12.htm*

had extremely high marginal tax rates (as high as 85%) and one thought that this could have the effect that people would take unpaid leave from work and do the maintenance work themselves rather than to hire a professional to do it.

For the present illustrative purpose I use the following variables.

MAINTNCE Average number of minutes per day spent on maintenance or repair

AGE in years

HOUSE = 1, if respondent lives in a house, = 0, otherwise

RECHOUSE = 1, if respondent owns a recreation house, = 0, otherwise

CAR = 1, if respondent owns a car, = 0, otherwise

SCHOOLYR Number of years of schooling

INCOME Disposable income of respondent's household in 1000 SEK

MARGTAX Respondent's marginal tax rate in %

The data are given in the file **MAINTENANCE.PSF** in the subfolder **Censor**. To estimate the regression of the censored variable **MAINTNCE** on the all the other variables, one can use the following PRELIS syntax file (**MAINTENANCE.PR2**):

```
Censored Regression of MAINTENANCE
SY=MAINTENANCE.PSF
CR MAINTNCE ON AGE - MARGTAX
OU
```

The output reveals the following information (slightly edited):

```
 Total Sample Size =   2021

 Univariate Summary Statistics for Continuous Variables

 Variable      Mean  St. Dev.   Minimum Freq.  Maximum Freq.
 --------      ----  --------   ------- -----  ------- -----
 MAINTNCE    29.937    65.196     0.000  1329  666.429     1
      AGE    44.339    14.393    18.000     1   80.000     1
    HOUSE     0.608     0.488     0.000   793    1.000  1228
 RECHOUSE     0.236     0.425     0.000  1544    1.000   477
      CAR     0.836     0.370     0.000   331    1.000  1690
 SCHOOLYR    10.588     3.538     1.000     1   30.000     1
   INCOME    65.724    29.741     0.712     1  256.730     1
  MARGTAX    48.152    12.579    21.000     1   84.000     5

 Variable MAINTNCE is censored below.
 It has    1329 (65.76%) values =        0.000
```

Of the 2021 respondents there are 1329 people or nearly 66% who spent no time on maintenance. Hence, MAINTNCE is a highly censored variable. Note that there was one person who spent as much as 666 minutes on maintenance (there are 720 minutes in a 12 hour day). Furthermore,

- The average age is 44 years. The youngest person in the sample is 18 years and the oldest is 80 years.
- 1228 persons, or 61%, own a house.
- 477 persons, or 24%, own a second house.
- 1690 persons, or 84% own a car.
- The average number of school years is 10.6
- The average household income after tax is SEK 65.700 but the highest disposable income in the sample is SEK 256.700.
- The average marginal tax rate is 48%. There are 5 persons in the sample with the highest marginal tax rate 84%. The lowest marginal tax rate in the sample is 21%

The estimated censored regression is

```
MAINTNCE =  - 202.106 + 1.124*AGE + 50.517*HOUSE + 10.932*RECHOUSE
             (25.958)  (0.308)     (8.380)        (8.875)
             -7.786     3.653       6.028          1.232

            + 42.790*CAR - 4.130*SCHOOLYR + 0.357*INCOME + 0.838*MARGTAX
             (11.703)     (1.276)          (0.167)        (0.407)
              3.656       -3.236            2.137          2.060

            + Error
```

Taking all t-values greater than 2 in absolute value to be statistically significant, it is seen that all explanatory variables except RECHOUSE have a significant effect on MAINTNCE. MAINTNCE increases with AGE, HOUSE, CAR, INCOME, and MARGTAX, and decreases with SCHOOLYR. I leave it to the reader to consider the implications of these results. Here I am merely interested in comparing the result of this censored regression with what would be obtained if MAINTNCE was treated as uncensored. This can be obtained by changing the word `CR` to `RG` in the PRELISsyntax file **MAINTENANCE.PR2**. **MAINTENANCE.PR2** The OLS regression estimates are

```
MAINTNCE =  - 11.726 + 0.465*AGE + 10.636*HOUSE + 1.904*RECHOUSE + 13.017*CAR
             (9.173)  (0.114)     (3.094)        (3.488)          (4.103)
             -1.278    4.065       3.438          0.546            3.173

            - 1.629*SCHOOLYR + 0.164*INCOME + 0.202*MARGTAX
             (0.478)          (0.0651)       (0.156)
             -3.405            2.517          1.290

            + Error
```

It is seen that there are considerable differences in the estimates of the regression parameters but the variables AGE, HOUSE, CAR, SCHOOLYR, and INCOME are significant also in the OLS regression. Note however that MARGTAX is not statistically significant.

2 Multivariate Censored Regression

Consider two ordinal variables y_1 and y_2 with underlying continuous variables $y_1^\star$ and $y_2^\star$, respectively. The equations to be estimated are

$$y_1^\star = \alpha_1 + \boldsymbol{\gamma}_1'\mathbf{x} + z_1 \ , \tag{1}$$

$$y_2^\star = \alpha_2 + \boldsymbol{\gamma}_2'\mathbf{x} + z_2 \ , \tag{2}$$

where α_1 and α_2 are intercept terms, $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are vectors of regression coefficients, and z_1 and z_2 are error terms. It is assumed that z_1 and z_2 are independent of $\mathbf{x}$ and have a bivariate normal distribution with means zero and covariance matrix

$$\begin{pmatrix} \psi_1^2 & \\ \psi_{21} & \psi_2^2 \end{pmatrix} .$$

y_1 and y_2 are assumed to be censored both above and below, such that

$$\begin{aligned} y_1 &= c_{1L} \ \text{ if } \ y_1^\star \le c_{1L} \\ &= y_1^\star \ \text{ if } \ c_{1L} < y_1^\star < c_{1U} \\ &= c_{1U} \ \text{ if } \ y_1^\star \ge c_{1U} \ , \end{aligned}$$

$$\begin{aligned} y_2 &= c_{2L} \ \text{ if } \ y_2^\star \le c_{2L} \\ &= y_2^\star \ \text{ if } \ c_{2L} < y_2^\star < c_{2U} \\ &= c_{2U} \ \text{ if } \ y_2^\star \ge c_{2U} \ , \end{aligned}$$

where c_{1L}, c_{1U}, c_{2L}, and c_{2U} are constants.

Let

$$z_1^\star = (y_1^\star - \alpha_1 - \boldsymbol{\gamma}_1'\mathbf{x})/\psi_1 \tag{3}$$

$$z_2^\star = (y_2^\star - \alpha_2 - \boldsymbol{\gamma}_2'\mathbf{x})/\psi_2 \tag{4}$$

Then $z_1^\star$ and $z_2^\star$ have a standard bivariate normal distribution with correlation $\rho = \psi_{21}/\psi_1\psi_2$. The density of $z_1^\star$ and $z_2^\star$ is the density $f(u, v)$ at the end in the Appendix with

$$a = (c_{1L} - \alpha_1 - \boldsymbol{\gamma}_1'\mathbf{x})/\psi_1 \tag{5}$$

$$b = (c_{1U} - \alpha_1 - \boldsymbol{\gamma}_1'\mathbf{x})/\psi_1 \tag{6}$$

$$c = (c_{2L} - \alpha_2 - \boldsymbol{\gamma}_2'\mathbf{x})/\psi_2 \tag{7}$$

$$d = (c_{2U} - \alpha_2 - \boldsymbol{\gamma}_2'\mathbf{x})/\psi_2 \tag{8}$$

This density depends on the parameter vector

$$\boldsymbol{\theta} = (\alpha_1, \boldsymbol{\gamma}_1', \psi_1, \alpha_2, \boldsymbol{\gamma}_2', \psi_2, \rho)' \tag{9}$$

as well as on $\mathbf{x}$.

Let $(y_{1i}, y_{2i}, \mathbf{x}_i)$, $i = 1, 2, \ldots, N$ be a random sample of size N. The likelihood L_i of this observation can be computed using the results in the Appendix. The maximum likelihood estimate of $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ that maximizes $\ln L = \sum_{i=1}^N \ln L_i$. For practical purposes the following two-step procedure is more convenient and computationally much faster.

Step 1: Estimate α_1, $\boldsymbol{\gamma}_1$, and ψ_1 by univariate censored regression of y_1 on $\mathbf{x}$ and α_2, $\boldsymbol{\gamma}_2$, and ψ_2 by univariate censored regression of y_2 on $\mathbf{x}$ as described in **UCR**.

Step 2: For given estimates obtained in Step 1, estimate ρ by maximizing $\ln L$. This gives $\hat{\rho}$. Compute the estimate of ψ_{21} as $\hat{\psi}_{21} = \hat{\psi}_1\hat{\psi}_2\hat{\rho}$.

The multivariate case is handled in a similar way. In Step 1, we estimate each univariate censored regression, including the standard deviations of the error terms ψ_i. In Step 2 we estimate the correlation of the error terms for each pair of variables. The covariance matrix of the error terms can then be computed from these estimates.

3 PRELIS Implementation

The general syntax for multivariate censored regression is

```
CR Y-varlist ON X-varlist
```

where `Y-varlist` is a list of censored variables and `X-varlist` is a list of explanatory variables. The meaning of this is that *each* variable in `Y-varlist` is regressed on *all* variables in `X-varlist`. If `Y-varlist` and/or `X-varlist` contains a set of consecutive variables one can use `-` to denote a range of variables. For example, suppose that the data set consists of the six variables `Y1 Y2 Y3 X1 X2 X3`. Then

```
CR Y1-Y3 ON X1-X3
```

will perform multivariate censored regression of Y1, Y2, and Y3 on X1, X2, and X3. All variables in the `Y-varlist` and `X-varlist` must be continuous variables.

PRELIS can distinguish between univariate and multivariate censored regression. For example,

```
CR Y1-Y3 ON X1-X3
```

will do multivariate censored regression, whereas

```
CR Y1 on X1-X3
CR Y2 on X1-X3
CR Y3 on X1-X3
```

will do three univariate censored regressions. The difference is that in the second case PRELIS will estimate the error variances but not the error covariances, whereas in the first case PRELIS will estimate the whole error covariance matrix called *Residual Covariance Matrix*.

Each of `Y-varlist` and `X-varlist` may contain a subset of variables from the data set. Further explanation is needed if one has `MA=CM` on the `OU` line in a PRELIS syntax file, for this refers to the covariance matrix of all the variables in the data set. If `MA=CM` on the `OU` line and the PRELIS syntax file contains `CA`, `CB`, `CE`, or `CR` lines, then PRELIS will treat all variables in the data set as censored variables. The reason for this is that PRELIS cannot estimate the covariance between a censored variable and an uncensored variable. However, this does no harm because PRELIS can handle an uncensored variable, *i.e.*, a variable that has only one observation at the minimum and maximum, as a special case of a censored variable. For the same reason, if `Y` is an uncensored variable,

```
CR Y ON X-varlist
```

will give the same result as

```
RG Y ON X-varlist
```

namely OLS regression.

The `CR` command can also be used without the `ON X-varlist`, thus

```
CR Y-varlist
```

This gives estimates of the mean vector and covariance matrix of the variables in `Y-varlist`.

4 Example: TESTSCORE

For this example I use a subset of the variables in the file **readspel.psf** described in **UCR**. This file contains scores on 11 reading and spelling tests for 90 school children used in a study of the meta-phonological character of the Swedish language. The variables used here are V01, V02, V07, V21, and V23. They are available in the file **TESTSCORE.PSF**. The sample size is 90.

I estimate the bivariate censored regression of V21 and V23 on V01, V02, and V07 using the following PRELIS syntax file (**TESTSCORE1.PR2**).

```
SY=TESTSCORE.PSF
CR V21 V23 ON V01 V02 V07
OU
```

The output gives the following information about the distribution of the variables.

```
 Univariate Summary Statistics for Continuous Variables

 Variable      Mean  St. Dev.   Minimum Freq.  Maximum Freq.
 --------      ----  --------   ------- -----  ------- -----
      V01   21.789     7.856     0.000     2   30.000     6
      V02   14.622     7.048     0.000     2   28.000     5
      V07   11.489     3.069     0.000     1   20.000     2
      V21   15.022     2.998     4.000     1   17.000    41
      V23   12.578     3.402     3.000     3   16.000    21
```

If we regard a variable to be censored if it has more than one observation at the minimum or maximum, then we see that all five variables are censored. But V21 and V23 are more severely censored because they have many observations at the maximum.

The results of the censored regression is given in the output file as

```
 Variable V21 is censored above
 It has 41 (45.56%) values = 17.000

 Estimated Mean and Standard Deviation based on 90 complete cases.
 Mean = 16.692 (0.361)
```

```
Standard Deviation = 4.759 (0.023)

Estimated Censored Regression based on 90 complete cases.
     V21 = 8.057 + 0.125*V01 + 0.243*V02 + 0.198*V07 + Error, R = 0.420
          (1.745) (0.0838)    (0.0933)    (0.198)
           4.617   1.497       2.598       1.002

Variable V23 is censored below and above.
It has 3 ( 3.33%) values = 3.000 and 21 (23.33%) values = 16.000

Estimated Mean and Standard Deviation based on 90 complete cases.
Mean = 13.142 (0.286)
Standard Deviation = 4.397 (0.021)

Estimated Censored Regression based on 90 complete cases.
     V23 = 5.043 + 0.0144*V01 + 0.269*V02 + 0.333*V07 + Error, R = 0.395
          (1.537) (0.0774)     (0.0841)    (0.171)
           3.281   0.187        3.199       1.944

Residual Correlation Matrix

               V21         V23
          --------    --------
     V21       1.000
     V23       0.215       1.000

Residual Covariance Matrix

               V21         V23
          --------    --------
     V21      13.124
     V23       2.664      11.692
```

Only V02 has a significant effect. V01 and V07 are not significant for either V21 or V23. The residual correlation is 0.215. As will be seen in the next example the correlation between V21 and V23 is 0.437 if we treat both as censored.

Now consider a different problem. Suppose we want to do a factor analysis of the five variables V01, V02, V07, V21, and V23 and test the hypothesis that there is one common factor. In the first step I treat all the variables as multivariate censored and estimate the mean vector and covariance matrix of all the variables. This is done using the following PRELIS syntax file (**TESTSCORE2.PR2**).

```
SY=TESTSCORE.PSF
CE ALL
OU MA=CM CM=TESTSCORE.CM
```

If the sample size had been larger, I would also estimate the asymptotic covariance matrix. The line `CE ALL` declares all variables as censored. This is OK even if some variables are not censored.

As explained in **UCR**, an uncensored variable is simply treated as a special case of censoring when there is only one observation at the maximum or minimum. The results are:

Correlation Matrix

	V01	V02	V07	V21	V23
V01	1.000				
V02	0.761	1.000			
V07	0.641	0.595	1.000		
V21	0.407	0.443	0.355	1.000	
V23	0.429	0.512	0.442	0.437	1.000

Covariance Matrix

	V01	V02	V07	V21	V23
V01	71.906				
V02	48.894	57.348			
V07	17.023	14.128	9.821		
V21	16.408	15.958	5.297	22.644	
V23	16.000	17.040	6.088	9.144	19.331

Means

V01	V02	V07	V21	V23
22.032	14.727	11.510	16.692	13.142

Standard Deviations

V01	V02	V07	V21	V23
8.480	7.573	3.134	4.759	4.397

These results are summarized in the bottom half of Table 1, where the numbers in parentheses are the number of observations at the minimum and maximum. For comparison, the corresponding statistics for the case when all variables are treated as uncensored, are given in the top part of the table.

The one-factor model can be estimated using the SIMPLIS syntax file (**TESTSCORE3.SPL**):

```
Estimating One-Factor Model for Test Score Data
Observed Variables: V01 V02 V07 V21 V23
Covariance Matrix from File TESTSCORE.CM
Sample Size: 90
Latent Variable: VerbAbil
Relationships:
V01 - V23 = VerbAbil
```

Treating all Variables as Uncensored

Var	Min	Max	Mean	St Dev	Correlations				
V01	0(2)	30(6)	21.789	7.836	1.000				
V02	0(2)	28(5)	14.622	7.048	0.755	1.000			
V07	0(1)	20(2)	11.489	3.069	0.634	0.599	1.000		
V21	4(1)	17(41)	15.022	2.998	0.554	0.568	0.393	1.000	
V23	3(3)	16(21)	12.578	3.402	0.518	0.596	0.484	0.564	1.000

Treating all Variables as Censored

Var	Min	Max	Mean	St Dev	Correlations				
V01	0(2)	30(6)	22.032	8.480	1.000				
V02	0(2)	28(5)	14.727	7.573	0.761	1.000			
V07	0(1)	20(2)	11.510	3.134	0.641	0.595	1.000		
V21	4(1)	17(41)	16.692	4.759	0.407	0.443	0.355	1.000	
V23	3(3)	16(21)	13.142	4.397	0.429	0.512	0.442	0.437	1.000

Table 1: Basic Statistics Estimated in Two Different Ways

```
Path Diagram
Options: SC
End of Problem
```

This run generates a file**TESTSCORE3.FIT** containing many fit measures. I list some of them here.

```
                    Degrees of Freedom = 5
          Minimum Fit Function Chi-Square = 8.03 (P = 0.15)
Normal Theory Weighted Least Squares Chi-Square = 7.90 (P = 0.16)
          Estimated Non-centrality Parameter (NCP) = 2.90
        90 Percent Confidence Interval for NCP = (0.0 ; 14.73)

                  Minimum Fit Function Value = 0.090
          Population Discrepancy Function Value (F0) = 0.033
         90 Percent Confidence Interval for F0 = (0.0 ; 0.17)
       Root Mean Square Error of Approximation (RMSEA) = 0.081
       90 Percent Confidence Interval for RMSEA = (0.0 ; 0.18)
         P-Value for Test of Close Fit (RMSEA < 0.05) = 0.26
```

For further information about these and other fit measures see my previous note *On Chi-Squares for the Independence Model and Fit Measures in LISREL*[2]. All fit measures indicate that the one-factor model fits reasonably well. The standardized factor loadings are

```
          VerbAbil
          --------
V01         0.86
```

[2]available at *www.ssicentral.com/lisrel/column14.htm*

```
V02       0.87
V07       0.71
V21       0.51
V23       0.57
```

References

Du Toit, M. & Du Toit, S. (2002) *Interactive LISREL: User's Guide*. Lincolnwood: Scientific Software International.

Fair, R. (1978) A theory of extramarital affairs. *Journal of Political Economy*, **86**, 45–61.

Appendix

Let $U^\star$ and $V^\star$ have a bivariate standard normal distribution with correlation ρ. The density and distribution functions are

$$\phi_2(u,v;\rho) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} e^{-\frac{1}{2(1-\rho^2)}(u^2-2\rho uv+v^2)} \ ,$$

and

$$\Phi_2(u,v;\rho) = \int_{-\infty}^{u}\int_{-\infty}^{v} \phi_2(s,t;\rho)dsdt \ ,$$

respectively. Note that

$$\phi_2(u,v;\rho) = \phi(u)\phi(\frac{v-\rho u}{\sqrt{1-\rho^2}}) = \phi(v)\phi(\frac{u-\rho v}{\sqrt{1-\rho^2}}) \ .$$

Furthermore, let

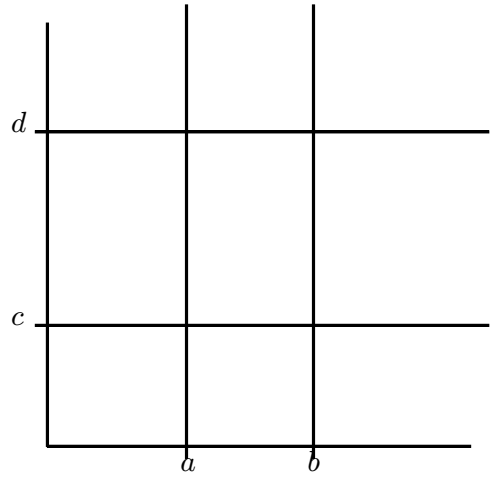
$$\Delta\Phi_2(a,b,c,d;\rho) = \int_a^b\int_c^d \phi_2(s,t;\rho)dsdt \ .$$

Let U and V be censored both above and below, such that

$$\begin{aligned} U &= a \text{ if } U^\star \le a \\ &= U^\star \text{ if } a < U^\star < b \\ &= b \text{ if } U^\star \ge b \ , \end{aligned}$$

$$\begin{aligned} V &= c \text{ if } V^\star \le c \\ &= V^\star \text{ if } c < V^\star < d \\ &= d \text{ if } V^\star \ge d \ , \end{aligned}$$

where a, b, c, and d are constants. The density of U and V is divided into nine regions as shown in following figure.



The density of U and V is

$$\begin{array}{rcll} f(u,v) & = & \Phi_2(a,c,\rho) & \text{if} \quad U=a\,, V=c \\ & = & \alpha\phi(u)\Phi(\frac{c-\rho u}{\alpha}) & \text{if} \quad a<U<b\,, V=c \\ & = & \Delta\Phi(b,+\infty,-\infty,c;\rho) & \text{if} \quad U=b\,, V=c \\ & = & \alpha\phi(v)\Phi(\frac{a-\rho v}{\alpha}) & \text{if} \quad U=a\,, c<V<d \\ & = & \phi_2(u,v;\rho) & \text{if} \quad a<U<b\,, c<V<d \\ & = & \alpha\phi(v)[1-\Phi(\frac{b-\rho v}{\alpha})] & \text{if} \quad U=b\,, c<V<d \\ & = & \Delta\Phi_2(-\infty,a,d,+\infty;\rho) & \text{if} \quad U=a\,, V=d \\ & = & \alpha\phi(u)[1-\Phi(\frac{d-\rho u}{\alpha})] & \text{if} \quad a<U<b\,, V=d \\ & = & \Delta\Phi_2(b,+\infty,d,+\infty;\rho) & \text{if} \quad U=b\,, V=d \end{array}$$

where $\alpha=\sqrt{1-\rho^2}$ and $\Phi(x)$ is the standard normal distribution function.