Censored Variables and Censored Regression

Karl G Jöreskog

3 December 2002

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 become smaller when the sample size increases. This note explains how maximum likelihood estimates can be obtained using PRELIS 2.53. The maximum likelihood estimates are consistent, *i.e.*, the bias is small in large samples.

Examples of censored variables are

Econometrics The first example of censored regression appears to be that of Tobin (1958). This is a study of the demand for capital goods such as automobiles or major household appliances. Households are asked whether they purchased such a capital good in the last 12 months. Many households report zero expenditures. However, among those households that made such an expenditure, there will be a wide variation in the amount of money spent.

Greene (2000) p. 205 lists several other examples of censored variables:

1. The number of extramarital affairs (Fair, 1978)
2. The number of hours worked by a woman in the labor force (Quester & Greene, 1982)
3. The number of arrests after release from prison (Witte, 1980)
4. Vacation expenditures (Melenberg & van Soest, 1996)

Biomedicine or Epidemiology Censored variables are common in biomedical, epidemiological, survival and duration studies. For example, in a five year follow-up study, time to death or time to recovery after surgery, medical treatment or diagnosis, are censored variables if, after five years, many patients are still alive or not yet recovered.

Educational Testing If a test is too easy or too difficult there will be a large number of examinees with all items or no items correctly answered.

In econometrics dependent censored variables are often called limited dependent variables and censored regression is sometimes called the tobit model[1].

[1]This model was first discussed by Tobin (1958). Goldberger (1964, p. 253) gave it this nickname (Tobin's probit) in analogy with the probit model.

1 Censored Normal Variables

A censored variable can be defined as follows. Let $y^\star$ be normally distributed with mean μ and variance σ^2. An observed variable y is censored below if

$$\begin{aligned} y &= c \text{ if } y^\star \leq c \\ &= y^\star \text{ otherwise} , \end{aligned}$$

where c is a given constant. This is illustrated in Figure 1.

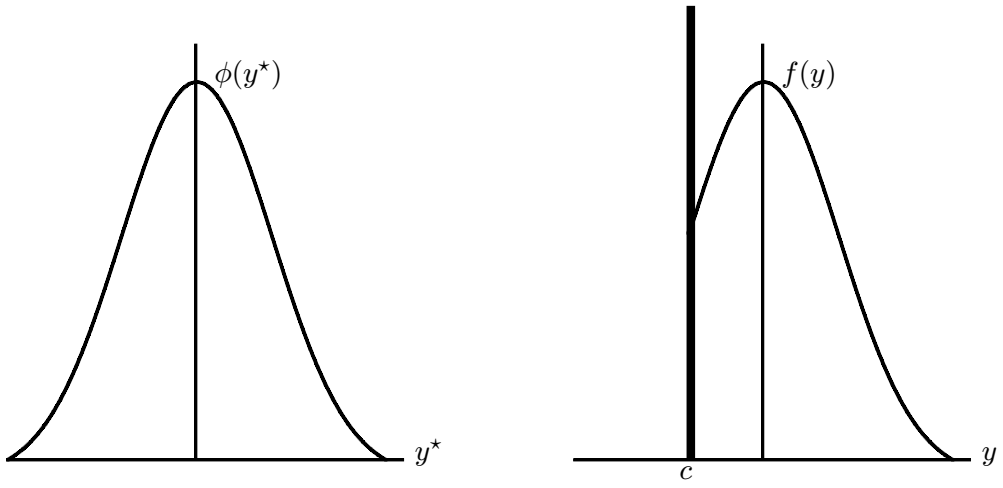

Figure 1: Normal Variable $y^\star$ and Censored Variable y

Let ϕ and Φ be the density and distribution functions of the standard normal distribution. The density function of y is

$$f(y) = \left[\Phi\left(\frac{c-\mu}{\sigma}\right)\right]^j \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2}\right]^{1-j} , \tag{1}$$

where $j = 1$ if $y = c$ and $j = 0$, otherwise. This may be regarded as a mixture of a binary and a normal variable.

The mean and variance of y are (see, *e.g.*, Greene, 2000, p. 907)

$$E(y) = \pi c + (1-\pi)(\mu + \lambda\sigma) , \tag{2}$$

$$Var(y) = (1-\pi)[(1-\delta) + (\alpha-\lambda)^2\pi]\sigma^2 , \tag{3}$$

where

$$\alpha = \frac{c-\mu}{\sigma} , \tag{4}$$

$$\pi = \Phi(\alpha) , \tag{5}$$

$$\lambda = \frac{\phi(\alpha)}{1-\Phi(\alpha)} , \tag{6}$$

$$\delta = \lambda^2 - \lambda\alpha . \tag{7}$$

A consequence of (2) and (3) is that the sample mean and variance of y are not consistent estimates of μ and σ^2. The bias of the mean $E(y) - \mu$ as a function of c is shown in Figure 2 for $\mu = 0$ and $\sigma = 1$.

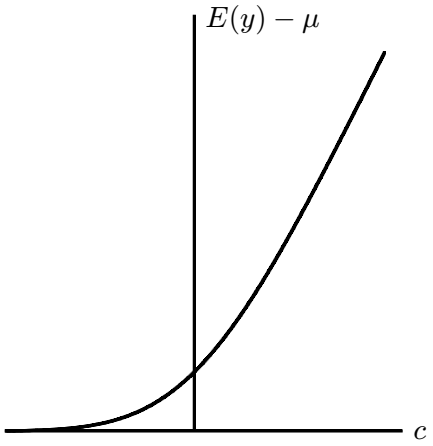

Figure 2: Bias $E(y) - \mu$ as a function of c

An observed variable y is censored above if

$$\begin{aligned} y &= c \text{ if } y^\star \geq c \\ &= y^\star \text{ otherwise} , \end{aligned}$$

A variable can be censored both below and above. In all three cases, the mean μ and variance σ^2 can be estimated by maximum likelihood as described in the Appendix.

2 Censored Normal Regression

Consider estimating the regression equation

$$y^\star = \alpha + \boldsymbol{\gamma}'\mathbf{x} + z , \tag{8}$$

where α is an intercept term and $\boldsymbol{\gamma}$ is a vector of regression coefficients on the explanatory variables $\mathbf{x}$. The error term z is assumed to be normally distributed with mean 0 and variance ψ^2. If $y^\star$ is observed as y throughout its entire range, the estimation of (8) is straightforward. However, if the observed variable y is censored below or above, then ordinary least squares (OLS) estimates of y on $\mathbf{x}$ are biased. However, α and $\boldsymbol{\gamma}$ can be estimated by maximum likelihood as described in the Appendix and these maximum likelihood estimates are unbiased in large samples.

3 PRELIS Implementation

The features described here have been implemented in PRELIS 2.53, released with LISREL 8.53 in December 2002.

I illustrate the case of 3 censored variables and 4 explanatory variables. Let `Y1 Y2 Y3` be the names of the censored variables and let `X1 X2 X3 X4` be the names of the regressors.

Censored regression of y_1 on x_1, x_2, x_3, x_4 is obtained by the PRELIS command

```
CR Y1 on X1 X2 X3 X4
```

One can select any subset of y-variables and any subset of x-variables to be included in the equation. Thus, one can obtain the regression for all the censored variables simultaneously. For example, the command

```
CR Y1 Y2 Y3 on X1 X2 X3 X4
```

will estimate three regression equations, namely the regression equation of each y_i on all x_j. Note the word `on` (or `ON`) separating the censored variables from the regressors.

One can have several `CR` commands in the same input file. For example,

```
CR Y1 on X1
CR Y1 on X1 X2
CR Y1 on X1 X2 X3
CR Y1 on X1 X2 X3 X4
```

will introduce one regressor at a time in the order x_1, x_2, x_3, x_4.

General rules:

- All y and x-variables appearing on CR lines must be declared continuous before the first `CR` command, or else they must have at least 16 different values.
- A censored regression equation can only be estimated from raw data. If there are missing values in the data, the estimation will be based on all cases without missing values on the variables included in the regression equation. Thus, the number of cases used depends on the variables included in the equation. Alternatively, one can impute missing values by multiple imputation[2] before estimating the regression equation.
- If several regression equations are estimated, the regression residuals in each equation are assumed to be uncorrelated.

4 Examples

I give four examples of censored regression. The starting point for each of these is a PRELIS system file (PSF file), see du Toit & du Toit (2001, pp. 384–385). You can create a PSF file from data in other formats such as Excel, SPSS and SAS. If the data is in text (ASCII) format it can be read into the PSF file directly from the Windows interface, or alternatively the PSF file can be created by running a simple PRELIS command file of the form

```
da ni=number-of-variables
la
labels-of-variables
ra=filename !filename for the data in text (ASCII) form
ou ra=filename.PSF
```

If the data requires a variable format statement, include the format as the first line(s) in the ASCII data file.

[2]See Du Toit & Du Toit, (2001, pp. 165–170). Note that this assumes multivariate normality and missingness at random.

4.1 Examples 1 and 2

I begin with two small examples based on generated data. Both of them consists of two variables y and x and a sample of 10000 observations. In the first example, y is censored below and in the second example y is censored both below and above. The data are in files **cr1.psf** and **cr2.psf**, respectively. Both of these were generated from the true regression line $E(y^\star \mid x) = 5 + x$. In the first example y was censored below at 3 and in the second example y was censored above at 6, in addition.

To estimate the censored regression of y on x for the first example, use the following PRELIS command file (**cr1.pr2**).

```
sy=ex1.psf
cr Y on X
ou
```

The output file reveals the following.

```
 Total Sample Size =  10000

 Univariate Summary Statistics for Continuous Variables

 Variable     Mean  St. Dev.   T-Value  Skewness  Kurtosis  Minimum Freq.  Maximum Freq.
 --------     ----  --------   -------  --------  --------  ------- -----  ------- -----
        Y    5.654     2.689   210.237     0.745    -0.439    3.000  3041   16.372     1
        X   -0.008     2.913    -0.278    -0.010    -1.222   -4.999     1    4.999     1

 Test of Univariate Normality for Continuous Variables

              Skewness         Kurtosis      Skewness and Kurtosis

 Variable Z-Score P-Value   Z-Score P-Value   Chi-Square P-Value

        Y  27.295   0.000    -8.957   0.000      825.252   0.000
        X  -0.402   0.687   -24.942   0.000      622.268   0.000

 Variable Y is censored below.
 It has    3041 (30.41%) values =        3.000

 Estimated Mean and Standard Deviation based on    10000 complete cases.
 Mean =                       4.963(       0.039)
 Standard Deviation =         3.609(       0.033)
```

The table `Univariate Summary Statistics for Continuous Variables` shows that the smallest value of y in the sample is 3.000 and this occurs 3041 times. A consequence of this is that y is highly non-normal. This table also gives the mean and standard deviation of y as 5.654 and 2.689, respectively. These are *wrong* values. Taking censoring into account gives the maximum likelihood estimates of the mean and standard deviation of y as 4.963 and 3.609, respectively. If we take the standard error estimates of these estimates into account, it is clear that the ordinary mean and standard deviations are far outside of the 95% confidence interval. This demonstrates that the ordinary mean and standard deviation can be considerably biased if a variable is highly censored.

The estimated regression equation is given in the output as

```
Estimated Censored Regression based on    10000 complete cases.

       Y = 5.022   + 0.995*X + Error, R2 = 0.542
          (0.0227)  (0.00828)
           221.344   120.238

Error Variance = 3.889
```

This can be compared with the OLS regression obtained by replacing `CR` by `RG` in **cr1.pr2**. The estimated OLS regression is

```
       Y = 5.659   + 0.737*X + Error, R2 = 0.637
          (0.0162)  (0.00556)
           349.494   132.576

Error Variance = 2.622
```

Hence, it is clear that the OLS estimates of the intercept, the regression coefficient, and the error variance are highly biased.

Note that PRELIS automatically determines

- The sample size.
- The smallest value of y and the degree of censoring.
- The maximum likelihood estimates of the mean μ and standard deviation σ of y and their asymptotic standard error estimates.
- The maximum likelihood estimates of the intercept α, the regression coefficient γ and the error standard deviation ψ with their asymptotic standard error estimates and t-values.

For example 2 run the PRELIS command file **cr2.pr2**. This is the same as **cr1.pr2** but with the name of the data file changed to `cr2.psf`. The output shows

```
Variable Y is censored below and above.
It has 3041 (30.41%) values = 3.000 and 4037 (40.37%) values = 6.000

Estimated Mean and Standard Deviation based on    10000 complete cases.
Mean =                      5.026(       0.045)
Standard Deviation =        3.972(       0.065)

Estimated Censored Regression based on    10000 complete cases.

       Y = 5.016   + 0.987*X + Error, R2 = 0.231
          (0.0257)  (0.0143)
           195.361   68.979

Error Variance = 3.770
```

The y-variable is censored both below and above. 30.41% of the observations are censored below and 40.37% are censored above. Thus there are only 29.22% uncensored observations. A consequence of this is that ordinary estimates of the mean and standard deviation of y are even more biased than in the previous example. Similarly, one can demonstrate that the OLS estimates are also more severely biased than in the previous example.

4.2 Example 3: Affairs

Fair (1978) published an example of censored regression. His study concerns the number of extramarital affairs and its determinants. From a large data set, the results of which were published in the July 1970 issue of *Psychology Today*, Fair extracted 601 observations on men and women who were then currently married for the first time. His data set consisting of 15 variables is available on the Internet at *http://fairmodel.econ.yale.edu/rayfair/workss.htm*. For present purposes the following nine variables are used.

GENDER	0 = female, 1 = male
AGE	in years
YEARS	number of years married[3]
CHILDREN	0 = no, 1 = yes
RELIGIOUS	1 = anti, ..., 5 = very religious
EDUCATION	number of years of schooling, 9 = grade school, 12 = high school, 20 = PhD
OCCUPATION	Hollingshead scale of 1 to 7
HAPPINESS	self rating of quality of marriage, 1 = very unhappy, ..., 5 = very happy
AFFAIRS	number of affairs in the past year, 1, 2, 3, 4–10 coded as 7, monthly, weekly, and daily coded as 12

I have selected these nine variables from the data set on the Internet. A text (ASCII) file is given in **affairs.dat** and the corresponding PSF file is given in **affairs1.psf**. Some of these variables are ordinal. For the purpose of data screening they are all declared ordinal in **affairs1.psf**. A data screening is obtained by running the following simple PRELIS command file (**affairs0.pr2**).

```
!Data Screening of Affairs Data
sy=affairs1.psf
ou
```

The data screening reveals interesting characteristics of the distribution of the variables.

AGE	Frequency	Percentage
17.5	6	1.0
22	117	19.5
27	153	25.5
32	115	19.1
37	88	14.6
42	56	9.3
47	23	3.8
52	21	3.5
57	22	3.7

YEARS	Frequency	Percentage
0.1	11	1.8
0.4	10	1.7
0.8	31	5.2
1.5	88	14.6
4	105	17.5
7	82	13.6
10	70	11.6
15	204	33.9

[3] I don't know how this was coded. In addition to integer values, there are values 0.12, 0.42, 0.75, and 1.50 on this variable.

GENDER	Frequency	Percentage
0	315	52.4
1	286	47.6

CHILDREN	Frequency	Percentage
0	171	28.5
1	430	71.5

RELIGIOU	Frequency	Percentage
1	48	8.0
2	164	27.3
3	129	21.5
4	190	31.6
5	70	11.6

EDUCATIO	Frequency	Percentage
9	7	1.2
12	44	7.3
14	154	25.6
16	115	19.1
17	89	14.8
18	112	18.6
20	80	13.3

OCCUPATI	Frequency	Percentage
1	113	18.8
2	13	2.2
3	47	7.8
4	68	11.3
5	204	33.9
6	143	23.8
7	13	2.2

HAPPINES	Frequency	Percentage
1	16	2.7
2	66	11.0
3	93	15.5
4	194	32.3
5	232	38.6

AFFAIRS	Frequency	Percentage
0	451	75.0
1	34	5.7
2	17	2.8
3	19	3.2
7	42	7.0
12	38	6.3

It is seen that 75% of the respondents report having no extramarital affairs. Thus, the dependent variable is highly censored at 0. It is also seen that 38 persons (6.3%) report having extramarital affairs monthly, weekly, or daily (code 12). To estimate censored regression equations, all variables in the equation must be continuous. This can be specified by including the line

```
co all
```

in the PRELIS command file. Furthermore, I have changed one value on AFFAIRS from 12.000 to 12.001, since I want to treat AFFAIRS as censored below[4]. Otherwise, PRELIS will treat AFFAIRS as censored both below and above. This change of a single data value has no other effects whatsoever. The data file with this change is given in **affairs2.psf** in which all variables are declared continuous.

One can use long names of variables but PRELIS truncates all variable names to 8 characters. To estimate the censored regression of AFFAIRS using all the other variables as explanatory variables, use the following PRELIS command file (**affairs1.pr2**).

```
Censored Regression of Affairs
sy=affairs2.psf
cr AFFAIRS on GENDER - HAPPINESS
ou
```

The output reveals the following.

[4]I do this primarily because this is the way the AFFAIRS variable has been treated earlier, see Fair (1978) and Greene (2000). In Table 1 I report the results for the case when AFFAIRS is treated as censored both below and above.

```
Total Sample Size =    601

Univariate Summary Statistics for Continuous Variables

Variable     Mean  St. Dev.   T-Value  Skewness  Kurtosis  Minimum Freq.  Maximum Freq.
--------     ----  --------   -------  --------  --------  ------- -----  ------- -----
  GENDER    0.476     0.500    23.340     0.097    -1.997    0.000   315    1.000   286
     AGE   32.488     9.289    85.742     0.889     0.232   17.500     6   57.000    22
   YEARS    8.178     5.571    35.984     0.078    -1.571    0.125    11   15.000   204
CHILDREN    0.715     0.452    38.843    -0.958    -1.087    0.000   171    1.000   430
RELIGIOU    3.116     1.168    65.440    -0.089    -1.008    1.000    48    5.000    70
EDUCATIO   16.166     2.403   164.959    -0.250    -0.302    9.000     7   20.000    80
OCCUPATI    4.195     1.819    56.519    -0.741    -0.776    1.000   113    7.000    13
HAPPINES    3.932     1.103    87.374    -0.836    -0.204    1.000    16    5.000   232
 AFFAIRS    1.456     3.299    10.820     2.347     4.257    0.000   451   12.001     1

Variable AFFAIRS is censored below.
It has     451 (75.04%) values =        0.000

Estimated Mean and Standard Deviation based on      601 complete cases.
Mean =                      -6.269(       0.774)
Standard Deviation =         9.420(       0.644)

Estimated Censored Regression based on      601 complete cases.

 AFFAIRS = 7.609 + 0.946*GENDER - 0.193*AGE + 0.533*YEARS + 1.019*CHILDREN
          (3.936) (1.071)        (0.0816)    (0.148)       (1.289)
           1.933   0.883         -2.362       3.610         0.791

           - 1.699*RELIGIOU + 0.0254*EDUCATIO + 0.213*OCCUPATI
            (0.409)          (0.229)           (0.324)
            -4.159            0.111             0.658

           - 2.273*HAPPINES + Error, R2 = 0.0203
            (0.419)
            -5.431

Error Variance = 69.239
```

As judged by the t-values, the effects of GENDER, CHILDREN, EDUCATION, and OCCUPATION are not statistically significant. The number of extramarital affairs seem to increase with number of years of marriage, and decrease when age, religiousness, and happiness increase.

To illustrate the concept of a fit file, I consider entering one variable at a time in the regression equation. The order of variables corresponds to the size of the t-values in the previous run. The PRELIS command file is (**affairs2.pr2**)

```
Sequential Censored Regression of Affairs
sy=affairs2.psf
cr AFFAIRS on HAPPINESS
cr AFFAIRS on HAPPINESS RELIGIOUS
cr AFFAIRS on HAPPINESS RELIGIOUS YEARS
cr AFFAIRS on HAPPINESS RELIGIOUS YEARS AGE
```

```
cr AFFAIRS on HAPPINESS RELIGIOUS YEARS AGE GENDER
cr AFFAIRS on HAPPINESS RELIGIOUS YEARS AGE GENDER CHILDREN
cr AFFAIRS on HAPPINESS RELIGIOUS YEARS AGE GENDER CHILDREN OCCUPATION
cr AFFAIRS on HAPPINESS RELIGIOUS YEARS AGE GENDER CHILDREN OCCUPATION EDUCATION
ou xu
```

The `xu` on the `ou` line tells PRELIS to skip the results of univariate data screening in the output. For each regression equation estimated, PRELIS produces a fit file with the same name as the PRELIS command file but with suffix FIT. In this case, the fit file **affairs2.fit** is:

```
 Variable       -2lnL  Chi-square   df  Covariates
 --------  ----------  ----------   --  ----------
  AFFAIRS    1168.622      45.172    1  HAPPINES
  AFFAIRS    1156.071      57.723    2  HAPPINES RELIGIOU
  AFFAIRS    1141.496      72.298    3  HAPPINES RELIGIOU YEARS
  AFFAIRS    1137.129      76.665    4  HAPPINES RELIGIOU YEARS AGE
  AFFAIRS    1134.924      78.870    5  HAPPINES RELIGIOU YEARS AGE GENDER
  AFFAIRS    1134.426      79.368    6  HAPPINES .....................CHILDREN
  AFFAIRS    1133.793      80.000    7  HAPPINES ........................ OCCUPATI
  AFFAIRS    1133.781      80.013    8  HAPPINES ............................ EDUCATIO
```

The first column gives the name of the dependent variable. The second column gives the minimum value of the deviance $-2 \ln L$. This value decreases as one adds explanatory variables in the regression equation. The third column gives a chi-square value for testing the hypothesis that all regression coefficients are zero. Thus, this chi-square is a test of the hypothesis that none of the covariates has any effect. The fourth column gives the degrees of freedom of this chi-square. This equals the number of explanatory variables. The remaining columns gives the names of the explanatory variables included in the equation.

The first line indicates that the effect of HAPPINESS is statistically significant. The second line shows that chi-square increases considerably when RELIGIOUS is added into the equation. The chi-square difference 57.723 – 45.172 = 12.551 is a chi-square with 1 degree of freedom for testing the hypothesis the regression coefficient of RELIGIOUS is zero when HAPPINESS is included in the equation. This hypothesis is rejected. Thus, both variables must be included. Note that this chi-square difference is the same as the difference between the two deviances in reverse order, *i.e.*, 1168.622 – 1156.071 = 12.551. In a similar way, one concludes that the effect of YEARS is also statistically significant when this variable is added into the equation that already includes HAPPINESS and RELIGIOUS, the chi-square difference being 72.298 – 57.723 = 14.575. By further adding AGE into the equation, the chi-square difference is 4.367 which is statistically significant at the 5% level but not at the 1% level. Adding more variables into the equation does not improve the fit significantly. For example, if all eight variables are included one obtains a chi-square of 80.013 which can be compared with 76.665 obtained if only the first four variables are included. The chi-square difference 3.348 with 4 degrees of freedom is far from statistically significant. The conclusion is that the equation which includes HAPPINESS, RELIGIOUS, YEARS, and AGE is the best equation for predicting AFFAIRS. The estimated regression coefficients for this equation, with their standard error estimates in parentheses, are given in column 2 of Table 1.

In this analysis, the AFFAIRS variable is treated as continuous. This means that the values 0, 1, 2, 3, 7, and 12 that this variable takes are assumed to be numbers on an interval scale. One could also treat AFFAIRS as censored both below and above, see file **affairs3.pr2**. Another way

is to treat AFFAIRS as an ordinal variable with six categories[5] or as a binary variable where 0 is used in one category and all values larger than 0 are used in the other category. One can then use logistic or probit regression. To do so with PRELIS, use **affairs1.psf**, include the lines

```
or AFFAIRS
co GENDER - HAPPINESS
```

and replace `cr` (censored regression) by `lr` (logistic regression) or `pr` (probit regression). To use AFFAIRS as a binary variable, include the line

```
re AFFAIRS old=1-12 new=1
```

see files **affairs4.pr2** - **affairs7.pr2**. The results obtained from these files are not directly comparable because the variable $y^\star$ is scaled differently for different methods. However, they can be made comparable by multiplying the regression coefficients and their standard error estimates by a suitable scale factor. The probit regressions (columns 4 and 6 in Table 1) are scaled such that the error variance is 1 (standard parameterization, see Jöreskog, 2002). Using this as a standard, we must scale the other solutions by $1/\hat{\psi}$. If AFFAIRS is treated as censored below (column 2 in Table 1), this scale factor is $1/\sqrt{69.031} = 0.12036$, see file **affairs2.out**. If AFFAIRS is treated as censored below and above (column 3 in Table 1), this scale factor is $1/\sqrt{123.39} = 0.09002$, see file **affairs3.out**. For the logistic regressions (columns 5 and 7 in Table 1) the scale factor is $\sqrt{3}/\pi = 0.55133$, because the variance of the standard logistic distribution is $\pi^2/3$. The t-values are not affected by this scaling. After this scaling the results are shown in Table 1.

Table 1: Estimated Regression Coefficients with Different Methods

	Censored		Ordinal		Binary	
Variable	Below	&Above	Probit	Logit	Probit	Logit
HAPPINESS	-.273(.049)	-.281(.052)	-.284(.049)	-.278(.047)	-.270(.052)	-.254(.049)
RELIGIOUS	-.207(.049)	-.208(.051)	-.209(.049)	-.200(.048)	-.187(.052)	-.181(.049)
YEARS	.065(.016)	.067(.017)	.067(.016)	.067(.016)	.058(.017)	.056(.016)
AGE	-.019(.009)	-.020(.010)	-.021(.010)	-.023(.009)	-.020(.010)	-.019(.010)

It is seen that all methods give similar results. For most practical purposes these results are the same. The binary methods does not make use of all information in the AFFAIRS variable. Nevertheless, the results are very close to the other methods which make use of all available information. Which of these methods should be used to estimate the model? The question arises because of the nature of the dependent variable. This is not fully continuous (as if it were an amount of money spent). Neither is it fully ordinal as if the responses were classified as never, sometimes, and often. It is somewhere in between. Censored regression is a method for continuous variables and probit and logit regressions are methods for ordinal variables.

4.3 Example 4: Reading and Spelling Tests

The file **readspel.psf** 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. It is of particular interest

[5]It may be better to use the exact counts (number of extramarital affairs) and treat this as a Poisson variable, but such data is not available to me.

to predict one of these tests, V23, using the other 10 variables as predictors and to determine which of these variables are the best predictors. However, a data screening of **readspel.psf** reveals that V23 may be censored both below and above. Hence, we must use censored regression to estimate the prediction equation. The PRELIS command file is (**readspel1.pr2**)

```
Eleven Reading and Spelling Tests
sy=READSPEL.PSF
co all
cr V23 on V01 - V22 V24 V25
ou
```

The output shows

```
 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(       0.483)
 Standard Deviation =         4.397(       0.412)

 Estimated Censored Regression based on       90 complete cases.

      V23 =  - 1.724 - 0.0731*V01 + 0.122*V02 + 0.384*V07 - 0.129*V08
              (2.414) (0.0815)     (0.0900)    (0.262)     (0.213)
              -0.714  -0.897        1.359       1.463      -0.605

             + 0.0954*V09 + 0.117*V10 + 0.208*V21 + 0.0679*V22 + 0.0276*V24
              (0.0688)     (0.0838)    (0.177)     (0.149)      (0.163)
               1.388        1.403       1.178       0.456        0.170

             + 0.208*V25 + Error, R2 = 0.374
              (0.158)
               1.319

 Error Variance = 9.848
```

None of the predictors are statistically significant. However, in terms of the t-values, the most important predictors seem to be V02, V07, V09, and V10. Using only these as predictors (see file **readspel2.pr2**) gives the following prediction equation.

```
      V23 = 1.652 + 0.210*V02 + 0.283*V07 + 0.0775*V09 + 0.169*V10
           (1.843) (0.0663)    (0.151)     (0.0656)     (0.0801)
            0.896   3.170       1.877       1.181        2.114

             + Error, R2 = 0.328

 Error Variance = 10.392
```

Here it is seen that the effects V02 and V10 are statistically significant.

Appendix: Computational Notes

The estimation of a censored regression equation is described in Chapter 6 of Maddala (1983) for the case of a variable that is censored below at 0. The development outlined here covers the cases

when the observed variable y is censored below, censored above, censored both below and above, and not censored at all. It also covers the case when there are no regressors in which case one can estimate the mean and standard deviation of y.

Changing notation slightly from Section 2, consider the estimation of the regression equation

$$y^\star = \alpha^\star + \boldsymbol{\gamma}^{\star\prime}\mathbf{x} + z \ , \tag{9}$$

where $\alpha^\star$ is the intercept term, $\boldsymbol{\gamma}^\star$ is the vector of regression coefficients, and $\mathbf{x}$ the regressors. The error term z is assumed to be normally distributed with mean 0 and variance $\psi^{\star 2}$. If there are no regressors, the second term in (9) is not included.

The observed variable

$$\begin{aligned} y &= c_1 \text{ if } y^\star \le c_1 \\ &= y^\star \text{ if } c_1 < y^\star < c_2 \\ &= c_2 \text{ if } y^\star \ge c_2 \ , \end{aligned}$$

where c_1 and c_2 are constants. If y is censored below set $c_2 = +\infty$. If y is censored above set $c_1 = -\infty$. If y is not censored set both $c_1 = -\infty$ and $c_2 = \infty$.

Let $(y_i, \mathbf{x}_i)$ be the observed values of y and $\mathbf{x}$ of case i in a random sample of N independent observations. The likelihood of $(y_i, \mathbf{x}_i)$ is

$$L_i = \left[\Phi(\frac{c_1 - \alpha^\star - \boldsymbol{\gamma}^{\star\prime}\mathbf{x}_i}{\psi^\star})\right]^{j_{1i}} \left[\frac{1}{\sqrt{2\pi}\psi^\star} e^{-\frac{1}{2}(\frac{y_i - \alpha^\star - \boldsymbol{\gamma}^{\star\prime}\mathbf{x}_i}{\psi^\star})^2}\right]^{1-j_{1i}-j_{2i}} \left[1 - \Phi(\frac{c_2 - \alpha^\star - \boldsymbol{\gamma}^{\star\prime}\mathbf{x}_i}{\psi^\star})\right]^{j_{2i}} \ ,$$

where $j_{i1} = 1$ if $y = c_1$ and $j_{i1} = 0$ otherwise and $j_{i2} = 1$ if $y = c_2$ and $j_{i2} = 0$ otherwise. Note that j_{i1} and j_{i2} cannot be 1 simultaneously.

The log likelihood is

$$\ln L = \sum_{i=1}^{N} \ln L_i \ .$$

This is to be maximized with respect to the parameter vector $\boldsymbol{\theta}^{\star\prime} = (\alpha^\star, \boldsymbol{\gamma}^{\star\prime}, \psi^\star)$.

First and second derivatives of $\ln L$ with respect to $\boldsymbol{\theta}^\star$ are very complicated. They will be considerably simplified and the maximization of $\ln L$ will be considerably more efficient if another parameterization due to Tobin (1958) is used.

This parameterization uses the parameter vector $\boldsymbol{\theta}' = (\alpha, \boldsymbol{\gamma}', \psi)$ instead of $\boldsymbol{\theta}^\star$, where $\alpha = \alpha^\star/\psi^\star$, $\boldsymbol{\gamma} = \boldsymbol{\gamma}^\star/\psi^\star$, and $\psi = 1/\psi^\star$.

Multiplication of (9) by $\psi = 1/\psi^\star$ gives

$$\psi y^\star = \alpha + \boldsymbol{\gamma}'\mathbf{x} + v \ , \tag{10}$$

where $v = \psi z = z/\psi^\star$ which is $N(0,1)$. Then

$$y = c_1 \leftrightarrow \ y^\star \le c_1 \leftrightarrow \psi y^\star \le \psi c_1 \leftrightarrow v \le \psi c_1 - \alpha - \boldsymbol{\gamma}'\mathbf{x} \ ,$$

$$y = c_2 \leftrightarrow \ y^\star \ge c_2 \leftrightarrow \psi y^\star \ge \psi c_2 \leftrightarrow v \ge \psi c_2 - \alpha - \boldsymbol{\gamma}'\mathbf{x} \ .$$

Hence the likelihood L_i becomes

$$L_i = \left[\Phi(\psi c_1 - \alpha - \boldsymbol{\gamma}'\mathbf{x}_i)\right]^{j_{1i}} \left[\frac{1}{\sqrt{2\pi}}\psi e^{-\frac{1}{2}(\psi y_i - \alpha - \boldsymbol{\gamma}'\mathbf{x}_i)^2}\right]^{1-j_{1i}-j_{2i}} \left[1 - \Phi(\psi c_2 - \alpha - \boldsymbol{\gamma}'\mathbf{x}_i)\right]^{j_{2i}} \ .$$

Let

$$\delta_i = \psi y_i - \alpha - \boldsymbol{\gamma}'\mathbf{x}_i \ . \tag{11}$$

Then $\ln L_i$ becomes

$$\ln L_i = -\ln\sqrt{2\pi} + (1 - j_{1i} - j_{2i})(\ln\psi - \frac{1}{2}\delta_i^2) + j_{1i}\ln\Phi(\delta_i) + j_{2i}\ln[1 - \Phi(\delta_i)] \ . \tag{12}$$

First and second derivatives of $\ln L_i$ are straightforward by noting that $\partial\delta_i/\partial\alpha = -1$, $\partial\delta_i/\partial\boldsymbol{\gamma} = -\mathbf{x}_i$, and $\partial\delta_i/\partial\psi = y_i$. Furthermore, $\Phi'(t) = \phi(t)$, $\phi'(t) = -t\phi(t)$ and if $A(t) = \phi(t)/\Phi(t)$, then $A'(t) = -A(t)[t + A(t)] = B(t)$, say.

Omitting index i, the required derivatives are

$$\partial\ln L/\partial\alpha \ = (1 - j_1 - j_2)\delta - j_1 A(\delta) + j_2 A(-\delta)$$

$$\partial\ln L/\partial\boldsymbol{\gamma} = (1 - j_1 - j_2)\delta\ \mathbf{x} - j_1 A(\delta)\mathbf{x} + j_2 A(-\delta)\mathbf{x}$$

$$\partial\ln L/\partial\psi = (1 - j_1 - j_2)(1/\psi - \delta y) + j_1 A(\delta)y - j_2 A(-\delta)y$$

$$\partial^2\ln L/\partial\alpha\partial\alpha = -(1 - j_1 - j_2) + j_1 B(\delta) + j_2 B(-\delta)$$

$$\partial^2\ln L/\partial\boldsymbol{\gamma}\partial\alpha = -(1 - j_1 - j_2)\mathbf{x} + j_1 B(\delta)\mathbf{x} + j_2 B(-\delta)\mathbf{x}$$

$$\partial^2\ln L/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}' = -(1 - j_1 - j_2)\mathbf{x}\mathbf{x}' + j_1 B(\delta)\mathbf{x}\mathbf{x}' + j_2 B(-\delta)\mathbf{x}\mathbf{x}'$$

$$\partial^2\ln L/\partial\psi\partial\alpha = (1 - j_1 - j_2)y - j_1 B(\delta)y - j_2 B(-\delta)y$$

$$\partial^2\ln L/\partial\psi\partial\boldsymbol{\gamma}' = (1 - j_1 - j_2)y\mathbf{x}' - j_1 B(\delta)y\mathbf{x}' - j_2 B(-\delta)y\mathbf{x}'$$

$$\partial^2\ln L/\partial\psi\partial\psi = -(1 - j_1 - j_2)(1/\psi^2 + y^2) + j_1 B(\delta)y^2 + j_2 B(-\delta)y^2$$

Maximizing $\ln L$ is equivalent to minimizing the fit function $F(\boldsymbol{\theta}) = -\ln L$. Let $\mathbf{g}(\boldsymbol{\theta}) = \partial F/\partial\boldsymbol{\theta}$ be the gradient vector and $\mathbf{H}(\boldsymbol{\theta}) = \partial^2 F/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ be the Hessian matrix. Amemiya (1973) proved that $\mathbf{H}$ is positive definite everywhere.

The fit function $F(\boldsymbol{\theta})$ is minimized using a Newton-Raphson procedure which converges very fast. The starting values $\boldsymbol{\theta}_0$ are the parameters estimated by OLS. Successive estimates are obtained by the formula

$$\boldsymbol{\theta}_{s+1} = \boldsymbol{\theta}_s - \mathbf{H}_s^{-1}\mathbf{g}_s \ , \tag{13}$$

where $\mathbf{g}_s = \mathbf{g}(\boldsymbol{\theta}_s)$ and $\mathbf{H}_s = \mathbf{H}(\boldsymbol{\theta}_s)$.

Let $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\boldsymbol{\gamma}}, \hat{\psi})$ be the maximum likelihood estimates of $\boldsymbol{\theta}$. The asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ is $\mathbf{E} = \mathbf{H}^{-1}(\boldsymbol{\theta})$ evaluated at the true parameter $\boldsymbol{\theta}$. Since the transformation from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^\star$ is one-to-one, the maximum likelihood estimates of $\boldsymbol{\theta}^\star$ is $\hat{\boldsymbol{\theta}}^\star = (\hat{\alpha}^\star, \hat{\boldsymbol{\gamma}}^\star, \hat{\psi}^\star)$, where $\hat{\alpha}^\star = \hat{\alpha}/\hat{\psi}$, $\hat{\boldsymbol{\gamma}}^\star = \boldsymbol{\gamma}/\hat{\psi}$, and $\hat{\psi}^\star = 1/\hat{\psi}$.

To obtain the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}^\star$, we evaluate the matrix $\partial\boldsymbol{\theta}^\star/\partial\boldsymbol{\theta}'$. This is

$$\partial\boldsymbol{\theta}^\star/\partial\boldsymbol{\theta}' = (1/\psi^2)\begin{pmatrix} \psi & \mathbf{0}' & -\alpha \\ \mathbf{0} & \psi\mathbf{1} & \boldsymbol{\gamma} \\ 0 & \mathbf{0}' & -1 \end{pmatrix} = \mathbf{A}(\boldsymbol{\theta}), \text{say} \ , \tag{14}$$

where $\mathbf{0}$ and $\mathbf{1}$ are column vectors of zeros and ones, respectively. The asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}^\star$ is $\mathbf{AEA}'$, where $\mathbf{A}$ and $\mathbf{E}$ are evaluated at the true parameter values. An estimate of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}^\star$ is $\mathbf{AEA}'$ obtained by using the estimated parameter values in $\mathbf{A}$ and $\mathbf{E}$. Asymptotic standard error estimates of the parameter estimates are obtained as the square roots of the diagonal elements of this matrix.

References

Amemiya, T. (1973) Regression analysis when the dependent variable is truncated normal. *Econometrica*, 41, 997–1016.

Du Toit, M., Du Toit, S. (2001) *Interactive LISREL: User's Guide*. Chicago: Scientific Software International.

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

Goldberger, A.S. (1964) *Econometric theory*. New York: Wiley.

Greene, W.H. (2000) *Econometric Analysis*. Fourth Edition. London: Prentice Hall International.

Jöreskog, K.G. (2002) Structural equation modeling with ordinal variables using LISREL. Available at *http://www.ssicentral.com/lisrel/corner.htm*.

Maddala, G.S. (1983) *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press.

Melenberg, B. & van Soest, A. (1996) Parametric and semi-parametric modeling of vacation expenditures. *Journal of Applied Econometrics*, 11, 59–76.

Quester, A. & Greene, W. (1982) Divorce risk and wifes' labor supply behavioral. *Social Science Quarterly*, 16–27.

Tobin J. (1958) Estimation of relationships for limited dependent variables. *Econometrica*, 26, 24–36.

Witte, A. (1980) Estimating an economic model of crime with individual data. *Quarterly Journal of Economics*, 94,57–84.