Multivariate censored regression

Contents

1. Introduction

Consider two ordinal variables y_1 and y_2 with underlying continuous variables y_1^* and y_2^* respectively. We need to estimate the equations

$$y_1^* = \alpha_1 + \boldsymbol{\gamma}_1'\mathbf{x} + z_1$$
$$y_2^* = \alpha_2 + \boldsymbol{\gamma}_2'\mathbf{x} + z_2,$$

where the αs represent intercept terms and the $\boldsymbol{\gamma}$s vectors of regression coefficients. The error terms are represented by z_1 and z_2. It is assumed that the error terms are independent of the regressors $\mathbf{x}$ and that they have a bivariate normal distribution with means equal to zero and a covariance matrix of the form

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

We further assume that the variables y_1 and y_2 are censored both above and below. In other words,

$$
\begin{aligned}
y_1 &= c_{1L} \; if \; y_1^* \le c_{1L} \\
&= y_1^* \; if \; c_{1L} < y_1^* < c_{1U} \\
&= c_{1U} \; if \; y_1^* \ge c_{1U}
\end{aligned}
$$

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

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

We now define two variables z_1^* and z_2^* as

$$
\begin{aligned}
z_1^* &= \left(y_1^* - \alpha_1 - \boldsymbol{\gamma}_1' \mathbf{x} \right) / \psi_1 \\
z_2^* &= \left(y_2^* - \alpha_2 - \boldsymbol{\gamma}_2' \mathbf{x} \right) / \psi_2
\end{aligned}
$$

The variables z_1^* and z_2^* have a standard bivariate normal distribution with correlation $\rho = \psi_{21} / \psi_1 \psi_2$.

As a first step, we estimate each univariate censored regression, including the standard deviations of the error terms ψ_i. Next, 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.

Censored regression is requested in PRELIS through the use of the CR command.

2. Example of bivariate censored regression

The data used here 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 to predict some of these tests using the other variables as predictors and to determine which of these variables are the best predictors. These data are given in **testscore.lsf**, of which the first few rows are shown below. This file can be found in the **MVABOOK Examples\Chapter 2** folder.

TESTSCORE.lsf

	V01	V02	V07	V21	V23
1	26.00	16.00	12.00	14.33	15.44
2	17.00	12.00	11.00	14.33	10.44
3	15.00	11.00	11.00	13.33	14.44
4	30.00	28.00	13.00	15.33	15.44
5	21.00	16.00	12.00	13.33	15.44
6	26.00	22.00	12.00	8.33	7.44
7	27.00	17.00	12.00	14.33	11.44
8	19.00	8.00	9.00	13.33	10.44
9	27.00	19.00	12.00	15.33	14.44
10	16.00	10.00	12.00	13.33	8.44
11	26.00	15.00	12.00	15.33	15.44
12	28.00	16.00	13.00	15.33	13.44
13	27.00	20.00	12.00	14.33	14.44
14	5.00	2.00	12.00	10.33	5.44
15	14.00	12.00	12.00	6.33	13.44
16	22.00	9.00	12.00	12.33	10.44
17	23.00	13.00	12.00	15.33	12.44
18	30.00	28.00	20.00	15.33	15.44
19	27.00	21.00	14.00	15.33	14.44
20	12.00	13.00	8.00	14.33	11.44

The covariance matrix for these variables, along with means and standard deviations, are given below. These are obtained by using the **Data Screening** option from the **Statistics** menu.

Covariance Matrix

	V01	V02	V07	V21	V23
V01	61.719				
V02	41.829	49.676			
V07	15.284	12.962	9.421		
V21	13.038	11.997	3.618	8.988	
V23	13.854	14.288	5.051	5.751	11.573

Correlation Matrix

	V01	V02	V07	V21	V23
V01	1.000				
V02	0.755	1.000			
V07	0.634	0.599	1.000		
V21	0.554	0.568	0.393	1.000	
V23	0.518	0.596	0.484	0.564	

Total Variance = 141.377 Generalized Variance = 257403.392

Largest Eigenvalue = 110.203 Smallest Eigenvalue = 3.958

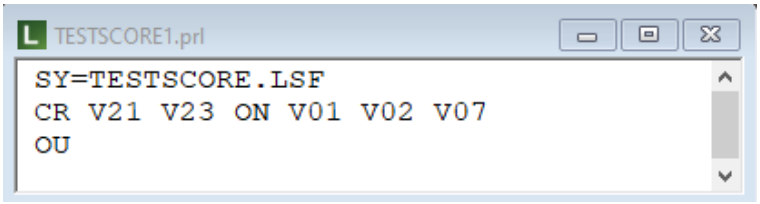
Condition Number = 5.277

Means

V01	V02	V07	V21	V23
21.789	14.622	11.489	13.352	12.013

```
 Standard Deviations

              V01        V02        V07        V21        V23
         --------   --------   --------   --------   --------
            7.856      7.048      3.069      2.998      3.402
```

The PRELIS syntax contained **in testscore1.prl** can be used to request a bivariate censored regression of the variables V21 and V23 on V01, V02 and V07:



The univariate summary statistics for the variables is given first:

```
Univariate Summary Statistics for Continuous Variables

 Variable      Mean   St. Dev.   Skewness   Kurtosis   Minimum Freq.   Maximum Freq.
 --------      ----   --------   --------   --------   ------- -----   ------- -----
      V01    21.789      7.856     -1.117      0.333     0.000     2    30.000     6
      V02    14.622      7.048     -0.173     -0.568     0.000     2    28.000     5
      V07    11.489      3.069     -0.175      2.673     0.000     1    20.000     2
      V21    13.352      2.998     -1.956      3.424     2.330     1    15.330    41
      V23    12.013      3.402     -1.141      0.735     2.436     3    15.436    21
```

We see that all the variables in the data have multiple values at either the minimum or the maximum value, or both. This indicates that all of them are censored. The two variables with the highest level of censoring are the last two, V21 and V23, the variables we are using as outcomes.

For the first censored regression, V21 on V01, V02, and V07 we obtain the following information.

```
Variable V21 is censored above
 It has 41 (45.56%) values = 15.330

 Estimated Mean and Standard Deviation based on 90 complete cases.

 Mean = 15.022 (0.324)
 Standard Deviation = 4.759 (0.023)

 Estimated Censored Regression based on 90 complete cases.

     V21 = 6.387 + 0.125*V01 + 0.243*V02 + 0.198*V07 + Error, R² = 0.420
 Standerr  (1.745) (0.0838)    (0.0933)    (0.198)
 z-values   3.660   1.497       2.598       1.002
 P-values   0.000   0.134       0.009       0.316
```

The output shows that 45.56% of the outcome variables is equal to the maximum value for this variable. In terms of the estimated equation, we note that the estimated coefficient for V02 is statistically significant, but those for V01and V07 are not.

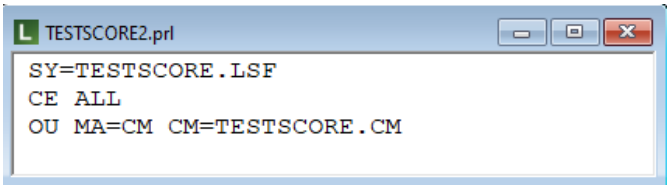
```
Variable V23 is censored below and above.
 It has 3 ( 3.33%) values = 2.436 and 21 (23.33%) values = 15.436

 Estimated Mean and Standard Deviation based on 90 complete cases.

 Mean = 12.013 (0.285)
 Standard Deviation = 3.402 (0.022)

 Estimated Censored Regression based on 90 complete cases.

      V23 = 5.981 + 0.0371*V01 + 0.206*V02 + 0.192*V07 + Error, R² = 0.361
 Standerr  (1.142) (0.0597)     (0.0642)    (0.125)
 z-values   5.238   0.621        3.211       1.539
 P-values   0.000   0.534        0.001       0.124
```

V23 has 23.33% of its values at the top end of observed values, but we see evidence of lesser censoring at the lower end as well. The same result obtained for V21 holds here – only the estimated coefficient for V02 is statistically significant.

```
Residual Correlation Matrix

               V21         V23
          --------    --------
     V21       1.000
     V23       0.262       1.000

 Residual Covariance Matrix

               V21         V23
          --------    --------
     V21      13.124
     V23       2.578       7.390
```

3. Treating all variables as censored

Given the evidence of censoring of all the variables in these data, we opt to treat them all as censored. We wish to examine the correlation matrix obtained when they are all treated as censored and compare it to that for the uncensored case.



The line

```
CE ALL
```

indicates that all the variables are considered censored, and we also request the writing of the covariance matrix to an external file through the use of the CM keyword on the OU command. Note that if a variable is declared as censored, but is really uncensored, LISREL will treat it as a special case of censoring with only one observation at the minimum or maximum of its range.

```
Correlation Matrix

              V01        V02        V07        V21        V23
         --------   --------   --------   --------   --------
V01        1.000
V02        0.761      1.000
V07        0.641      0.595      1.000
V21        0.485      0.653      0.607      1.000
V23        0.492      0.576      0.478      0.553      1.000


 Covariance Matrix

              V01        V02        V07        V21        V23
         --------   --------   --------   --------   --------
V01       61.719
V02       41.829     49.676
V07       15.284     12.962      9.421
V21       13.038     11.997      3.618      8.988
V23       13.854     14.288      5.051      5.751     11.573

 Total Variance = 141.377 Generalized Variance = 257403.392

 Largest Eigenvalue = 110.203 Smallest Eigenvalue = 3.958

 Condition Number = 5.277

 Means

              V01        V02        V07        V21        V23
         --------   --------   --------   --------   --------
          21.789     14.622     11.489     13.352     12.013

 Standard Deviations

              V01        V02        V07        V21        V23
         --------   --------   --------   --------   --------
           7.856      7.048      3.069      2.998      3.402
```

When these results are compared with those obtained for the exploratory data screening of this data set, we note that though similar, differences in statistics are observed. This is summarized in the table below.

Censored variables:

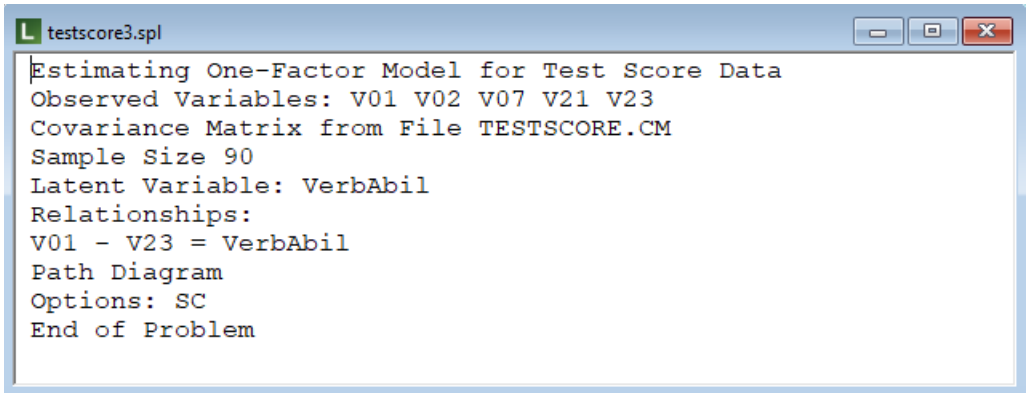
	Mean	Std. Dev.	Correlations				
V01	21.789	7.856	1.000				
V02	14.622	7.048	0.761	1.000			
V07	11.489	3.069	0.641	0.595	1.000		
V21	13.352	2.998	0.485	0.653	0.607	1.000	
V23	12.013	3.402	0.492	0.576	0.478	0.553	1.000

Uncensored variables:

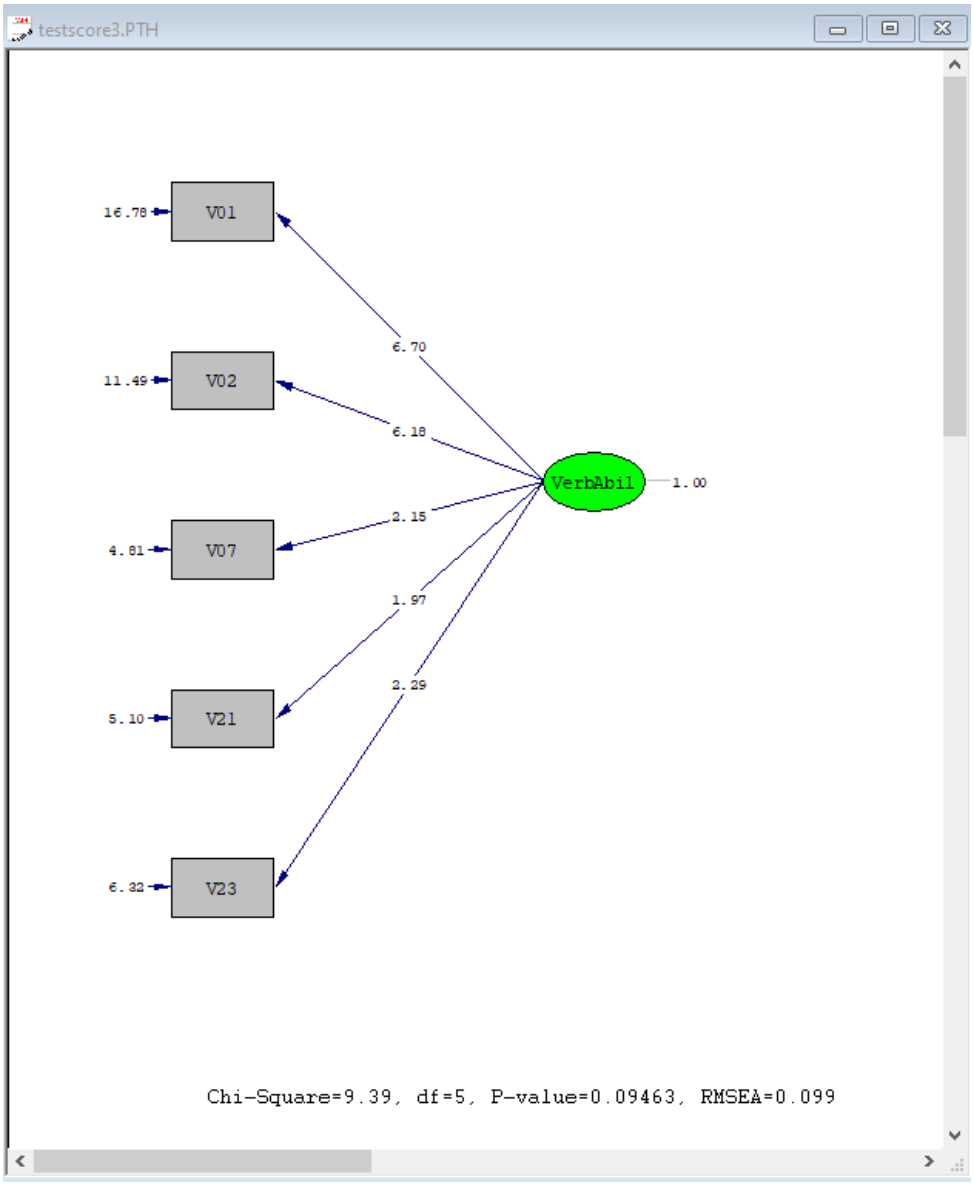
	Mean	Std. Dev.	Correlations				
V01	21.789	7.836	1.000				
V02	14.622	7.048	0.755	1.000			
V07	11.489	3.069	0.634	0.599	1.000		
V21	15.022	2.998	0.554	0.568	0.393	1.000	
V23	12.578	3.402	0.518	0.596	0.484	0.564	1.000

4. Estimating a one-factor model

We now do a factor analysis of the five variables and test the hypothesis that there is one common factor. The syntax for doing this analysis is given in **testscore3.spl**.



The path diagram for this model is shown below.



The following results are obtained for this model.

Measurement Equations

V01 = 6.703*VerbAbil, Errorvar.= 16.783, R^2 = 0.728

Standerr	(0.698)	(3.832)
Z-values	9.601	4.379
P-values	0.000	0.000

V02 = 6.179*VerbAbil, Errorvar.= 11.495, R^2 = 0.769

Standerr	(0.618)	(2.959)
Z-values	10.004	3.885
P-values	0.000	0.000

```
     V07 = 2.147*VerbAbil, Errorvar.= 4.810 , R² = 0.489
 Standerr  (0.296)                   (0.815)
 Z-values   7.247                    5.900
 P-values   0.000                    0.000

     V21 = 1.972*VerbAbil, Errorvar.= 5.101 , R² = 0.433
 Standerr  (0.295)                   (0.841)
 Z-values   6.682                    6.062
 P-values   0.000                    0.000

     V23 = 2.292*VerbAbil, Errorvar.= 6.321 , R² = 0.454
 Standerr  (0.332)                   (1.052)
 Z-values   6.894                    6.005
 P-values   0.000                    0.000
```

We note that the estimated loadings of the first two variables on the factor VerbAbil are the highest.

Log-likelihood Values

	Estimated Model	Saturated Model
Number of free parameters(t)	10	15
-2ln(L)	1580.655	1571.269
AIC (Akaike, 1974)*	1600.655	1601.269
BIC (Schwarz, 1978)*	1625.653	1638.766

*LISREL uses AIC= 2t - 2ln(L) and BIC = tln(N)- 2ln(L)

Goodness-of-Fit Statistics

Degrees of Freedom for (C1)-(C2)	5
Maximum Likelihood Ratio Chi-Square (C1)	9.386 (P = 0.0946)
Browne's (1984) ADF Chi-Square (C2_NT)	8.832 (P = 0.1160)

The test Chi-Square test statistic value and accompanying *p*-value indicate an acceptable fit at a significant level of 10%. It can therefore be concluded that the single latent variable Verbabil is an adequate indicator of the variation in the five observed variables.