Regression of GNP

Goldberger (1964, p.187) presented raw data on gross national product in billions of dollars (y = GNP), labor inputs in millions of man-years (x_1 = LABOR), real capital in billions of dollars (x_2 = CAPITAL), and the time in years measured from 1928 (x_3 = TIME). A path diagram for the regression of y on x_1, x_2, and x_3 is shown in the figure below. The data consists of 23 annual observations for the United States during 1929 – 1941 and 1946 – 1955. The covariance matrix of the variables is given in Table 1.

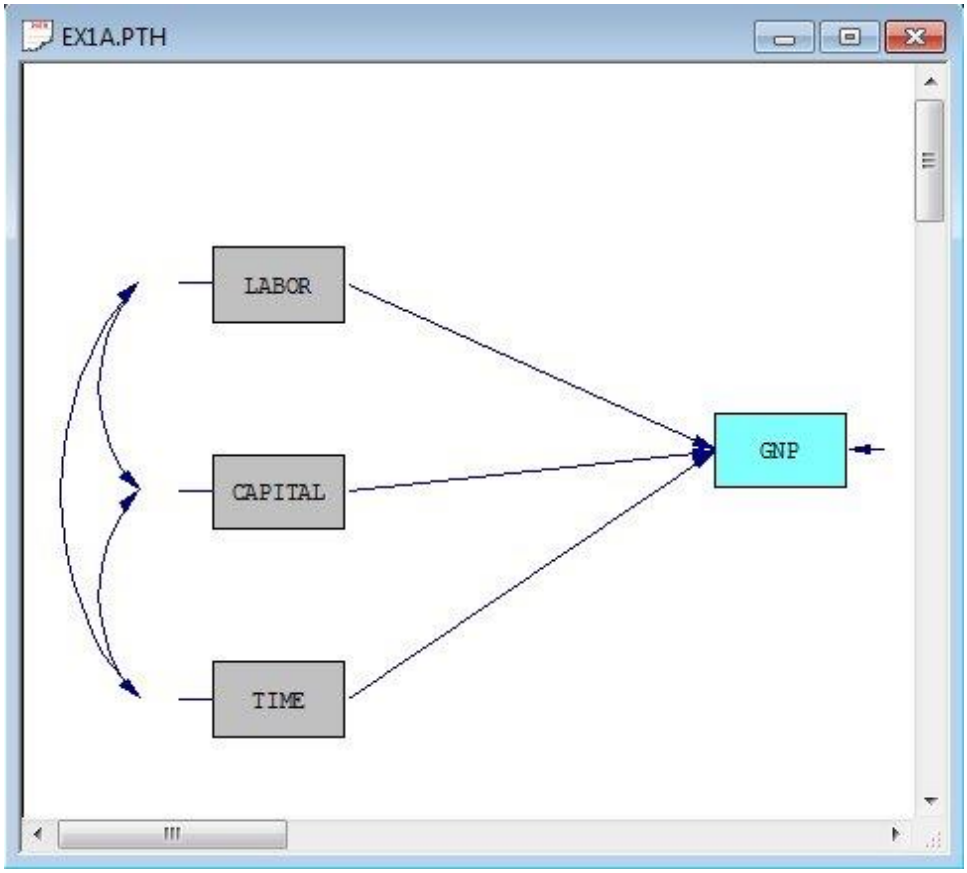


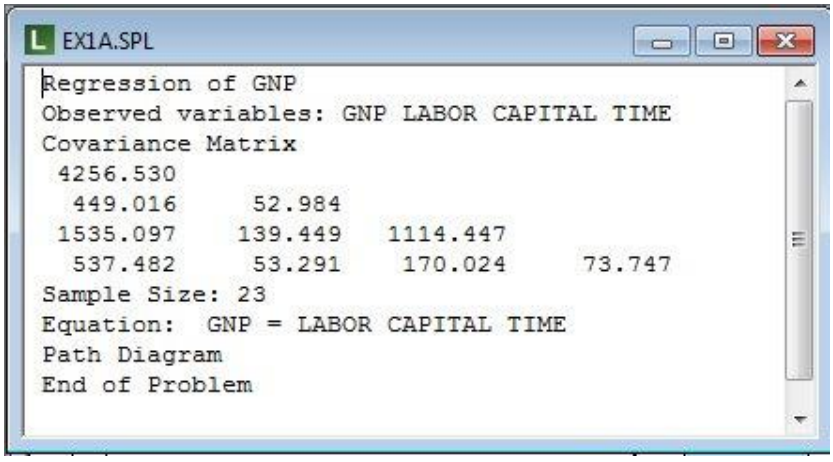
Table 1: Covariance matrix for GNP data

	y	x_1	x_2	x_3
GNP	4256.530			
LABOR	449.016	52.984		
CAPITAL	1535.097	139.449	1114.447	
TIME	537.482	53.291	170.024	73.747
	180.435	45.565	50.087	13.739

As the path diagram indicates, LABOR, CAPITAL, and TIME are supposed to influence GNP. This is indicated by the three one-way (unidirected) arrows pointing towards GNP. The three two-way arrows on the left indicate that the three independent variables may be correlated. The one-way arrow on the right represents the effect of the error term z .

The means and the covariance matrix may be computed using the PRELIS program, which can take missing values and other data problems into account.

The SIMPLIS input file **EX1A.SPL**:



The first line is a title line. Any number of title lines may be used and one may write anything on the title lines as of as they do not begin with the words *Observed Variables* or *Labels.*

The second line defines the names of the variables. The order of the names of the variables must correspond to the order of the variables in the covariance matrix.

The following five lines define the covariance matrix. Only the lower half of the covariance matrix is given. The elements of the covariance matrix are given in free format, i.e., with blanks between them.

The next line specifies the sample size, i.e., the number of cases on which the covariance matrix is based.

The line

```
Equation: GNP = LABOR CAPITAL TIME
```

specifies the regression equation to be estimated and is interpreted as "GNP depends on LABOR, CAPITAL, and TIME". The may also be specified as

```
Paths: LABOR – TIME -> GNP
```

which is interpreted as "There is a path from each of the variables LABOR, CAPITAL, and TIME to GNP".

The line `End of Problem`, which is optional, may be used to specify the end of the problem. In this case, this is also the end of the output file.

The estimated regression equation shows up in the output file as:

```
GNP = 3.819*LABOR + 0.322*CAPITAL + 3.786*TIME, Errorvar.= 12.470, R² = 0.997
 Standerr  (0.210)       (0.0298)        (0.181)                (3.943)
 Z-values   18.158        10.814          20.880                 3.162
 P-values   0.000          0.000           0.000                 0.002
```

The estimated regression coefficients appear in front of the * before each variable. The estimated partial regression coefficient of LABOR is 3.82. This is interpreted as follows. If LABOR increases one unit while CAPITAL and TIME are held fixed, the expected increase of GNP is 3.82 units on the average.

The error variance is 12.47, whish is small compared to the total variance 4256.53 of GNP. This indicates that the three independent variables LABOR, CAPITAL, and TIME account for almost all of the variance of GNP.

The numbers below the regression coefficients are the standard errors of the estimates. Each standard error is a measure of the precision of the parameter estimate. Below the standard errors are the *t*-values. The *t*-value is the ratio between the estimate and its standard error. If a *t*-value exceeds a certain level, we say that the corresponding parameter is *significant* which means that one can be fairly confident that the corresponding variable really influences GNP. In small samples, if the residuals are normally distributed, one can use a formal *F* test to test whether a specified regression coefficient is zero in the population. In our example, all the regression coefficients are highly significant.

The squared multiple correlation, R^2 is also given for each relationship. This is a measure of the strength of the linear relationship. A format test of the significance of the whole regression equation, i.e., a test of the hypothesis that all γ s are zero, can be obtained by computing

$$F = \frac{R^2 / q}{\left(1 - R^2\right) / \left(N - q - 1\right),}$$

where R^2 is the squared multiple correlation listed in the output file, *N* is the total sample size and *q* is the number of genuine *x*-variables. *F* is used as an *F*-statistic with *q* and *N-q-1* degrees of freedom. In our example, R^2 = 0.997 and *F* = 2104.8 with 2 and 19 degrees of freedom. Note that the number of decimals can be changed by including the line

```
Number of Decimals = 3
```

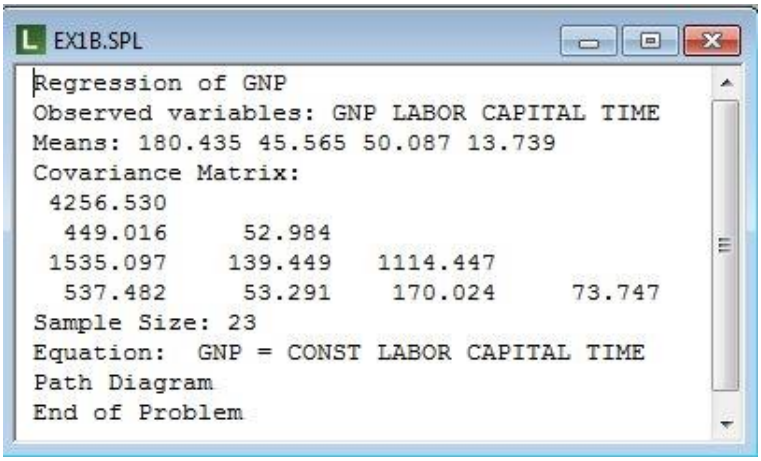
in the input file.

Observe that the regression model is scale-invariant in the following sense. If y is replaced by $y^* = c_0 y$ and x_i with $x_i^* = c_i x_i, i = 1, 2, \ldots, q$ where the *c*'s are arbitrary non-zero constants, the analysis of these scaled variables will yield regression coefficients

$$\hat{\gamma}_i^* = \left(c_0 / c_i\right) \hat{\gamma}_i .$$

The standard error will change similarly, but the *t*-values are invariant under such scalings of the variables.

The constant or intercept term α is the mean of *y* when all *x*-variables are zero. When only the covariance matrix is given in the input file, LISREL assumes that all variables are measured in deviations from their means and that α is zero. To estimate α, all one needs to do is to include the means of the variables in the input file.



```
GNP = - 61.727 + 3.819*LABOR + 0.322*CAPITAL + 3.786*TIME, Errorvar.= 12.470, R²
= 0.997
Standerr     (7.574)  (0.210)       (0.0298)        (0.181)
(3.943)
Z-values     -8.150    18.158        10.814          20.880                  3.162
P-values      0.000    0.000         0.000           0.000                   0.002
```

The estimated regression equation is shown above. Note that the estimated regression parameters and their standard errors are the same as for the previous example (**1A**). The intercept term is estimated at -61.73 with a standard error of 7.77. This indicates that GNP would be highly negative if LABOR, CAPITAL, and TIME were all zero. This is an extrapolation which assumes that the relationship is linear over the entire range of values of the *x*-variables.

Analysis of variance (ANOVA) and analysis of covariance (ANCOVA) can also be done with regression analysis by including dummy variables in the regression equation (see Huitema, 1980, or Jöreskog & Sörbom, 1989, pp. 112 – 116).