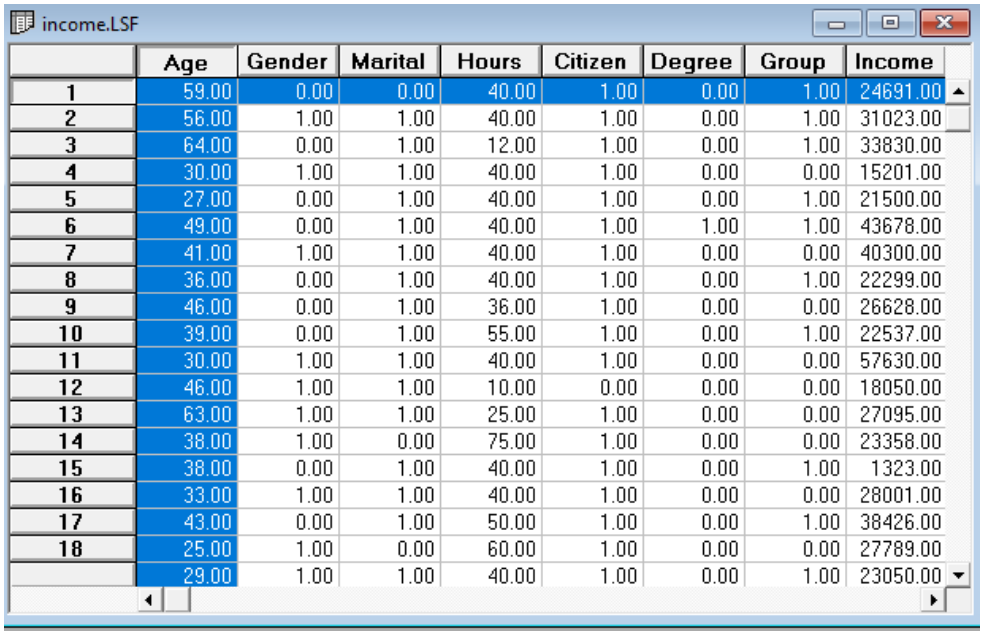
Univariate regression

Contents

1. Introduction

Data from the USA March 1995 Population Survey is used. The file **income.lsf** shown below contains data on a sample of persons between the ages of 21 and 65 who were employed full time in 1994 and had an annual income of US$1 or more. The data and syntax files can be found in the **MVABOOK examples\Chapter2** folder.



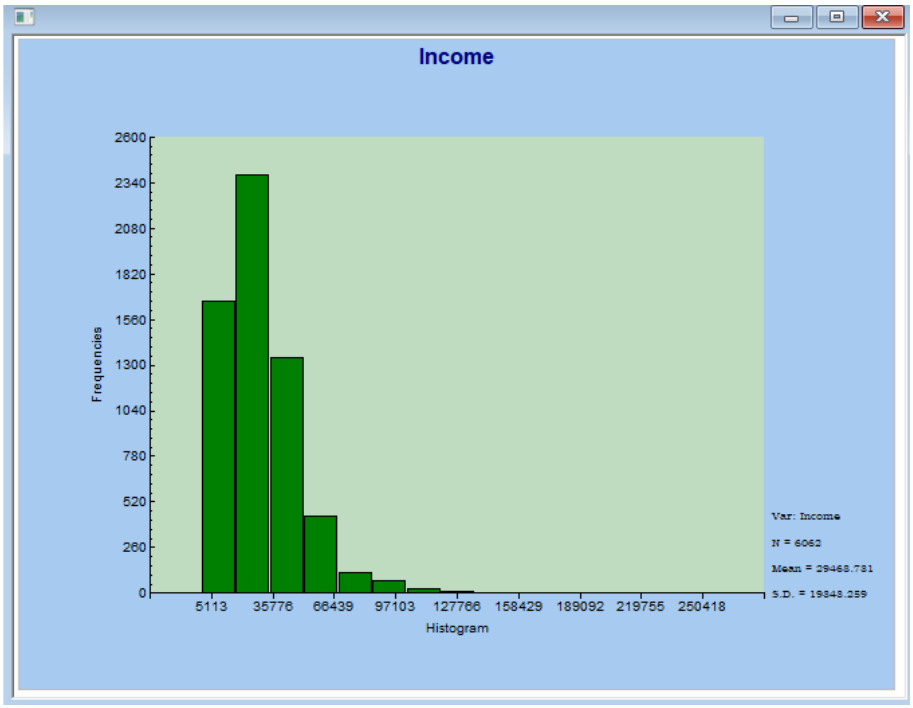
income.LSF

	Age	Gender	Marital	Hours	Citizen	Degree	Group	Income
1	59.00	0.00	0.00	40.00	1.00	0.00	1.00	24691.00
2	56.00	1.00	1.00	40.00	1.00	0.00	1.00	31023.00
3	64.00	0.00	1.00	12.00	1.00	0.00	1.00	33830.00
4	30.00	1.00	1.00	40.00	1.00	0.00	0.00	15201.00
5	27.00	0.00	1.00	40.00	1.00	0.00	1.00	21500.00
6	49.00	0.00	1.00	40.00	1.00	1.00	1.00	43678.00
7	41.00	1.00	1.00	40.00	1.00	0.00	0.00	40300.00
8	36.00	0.00	1.00	40.00	1.00	0.00	1.00	22299.00
9	46.00	0.00	1.00	36.00	1.00	0.00	0.00	26628.00
10	39.00	0.00	1.00	55.00	1.00	0.00	1.00	22537.00
11	30.00	1.00	1.00	40.00	1.00	0.00	0.00	57630.00
12	46.00	1.00	1.00	10.00	0.00	0.00	0.00	18050.00
13	63.00	1.00	1.00	25.00	1.00	0.00	0.00	27095.00
14	38.00	1.00	0.00	75.00	1.00	0.00	0.00	23358.00
15	38.00	0.00	1.00	40.00	1.00	0.00	1.00	1323.00
16	33.00	1.00	1.00	40.00	1.00	0.00	0.00	28001.00
17	43.00	0.00	1.00	50.00	1.00	0.00	1.00	38426.00
18	25.00	1.00	0.00	60.00	1.00	0.00	0.00	27789.00
	29.00	1.00	1.00	40.00	1.00	0.00	1.00	23050.00

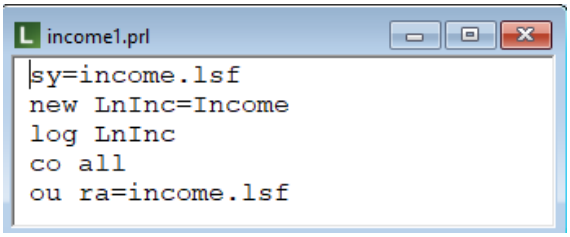
The variables are

- Age age in years
- Gender = 0 for females, 1 for males
- Marital = 1 if married, 0 otherwise
- Hours represents the number of hours worked the last week at all jobs
- Citizen = 1 for native born Americans, 0 for foreign born
- Degree = 1 for master's degree, professional school degree, or doctoral degree and 0 otherwise
- Group = 1 for respondents with professional specialty in the education sector; 0 for workers in the construction sector
- Income is the personal income of the person in thousands of US$ during 1994.

It is well known that income is not a normally distributed variable. Requesting a univariate graph of the variable Income via the **Graphs** menu produces the histogram below. It is clear that this variable is not normally distributed and that some transformation of it should be considered.



As a first step, we need to calculate the natural logarithm of personal income. To do so, we use the PRELIS command file **income1.prl**. It creates a new variable LnInc using the New command and append it to the LSF file as shown on the Output command.



The amended LSF file is shown below.

income.LSF

	Age	Gender	Marital	Hours	Citizen	Degree	Group	Income	LnInc
1	59.00	0.00	0.00	40.00	1.00	0.00	1.00	24691.00	10.11
2	56.00	1.00	1.00	40.00	1.00	0.00	1.00	31023.00	10.34
3	64.00	0.00	1.00	12.00	1.00	0.00	1.00	33830.00	10.43
4	30.00	1.00	1.00	40.00	1.00	0.00	0.00	15201.00	9.63
5	27.00	0.00	1.00	40.00	1.00	0.00	1.00	21500.00	9.98
6	49.00	0.00	1.00	40.00	1.00	1.00	1.00	43678.00	10.68
7	41.00	1.00	1.00	40.00	1.00	0.00	0.00	40300.00	10.60
8	36.00	0.00	1.00	40.00	1.00	0.00	1.00	22299.00	10.01
9	46.00	0.00	1.00	36.00	1.00	0.00	0.00	26628.00	10.19
10	39.00	0.00	1.00	55.00	1.00	0.00	1.00	22537.00	10.02
11	30.00	1.00	1.00	40.00	1.00	0.00	0.00	57630.00	10.96
12	46.00	1.00	1.00	10.00	0.00	0.00	0.00	18050.00	9.80
13	63.00	1.00	1.00	25.00	1.00	0.00	0.00	27095.00	10.21
14	38.00	1.00	0.00	75.00	1.00	0.00	0.00	23358.00	10.06
15	38.00	0.00	1.00	40.00	1.00	0.00	1.00	1323.00	7.19
16	33.00	1.00	1.00	40.00	1.00	0.00	0.00	28001.00	10.24
17	43.00	0.00	1.00	50.00	1.00	0.00	1.00	38426.00	10.56
18	25.00	1.00	0.00	60.00	1.00	0.00	0.00	27789.00	10.23
19	29.00	1.00	1.00	40.00	1.00	0.00	1.00	23050.00	10.05
20	44.00	1.00	1.00	38.00	1.00	1.00	1.00	47231.00	10.76

2. Using LISREL syntax

To run the regression of LnInc on Age to Group we use LISREL syntax (see **income2b1.lis**).

income2b1.lis

```
Regression of LnInc
DA NI=9
RA=income.lsf
SE
LnInc Age Gender Marital Hours Citizen Degree Group /
MO NY=1 NX=7 AL=FR KA=FR
OU
```

The first line is a title line. The DA line indicates the number of variables in the data set, while the RA line specifies the name and path to the data file. The outcome variable of interest is the last variable in the LSF. LISREL requires it to be the first listed variable, so the SE command is used to reorder the variables so that the outcome lnInc appears first. Note that SE appears on it's own on a line, followed by the required order of variables in the next line.

The MO line specifies, first of all, the number of *y*- and *x*-variables. The LISREL matrices BE, GA, PH, and PS are default on the MO line, so that $\mathbf{B}(1\times1) = 0$, $\mathbf{\Gamma}(1\times7)$, representing the regression coefficients, is a vector of free parameters to be estimated. $\mathbf{\Phi}(7\times7)$, the covariance matrix of the *x*-variables, is a free symmetric matrix to be estimated and $\mathbf{\Psi}(1\times1)$, the residual variance, is a free parameter to be estimated. To estimate the intercept α and the mean vector $\boldsymbol{\kappa}$ of the *x*-variables one must specify AL = FR and KA = FR on the MO line. If that is not done, these would be assumed zero.

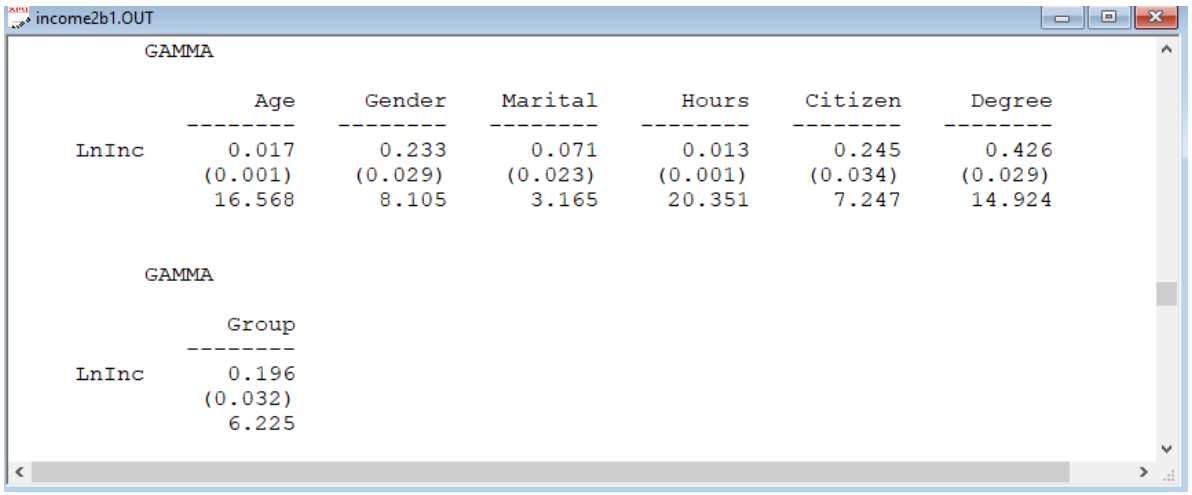
The last line in a LISREL syntax file should always be the OU line. This command is used to request additional output.

When SIMPLIS or LISREL syntax is used, LISREL will regard the model as a mean and covariance structure and fit all free parameters of the model to the sample mean vector $\left(\bar{\mathbf{y}}, \bar{\mathbf{x}}\right)$ and sample covariance matrix

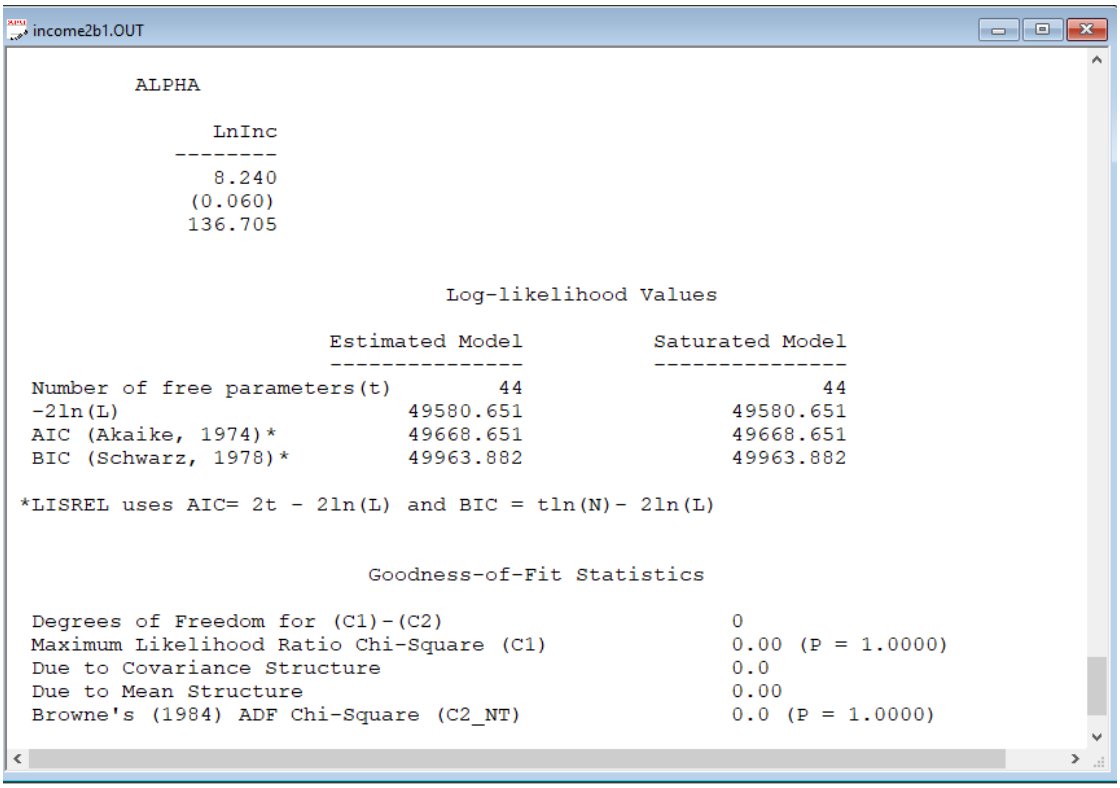
$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{yy} & \\ \mathbf{S}_{xy} & \mathbf{S}_{xx} \end{bmatrix}$$

of $\left(\bar{\mathbf{y}}, \bar{\mathbf{x}}\right)$ by minimizing some fit function. The default fit function is ML which gives maximum likelihood estimates under the assumption that $\left(\bar{\mathbf{y}}, \bar{\mathbf{x}}\right)$ has a multivariate normal distribution. As a mean and covariance structure, the regression model is a saturated model because the number of free parameters to be estimated is the same as the number independent variables in $\left(\bar{\mathbf{y}}, \bar{\mathbf{x}}\right)$ and **S**. In this case there is an explicit solution for the parameter estimates.

Results for this analysis are as follows:



GAMMA contains the estimated regression coefficients with associated standard errors and *z*-values. ALPHA provides the estimate of the intercept.



It should be noted that it is a lot easier to fit this model using PRELIS syntax. The PRELIS syntax

```
Systemfile income.lsf
Regress lnInc on Age – Group
```

will fit the same model. There may be small differences in estimated standard errors between LISREL and PRELIS, due to PRELIS using more exact formulas while LISREL uses an asymptotic formula.

3. Hypothesis testing

Suppose we want to test the hypotheses

$$H_0 : \gamma_2 = \gamma_5$$
$$H_1 : \gamma_2 \neq \gamma_5$$

This can be done by modifying the LISREL syntax file previously used to

income2b2.lis

```
Regression of LnInc
DA NI=9
RA=income.lsf
SE
LnInc Age Gender Marital Hours Citizen Degree Group /
MO NY=1 NX=7 AL=FR KA=FR
EQ GA(1,2) GA(1,5)
OU
```

The EQ line indicates that the second and fifth elements of GAMMA are set equal to each other.

income2b2.OUT

```
                          Log-likelihood Values

                     Estimated Model          Saturated Model
                     ---------------          ---------------
 Number of free parameters(t)         43                    44
 -2ln(L)                       49580.723             49580.651
 AIC (Akaike, 1974)*           49666.723             49668.651
 BIC (Schwarz, 1978)*          49955.244             49963.882

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


                       Goodness-of-Fit Statistics

 Degrees of Freedom for (C1)-(C2)                      1
 Maximum Likelihood Ratio Chi-Square (C1)              0.0722  (P = 0.7882)
 Due to Covariance Structure                           0.0722
 Due to Mean Structure                                 0.0
 Browne's (1984) ADF Chi-Square (C2_NT)                0.0722  (P = 0.7882)
```

To test the hypothesis one can use the chi-square = 0.0722 with 1 degree of freedom. The associated *p*-value is 0.7882 so the hypothesis cannot be rejected.

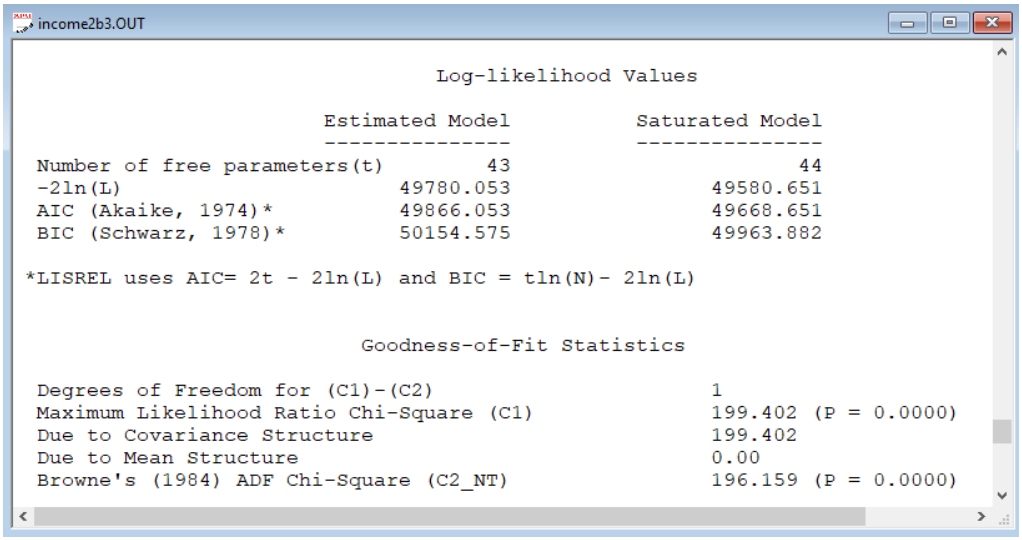
To test the hypotheses

$$H_0 : \gamma_1 = \gamma_6$$
$$H_1 : \gamma_1 \neq \gamma_6$$

the EQ line is amended to

```
EQ GA(1,1) GA(1,6)
```

as done in **income2b3.lis**. For this analysis, we find a reported chi-square of 199.4. The associated *p*-value indicates that the hypothesis is strongly rejected.



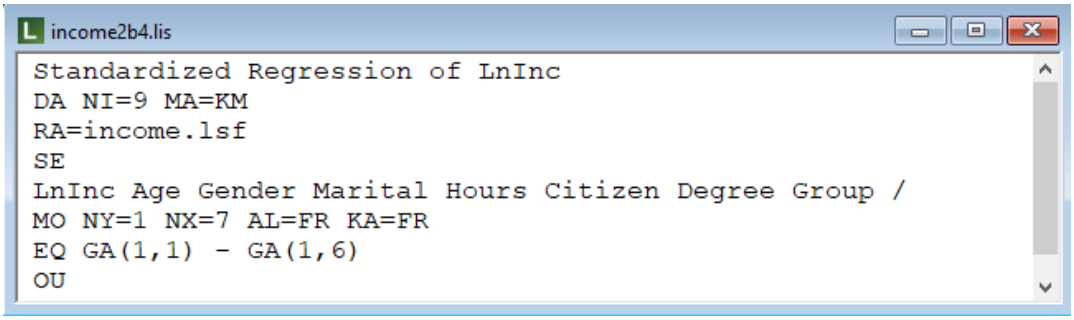
These two tests indicate that not all regression coefficients are equal. Another question worth investigating is whether they would be equal if all variables are standardized. This can be tested by amending the LISREL syntax file.

In the syntax file income2b4.lis the EQ line has been amended from

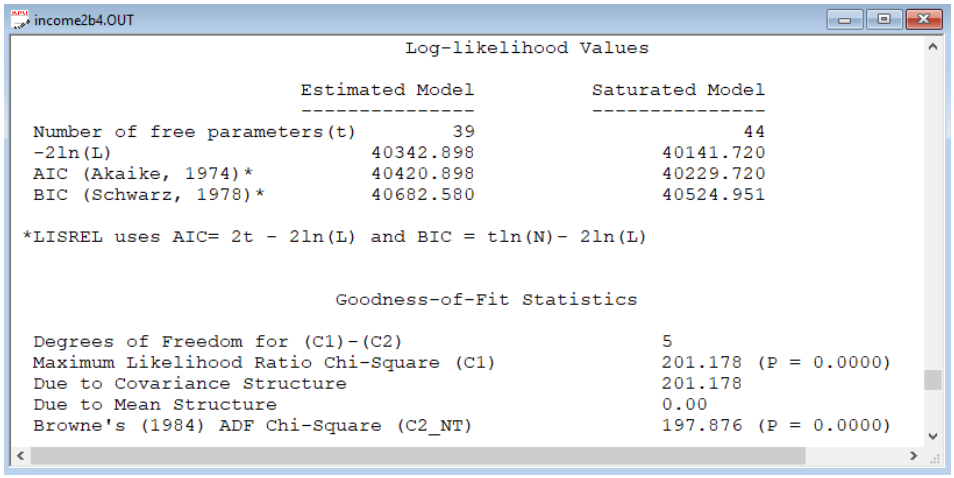
```
EQ GA(1,1) GA(1,6)
```

to

```
EQ GA(1,1) - GA(1,6).
```



We also change the DA line to include the option MA = KM, indicating that the correlation matrix is to be used instead of the default covariance matrix (MA = CM).



From the results of this analysis, we obtain a chi-square of 201.178 with 6 degrees of freedom. Judging by the *p*-value of 0.000, the hypothesis is strongly rejected.