Linear regression

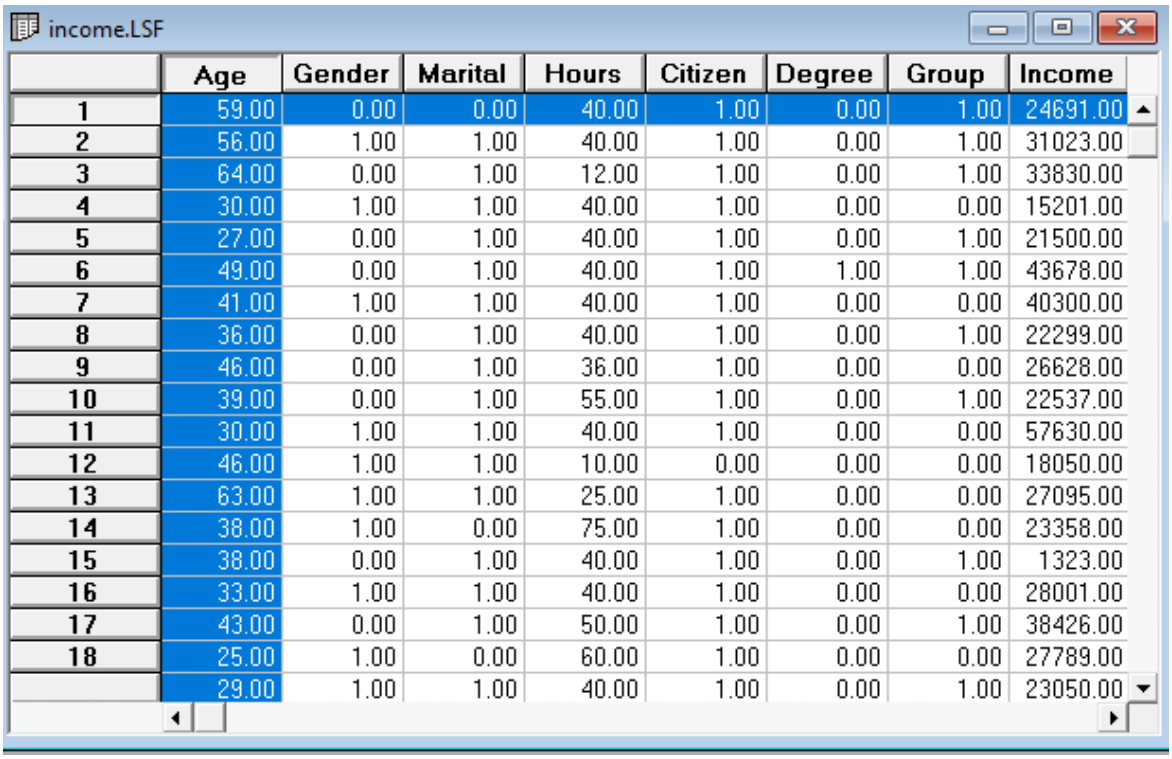
Contents

1. Linear regression 1

2. Calculating expected income based on linear regression results 4

1. Linear regression

In this example of linear regression, 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.



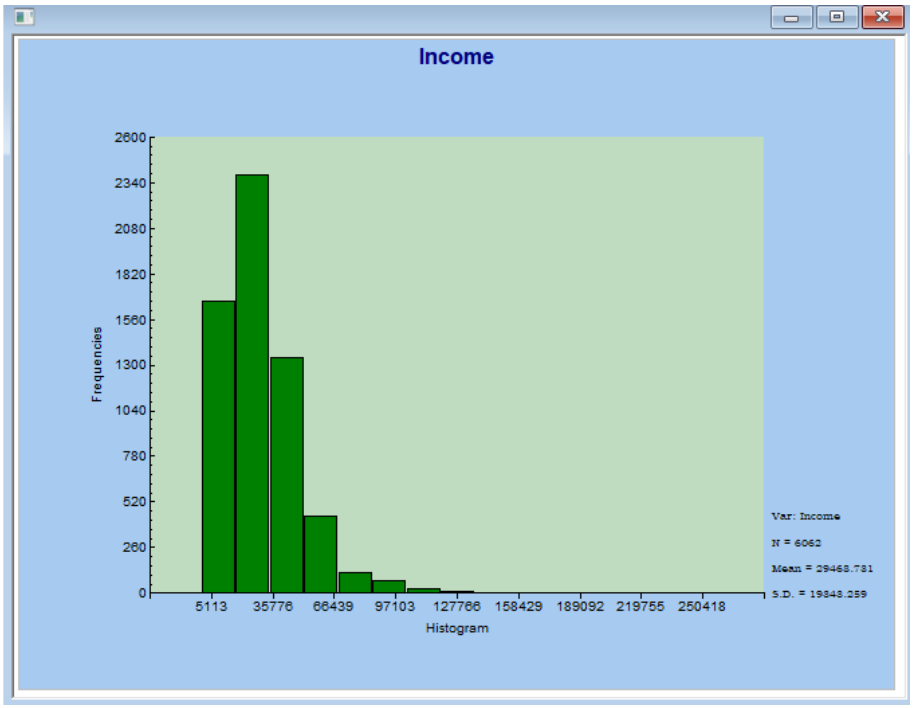
income.LSF

	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

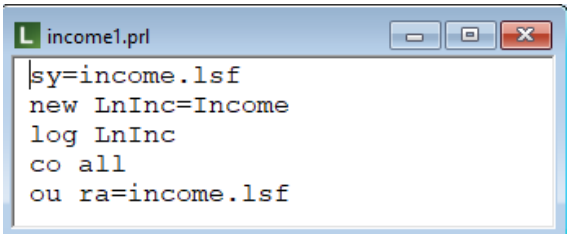
The variables are

- 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.

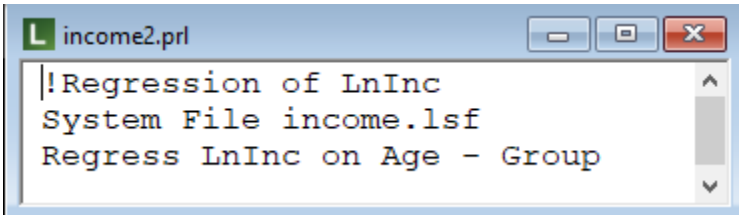


The amended LSF file is shown below.

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

We now fit a regression line using the natural log of personal income as outcome variable, and all other demographic variables as predictors.



Results are as follows:

```
Estimated Equations

    LnInc = 8.240   + 0.0169*Age + 0.233*Gender + 0.0713*Marital + 0.0133*Hours
 Standerr  (0.0603)  (0.00102)    (0.0288)       (0.0225)         (0.000652)
 t-values   136.694   16.567       8.104          3.165            20.350
 P-values   0.000     0.000        0.000          0.002            0.000

             + 0.245*Citizen + 0.426*Degree + 0.196*Group + Error, R² = 0.214
              (0.0338)        (0.0285)       (0.0316)
               7.247           14.923         6.225
               0.000           0.000          0.000

 Error Variance = 0.620
```

We see that all estimated coefficients are highly significant. When all other variables are held constant, being male is expected to lead to an increase of 0.233 in the outcome. Similarly, natural born citizens are expected to earn more than foreign born citizens, again with all other variables held constant. For the variables coded 0,1 (Gender, Marital, Citizen, Degree and Group) it seems as if having a professional degree has the biggest positive impact on the outcome, followed by whether the person is a natural or foreign born citizen.

2. Calculating expected income based on linear regression results

One can use PRELIS to calculate expected outcomes for subsets of persons in the data. First, create a txt file named, say, **profiles.dat** containing columns equal in number to the number of predictors in the model. For our example, we need 7 columns. Each row of this file represents a person. We now populate the file by assigning values for the persons for each of the observed variables.

From the univariate summary statistics we know, for example, that Age varies between 21 and 65. Similarly, the hours worked varies between 0 and 99. Values assigned should ideally fall within these ranges.

```
Univariate Summary Statistics for Continuous Variables

 Variable      Mean   St. Dev.   Skewness   Kurtosis   Minimum Freq.   Maximum Freq.
 --------      ----   --------   --------   --------   ------- -----   ------- -----
      Age    39.859     10.615      0.227     -0.771    21.000    70    65.000    30
   Gender     0.624      0.484     -0.512     -1.738     0.000  2279     1.000  3783
  Marital     0.687      0.464     -0.805     -1.352     0.000  1899     1.000  4163
    Hours    37.111     15.701     -0.565      0.967     0.000   347    99.000     2
  Citizen     0.897      0.304     -2.617      4.850     0.000   623     1.000  5439
   Degree     0.249      0.433      1.161     -0.653     0.000  4552     1.000  1510
```

```
Group     0.532     0.499     -0.126     -1.985     0.000  2840     1.000  3222
```

An example of such a file is shown below. A value of either 30 or 40 is assigned to Age and hours worked is set to either 35 or 40 hours, in other words in the middle of the range for this variable.

```
30  0  0  40  0  0  0
30  1  1  40  1  1  1
40  0  0  35  0  0  0
40  1  0  35  1  0  1
40  1  1  35  1  1  1
40  0  1  35  1  1  1
```

The syntax file **profiles1.prl** is now used to calculate the expected incomes for these persons. The estimates obtained are entered as part of the syntax and PRELIS is requested to create a new variable equal to the expected natural logarithm of income (LNY). In the next step, the actual income in US$ is calculated as well. Results are then written to the file **profiles_extended.dat** as requested on the Output command.

profiles1.prl

```
!Profiles Calculation
da ni=7
la
Age Gender Marital Hours Citizen Degree Group
ra=profiles.dat
new A=8.240+0.0169*Age+0.233*Gender+0.0713*Marital
new B=0.0133*Hours+0.245*Citizen+0.426*Degree+0.196*Group
new LNY=A+B
new EST=LNY
exp EST
new INCOME=1.3634*EST
co all
sd A B
ou ra=profiles_extended.dat
```

The contents of this file are shown below.

profiles_extended.DAT - Notepad

File Edit Format View Help

```
    30.000      0.000      0.000     40.000      0.000      0.000      0.000      9.279  10710.713  14602.986
    30.000      1.000      1.000     40.000      1.000      1.000      1.000     10.450  34554.730  47111.919
    40.000      0.000      0.000     35.000      0.000      0.000      0.000      9.381  11866.798  16179.193
    40.000      1.000      0.000     35.000      1.000      0.000      1.000     10.055  23283.488  31744.707
    40.000      1.000      1.000     35.000      1.000      1.000      1.000     10.553  38284.474  52197.052
    40.000      0.000      1.000     35.000      1.000      1.000      1.000     10.320  30327.183  41348.081
```

When we look at the last 2 records in this data file, we see that the expected income for a married, natural born male with a professional degree working in the educational sector is $52,197. For a female with the same profile the expected income is considerably lower at $41,348. The same seems to be true for unmarried males (the 3rd last record) when compared to a female with similar profile (4th last record).