Linear regression

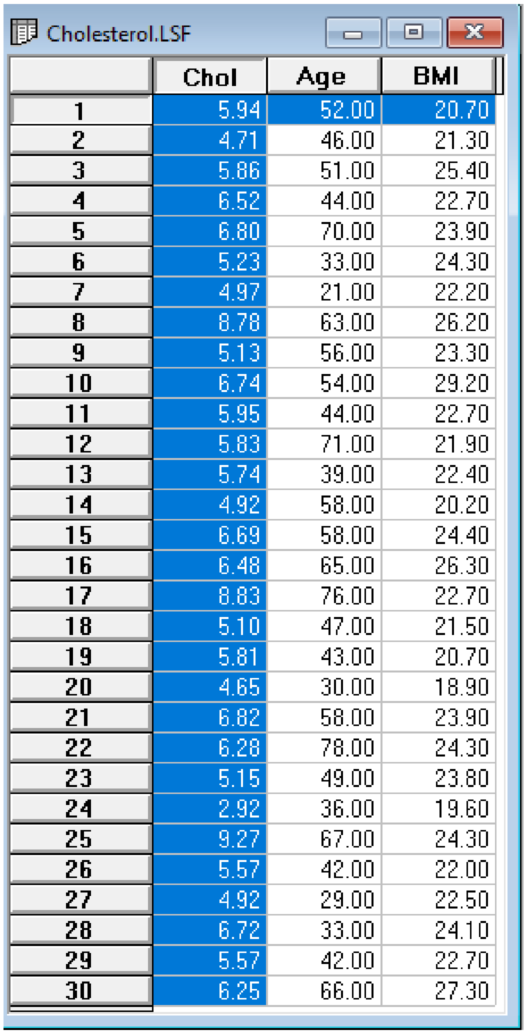
Contents

1. Introduction

Data on the serum cholesterol, age and body mass index for 30 women are given in the file **cholesterol.lsf**. The entire file is shown below. The data and syntax files can be found in the **MVABOOK examples\Chapter2** folder.

Cholesterol, represented by the variable Chol, is measured in millimoles per liter. BMI represents the body mass index and is the weight in kilograms divided by the square of height in meters.

We would like to explore the relationship between the level of cholesterol and age, simultaneously checking whether there is a BMI effect after controlling for age.

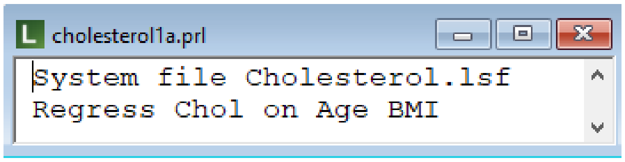


	Chol	Age	BMI
1	5.94	52.00	20.70
2	4.71	46.00	21.30
3	5.86	51.00	25.40
4	6.52	44.00	22.70
5	6.80	70.00	23.90
6	5.23	33.00	24.30
7	4.97	21.00	22.20
8	8.78	63.00	26.20
9	5.13	56.00	23.30
10	6.74	54.00	29.20
11	5.95	44.00	22.70
12	5.83	71.00	21.90
13	5.74	39.00	22.40
14	4.92	58.00	20.20
15	6.69	58.00	24.40
16	6.48	65.00	26.30
17	8.83	76.00	22.70
18	5.10	47.00	21.50
19	5.81	43.00	20.70
20	4.65	30.00	18.90
21	6.82	58.00	23.90
22	6.28	78.00	24.30
23	5.15	49.00	23.80
24	2.92	36.00	19.60
25	9.27	67.00	24.30
26	5.57	42.00	22.00
27	4.92	29.00	22.50
28	6.72	33.00	24.10
29	5.57	42.00	22.70
30	6.25	66.00	27.30

The regression of Chol on Age and BMI can be expressed mathematically as

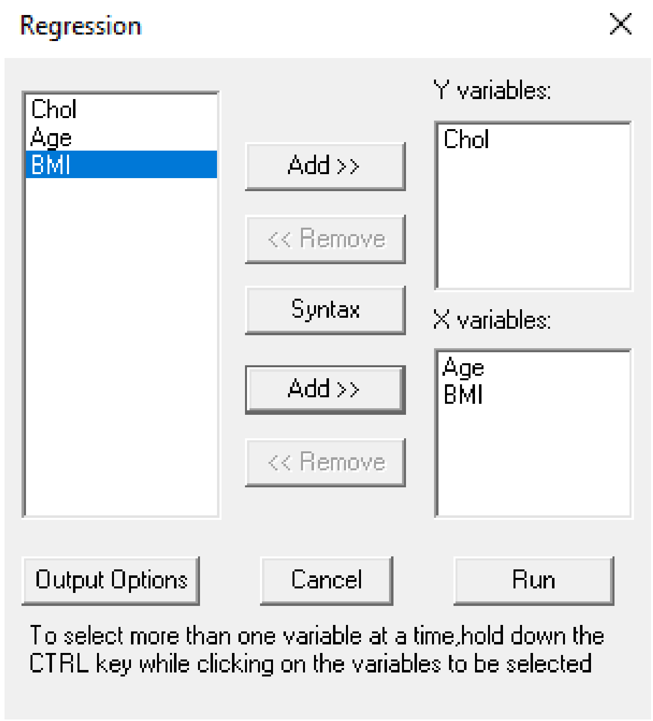
$$Chol_i = \alpha + \gamma_1 Age_i + \gamma_2 BMI_i + z_i, \qquad i = 1, 2, ..., 30.$$

This model corresponds with the syntax in the file **Cholesterol1a.prl**



```
System file Cholesterol.lsf
Regress Chol on Age BMI
```

but can also be done by selecting the **Regressions** option from the **Statistics** menu, completing the Regression dialog box as shown below, and simply clicking **Run**.



Regression

Chol
Age
BMI

Y variables:
Chol

X variables:
Age
BMI

To select more than one variable at a time,hold down the CTRL key while clicking on the variables to be selected

The output is as follows. Univariate summary statistics for the variables are followed by the estimated equation.

```
Univariate Summary Statistics for Continuous Variables

Variable     Mean  St. Dev.  Skewness  Kurtosis  Minimum Freq.  Maximum Freq.
--------     ----  --------  --------  --------  ------- -----  ------- -----
    Chol    6.005     1.309     0.671     1.591    2.920     1    9.270     1
     Age   50.700    14.770     0.028    -0.718   21.000     1   78.000     1
     BMI   23.180     2.268     0.526     0.659   18.900     1   29.200     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

    Chol   1.589   0.112     1.670   0.095        5.311   0.070
     Age   0.070   0.944    -0.976   0.329        0.957   0.620
     BMI   1.268   0.205     0.955   0.340        2.520   0.284

Estimated Equations

    Chol =  - 0.740 + 0.0410*Age + 0.201*BMI + Error, R² = 0.465
Standerr     (1.896) (0.0136)     (0.0888)
t-values     -0.390   3.006        2.269
P-values      0.699   0.006        0.031

Error Variance = 0.984
```

We note that the mean cholesterol value is 6.005 and ranges between 2.920 and 9.270. The ages of the 30 women vary considerably, between 21 and 78. From the estimated equation's *p*-values we conclude that both BMI and Age have a significant relationship with the observed cholesterol.

Interpreting the estimate for the intercept in this situation is problematical, as the estimated coefficient of -0.740 represents the expected average cholesterol level at Age and BMI equal to zero.

The positive estimate of γ_1 indicates a positive relationship between Age and cholesterol. Holding BMI constant, the oldest respondent (at 78) would be expected to have $(78-21)*0.0410=2.337$ millimoles per liter higher than the youngest (at 21). Increased BMI also leads to higher cholesterol, as indicated by the positive estimated coefficient of 0.201. BMI ranges between roughly 19 and 29, which means an additional 2 millimoles in expected cholesterol level between a person with a BMI of 19 and one with a BMI of 29.

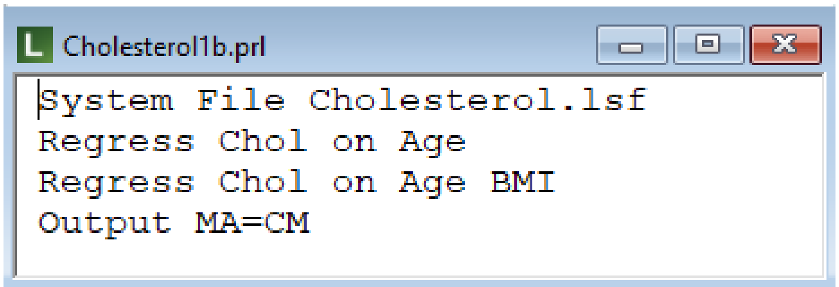
```
 The following chi-squares test the hypothesis that all
 regression coefficients are zero except the intercept.

 Variable       -2lnL  Chi-square   df  Covariates
 --------  ----------  ----------   --  ----------
     Chol      81.495      18.788    2   Age BMI

Analysis of Variance Table

   Regression d.f.    Residual d.f.           F  Covariates
   ---------------    -------------           -  ----------
       23.132    2      26.571   27      11.753  Age BMI

 Covariance Matrix

                Chol        Age        BMI
            --------   --------   --------
     Chol      1.714
      Age     11.658    218.148
      BMI      1.589     13.511      5.144

 Total Variance = 225.007 Generalized Variance = 860.977
 Largest Eigenvalue = 219.634 Smallest Eigenvalue = 0.871

 Condition Number = 15.882

 Means

                Chol        Age        BMI
            --------   --------   --------
               6.005     50.700     23.180

 Standard Deviations

                Chol        Age        BMI
            --------   --------   --------
               1.309     14.770      2.268
```

2. Hypothesis testing

As it seems as if BMI does not have as strong an effect on cholesterol levels as age, we test the hypotheses

$$H_0 : \gamma_2 = 0$$
$$H_1 : \gamma_2 \neq 0$$

This can be done in PRELIS, using the syntax file **Cholesterol1b.prl** below.



```
System File Cholesterol.lsf
Regress Chol on Age
Regress Chol on Age BMI
Output MA=CM
```

When only Age is used as predictor of Chol, the regression coefficient for Age is 0.0534, compared to 0.041 when BMI is included as a regressor, as Age and BMI are correlated.

```
Estimated Equations

     Chol = 3.296 + 0.0534*Age + Error, R² = 0.363
 Standerr  (0.705) (0.0134)
 t-values   4.676   3.999
 P-values   0.000   0.000

 Error Variance = 1.130


     Chol =  - 0.740 + 0.0410*Age + 0.201*BMI + Error, R² = 0.465
 Standerr     (1.896) (0.0136)     (0.0888)
 t-values     -0.390   3.006        2.269
 P-values      0.699   0.006        0.031

 Error Variance = 0.984
```

The estimated equations are followed by hypothesis test information. For the first equation, a chi-square of 13.553 is reported with 1 degree of freedom. This tests the hypothesis that the effect of Age is zero. For the second, the chi-square is 18.788 (2 degrees of freedom). Here the hypothesis being tested is $\gamma_1 = \gamma_2 = 0$. This chi-square is highly significant too. The difference between the two chi-squares, i.e. 5.235 with 1 degree of freedom and is a test that the additional effect of BMI is zero. This chi-square is statistically significant at the 5% level of significance, but not at the 1% level.

```
 The following chi-squares test the hypothesis that all
 regression coefficients are zero except the intercept.
 Variable       -2lnL  Chi-square   df  Covariates
 --------  ----------  ----------   --  ----------
     Chol      86.729      13.553    1   Age
     Chol      81.495      18.788    2   Age BMI
```

In the Analysis of Variance Table, the two lines of data correspond to the two regressions fitted here. The first two columns represent the sum of squares due to regression and their associated degrees of freedom. The next two columns contain the residual sum of squares and its degrees of freedom. The *F*-value is computed as

$$F = \frac{(RSS_0 - RSS)/q_0}{RSS/(N-q-1)} = \frac{(31.636 - 26.571)/1}{26.571/27} = 5.15$$

which is significant at a 5% level but not at a 15 level of significance. For samples less than 30 in size, the *F*-statistic is more accurate than the chi-square test, and so probably more appropriate in this specific situation.

From these results we conclude that there is a BMI effect, even after controlling for Age. Given the same size, it would be advisable to repeat these tests with a larger data set to verify this.

```
Analysis of Variance Table
   Regression d.f.    Residual d.f.           F  Covariates
   ---------------    -------------           -  ----------
       18.067    1      31.636   28      15.990  Age
       23.132    2      26.571   27      11.753  Age BMI
```

At the end of the output file, the covariance matrix, means and standard deviations are also given.

```
 Covariance Matrix

               Chol        Age        BMI
           --------   --------   --------
     Chol     1.714
      Age    11.658    218.148
      BMI     1.589     13.511      5.144

 Total Variance = 225.007 Generalized Variance = 860.977

 Largest Eigenvalue = 219.634 Smallest Eigenvalue = 0.871

 Condition Number = 15.882

 Means

               Chol        Age        BMI
           --------   --------   --------
              6.005     50.700     23.180
 Standard Deviations

               Chol        Age        BMI
           --------   --------   --------
              1.309     14.770      2.268
```

3. Checking assumptions

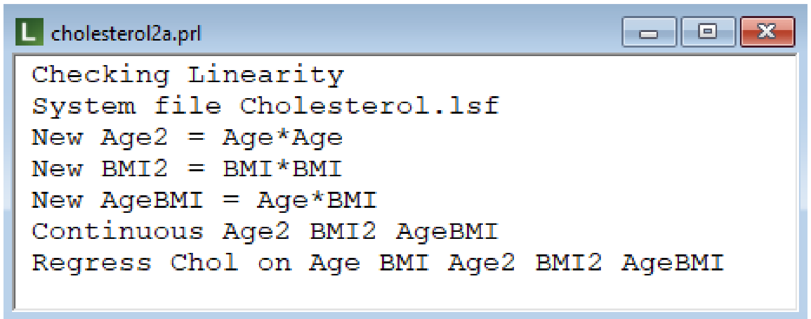
Underlying a linear regression model are the assumptions that the error term is independent of the covariates and that the mean of the outcome variable is linear in the predictors. It is also assumed to the residuals are i.i.d $N(0,\sigma^2)$.

In this section, we illustrate how LISREL may be used to check the linearity assumption and the assumption of normality of the residuals.

3.1 Checking linearity

The simplest way to test this is to evaluate a quadratic form of the model. If we add quadratic terms to the current model while it is appropriate to the data, we would expect the effects associated with the quadratic terms to be non-significant.

In our current example, this means adding a quadratic term for Age and a similar term for BMI. In the syntax shown below, BMI * BMI represents the quadratic term for BMI and Age * Age that for Age. We also include a interaction term denoted by Age * BMI in the model. Note that by running this syntax the newly created variables would automatically be added to the LSF file.



```
Checking Linearity
System file Cholesterol.lsf
New Age2  = Age*Age
New BMI2  = BMI*BMI
New AgeBMI  = Age*BMI
Continuous Age2 BMI2 AgeBMI
Regress Chol on Age BMI Age2 BMI2 AgeBMI
```

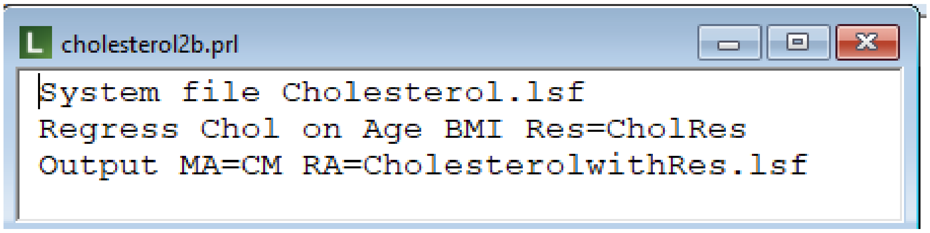
The estimated equation for this model is as given below. We conclude that none of the three additional terms are statistically significant and that the linear model previously fitted describes the data better than this model.

```
Estimated Equations

     Chol =  - 18.908 + 0.0129*Age + 1.795*BMI + 0.000501*Age2 - 0.0322*BMI2
 Standerr     (14.865) (0.202)      (1.283)     (0.000852)       (0.0313)
 t-values     -1.272    0.0638       1.400       0.587           -1.028
 P-values      0.215    0.950        0.174       0.562            0.314
```

```
            - 0.00110*AgeBMI + Error, R² = 0.513
             (0.00925)
             -0.119
              0.906

 Error Variance = 1.008
```

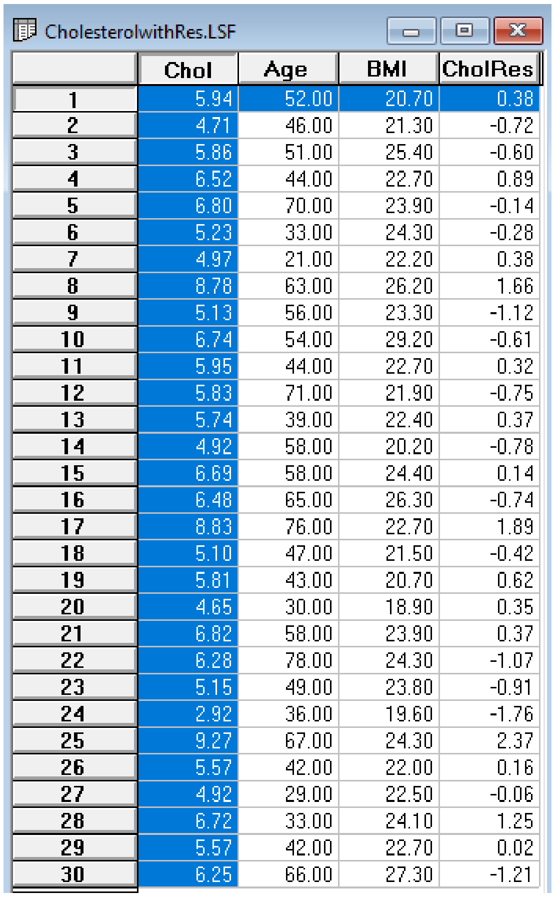
3.2 Normality of the residuals

To obtain the residuals for the linear model, we add the variable CholRes, representing the residuals, to the data file requested on the Output line.



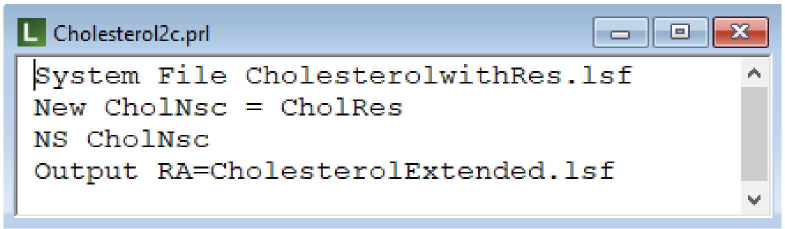
```
System file Cholesterol.lsf
Regress Chol on Age BMI Res=CholRes
Output MA=CM RA=CholesterolwithRes.lsf
```

The new data file is shown below.



	Chol	Age	BMI	CholRes
1	5.94	52.00	20.70	0.38
2	4.71	46.00	21.30	-0.72
3	5.86	51.00	25.40	-0.60
4	6.52	44.00	22.70	0.89
5	6.80	70.00	23.90	-0.14
6	5.23	33.00	24.30	-0.28
7	4.97	21.00	22.20	0.38
8	8.78	63.00	26.20	1.66
9	5.13	56.00	23.30	-1.12
10	6.74	54.00	29.20	-0.61
11	5.95	44.00	22.70	0.32
12	5.83	71.00	21.90	-0.75
13	5.74	39.00	22.40	0.37
14	4.92	58.00	20.20	-0.78
15	6.69	58.00	24.40	0.14
16	6.48	65.00	26.30	-0.74
17	8.83	76.00	22.70	1.89
18	5.10	47.00	21.50	-0.42
19	5.81	43.00	20.70	0.62
20	4.65	30.00	18.90	0.35
21	6.82	58.00	23.90	0.37
22	6.28	78.00	24.30	-1.07
23	5.15	49.00	23.80	-0.91
24	2.92	36.00	19.60	-1.76
25	9.27	67.00	24.30	2.37
26	5.57	42.00	22.00	0.16
27	4.92	29.00	22.50	-0.06
28	6.72	33.00	24.10	1.25
29	5.57	42.00	22.70	0.02
30	6.25	66.00	27.30	-1.21

A good graphical way of checking the normality assumption for the residuals is to plot the residuals against normal scores. To calculate the normal scores, we extend the recently created LSF file to include these as well.

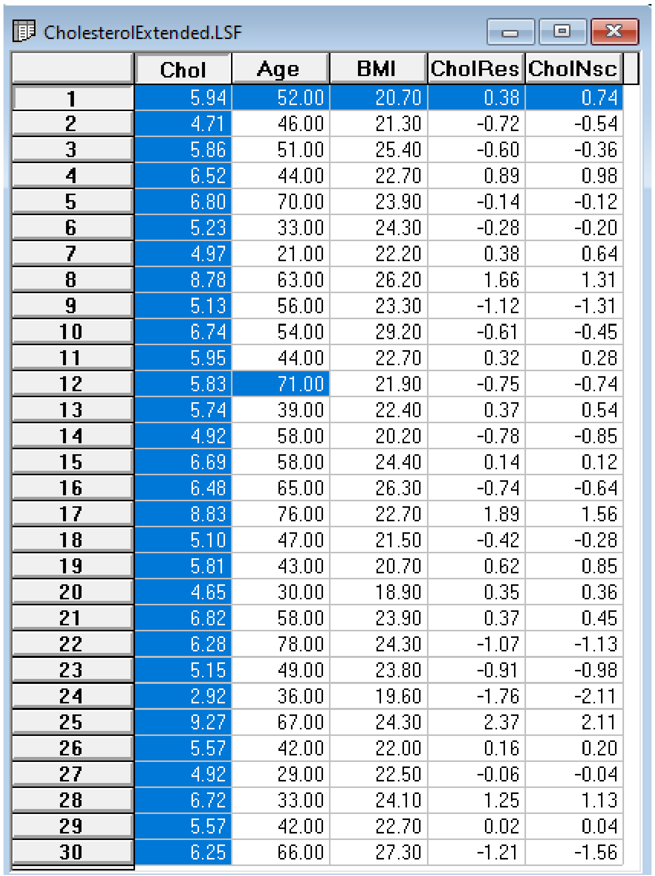


```
System File CholesterolwithRes.lsf
New CholNsc = CholRes
NS CholNsc
Output RA=CholesterolExtended.lsf
```

The univariate summary statistics for the recently created CholRes and CholNsc are reported in the output file.

```
Univariate Summary Statistics for Continuous Variables

 Variable      Mean  St. Dev.  Skewness  Kurtosis  Minimum Freq.  Maximum Freq.
 --------      ----  --------  --------  --------  ------- -----  ------- -----
     Chol     6.005     1.309     0.671     1.591    2.920     1    9.270     1
      Age    50.700    14.770     0.028    -0.718   21.000     1   78.000     1
      BMI    23.180     2.268     0.526     0.659   18.900     1   29.200     1
  CholRes    -0.000     0.957     0.632     0.310   -1.762     1    2.372     1
  CholNsc     0.000     0.957     0.000    -0.073   -2.106     1    2.106     1
```

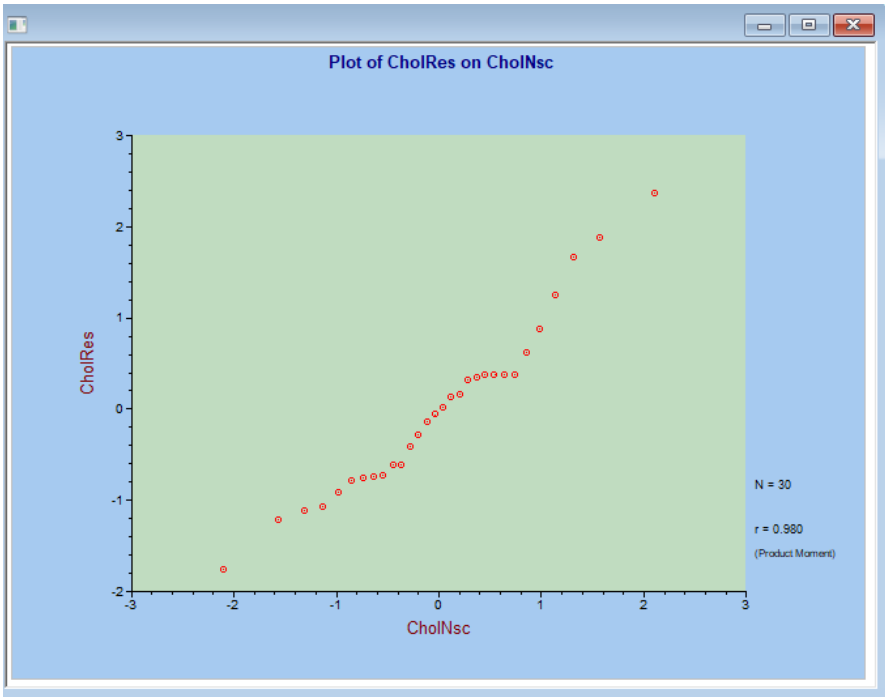
and the extended data file now contains 5 variables:



CholesterolExtended.LSF

	Chol	Age	BMI	CholRes	CholNsc
1	5.94	52.00	20.70	0.38	0.74
2	4.71	46.00	21.30	-0.72	-0.54
3	5.86	51.00	25.40	-0.60	-0.36
4	6.52	44.00	22.70	0.89	0.98
5	6.80	70.00	23.90	-0.14	-0.12
6	5.23	33.00	24.30	-0.28	-0.20
7	4.97	21.00	22.20	0.38	0.64
8	8.78	63.00	26.20	1.66	1.31
9	5.13	56.00	23.30	-1.12	-1.31
10	6.74	54.00	29.20	-0.61	-0.45
11	5.95	44.00	22.70	0.32	0.28
12	5.83	71.00	21.90	-0.75	-0.74
13	5.74	39.00	22.40	0.37	0.54
14	4.92	58.00	20.20	-0.78	-0.85
15	6.69	58.00	24.40	0.14	0.12
16	6.48	65.00	26.30	-0.74	-0.64
17	8.83	76.00	22.70	1.89	1.56
18	5.10	47.00	21.50	-0.42	-0.28
19	5.81	43.00	20.70	0.62	0.85
20	4.65	30.00	18.90	0.35	0.36
21	6.82	58.00	23.90	0.37	0.45
22	6.28	78.00	24.30	-1.07	-1.13
23	5.15	49.00	23.80	-0.91	-0.98
24	2.92	36.00	19.60	-1.76	-2.11
25	9.27	67.00	24.30	2.37	2.11
26	5.57	42.00	22.00	0.16	0.20
27	4.92	29.00	22.50	-0.06	-0.04
28	6.72	33.00	24.10	1.25	1.13
29	5.57	42.00	22.70	0.02	0.04
30	6.25	66.00	27.30	-1.21	-1.56

A scatter plot of CholRes against CholNsc shows that the residuals are not perfectly normal. If they were, the points in this scatterplot would fall on a straight line.



The standard output file also includes tests of the univariate normality for continuous variables in the model. None of the *p*-values in the table for the CholRes variable is smaller than 0.05, suggesting that the assumption of normality of the residuals seems to hold.

```
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

     Chol   1.589   0.112     1.670   0.095        5.311   0.070
      Age   0.070   0.944    -0.976   0.329        0.957   0.620
      BMI   1.268   0.205     0.955   0.340        2.520   0.284
  CholRes   1.504   0.133     0.603   0.546        2.626   0.269
  CholNsc   0.000   1.000     0.135   0.892        0.018   0.991
```

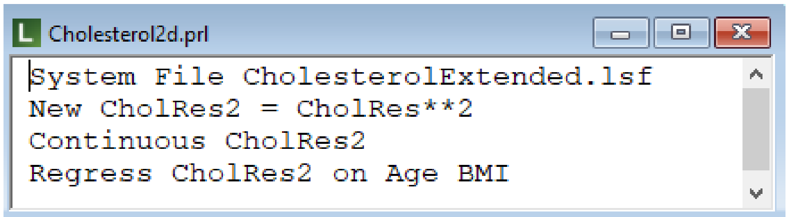
3.3 Checking homoscedasticity

Implied in the assumption that the residuals are i.i.d. $N(0,\sigma^2)$ is the assumption that the residual variance is constant over all observations. This can be mathematically expressed as

$$E(z_i^2 \mid \mathbf{x}) = \sigma^2$$

If this is not the case, the residuals are said to be heteroscedastic. To check for heteroscedasticity of the residuals, we again employ PRELIS to run a regression of the estimated squared residuals $\hat{z}_i^2$ on all covariates.

The syntax file to do so is very simple:

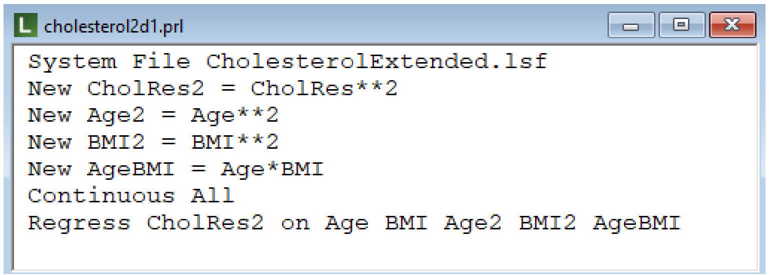


```
System File CholesterolExtended.lsf
New CholRes2 = CholRes**2
Continuous CholRes2
Regress CholRes2 on Age BMI
```

The estimated regression equation for this model is

```
Estimated Equations

 CholRes2 =  - 0.305 + 0.0339*Age - 0.0228*BMI + Error, R² = 0.139
 Standerr     (2.381) (0.0171)     (0.111)
 t-values     -0.128   1.981       -0.205
 P-values      0.899   0.057        0.839

 Error Variance = 1.552
```

Neither BMI nor Age has a statistically significant estimated coefficient. This supports the assumption of homoscedasiticity. We could also check the effect of the quadratic terms Age2 and BMI2, along with the interaction term AgeBMI by amending the syntax to



```
System File CholesterolExtended.lsf
New CholRes2 = CholRes**2
New Age2 = Age**2
New BMI2 = BMI**2
New AgeBMI = Age*BMI
Continuous All
Regress CholRes2 on Age BMI Age2 BMI2 AgeBMI
```

The results for this analysis show no significant *p*-value for any of the five coefficients, lending more weight for the homoscedasticity assumption.

```
Estimated Equations

 CholRes2 = 6.564  - 0.227*Age - 0.0744*BMI + 0.000774*Age2 - 0.00734*BMI2
 Standerr  (18.997) (0.258)     (1.639)      (0.00109)       (0.0400)
 t-values   0.346   -0.878      -0.0454       0.711          -0.184
 P-values   0.733    0.388       0.964        0.484           0.856

             + 0.00788*AgeBMI + Error, R² = 0.188
              (0.0118)
               0.666
               0.511

 Error Variance = 1.647
```

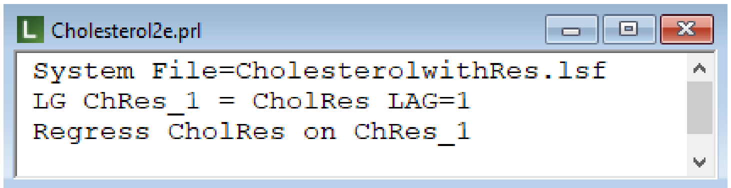
3.4 Checking autocorrelation

To test the assumption of independence among the residuals, we test for autocorrelation in the error term. To do so, we illustrate how one can lag a variable and estimate the regression on the lagged variable.

The following table shows the estimated residuals $\hat{z}_i$ and $\hat{z}_{i-1}$:

z_1	
z_2	z_1
z_3	z_2
$\vdots$	$\vdots$
z_N	z_{N-1}

The syntax file **Cholesterol2e.prl** shows how to obtain a lagged residual using the LG command.



```
System File=CholesterolwithRes.lsf
LG ChRes_1 = CholRes LAG=1
Regress CholRes on ChRes_1
```

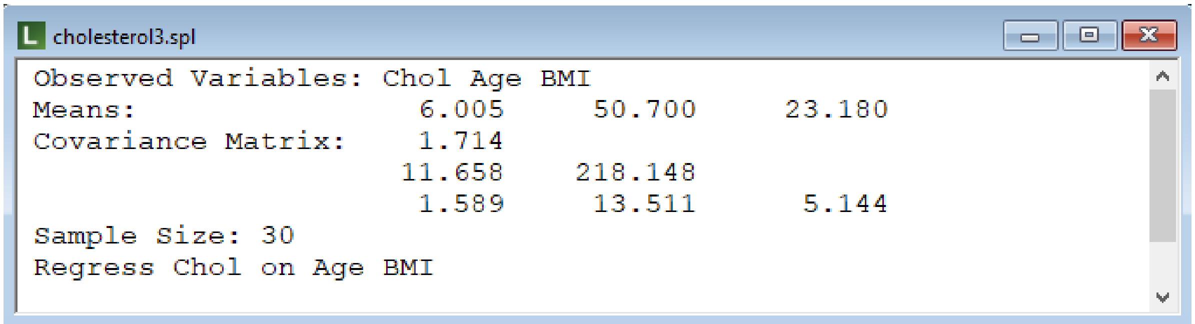
The results for the autoregression are

```
Estimated Equations

  CholRes = 0.0521 - 0.243*ChRes_1 + Error, R² = 0.0577
 Standerr  (0.176)  (0.189)
 t-values   0.296   -1.286
 P-values   0.769    0.209

 Error Variance = 0.895
```

The estimated autoregression coefficient of -0.243 is not statistically significant, indicating the absence of autocorrelation in the residual CholRes. Since the variances of the residual and the lagged residual are approximately equal, -0.243 is an approximate estimate of the residual autocorrelation.

3.5 Regression using means, variances and covariances

Instead of using the raw data, we can also use the means, variances and covariances of the variables to obtain regression estimates. These statistics are so-called sufficient statistics, which means that the individual data provides no further information that is already captured in these three statistics. This holds under the assumption of normality. To illustrate this, we use the SIMPLIS syntax file **cholesterol3.spl**.



```
Observed Variables: Chol Age BMI
Means:                 6.005     50.700     23.180
Covariance Matrix:     1.714
                      11.658    218.148
                       1.589     13.511      5.144
Sample Size: 30
Regress Chol on Age BMI
```

The mean and covariance matrix can also be stored in an external file instead of in the body of the SIMPLIS syntax. The SIMPLIS file **cholesterol4.spl** illustrates this and is equivalent to the one considered here.

```
Estimated Equations

     Chol =  - 0.738 + 0.0410*Age + 0.201*BMI + Error, R² = 0.465
 Standerr     (1.897) (0.0136)     (0.0888)
 Z-values     -0.389   3.006        2.267
 P-values      0.697   0.003        0.023
 Error Variance = 0.916
```

The output obtained for this run is exactly the same as obtained for the very first model we fitted in this document.