1 Objectif

Linear Regression. Reading the results.

The aim of the multiple regression is to predict the values of a continuous dependent variable Y from a set of continuous or binary (or binarized) independent variables (X1, … , Xp).

We written the model as follows (n is the number of available instances)

$$y_i = a_1 x_{i1} + a_2 x_{i2} + \cdots + a_p x_{ip} + a_{p+1} + \varepsilon_i \;,\; i = 1, \ldots, n$$

ε_i is the error term of the model. It reflects all the information about Y that is not provided by independent variables.

The coefficients $(a_1, a_2, \ldots, a_p, a_{p+1})$ of the model are estimated from a learning set. We want that the fitted values (the predicted values of the model) are as close as possible of the observed values of the dependent variable. For that, the criterion used is the residual sum of squares (RSS; SCR "somme des carrés des résidus" in French). The estimated coefficients of the regression minimize this criterion.

$$SCR = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

For the instance n°i, the residual ($\hat{\varepsilon}_i = \hat{y}_i - y_i$) is the difference between the predicted value and the observed value of the dependent variable.

We can analyze the model in various ways: (1) what is the global quality of the model? (2) The model is globally significant? (3) What coefficients are significant i.e. what are the independent variables which provide useful information about the dependent variable? (4) Can we test simultaneously the influence of several independent variables? (5) How can we compute a prediction interval for a predicted value from the model?

In this tutorial, we want to model the relationship between the cars consumption and their weight, engine-size and horsepower. We describe the outputs of Tanagra by associating them with the used formulas. We highlight the importance of the unscaled covariance matrix of the estimated coefficients [(X'X)$^{-1}$] (**Tanagra 1.4.38 and later**). It is used for the subsequent analysis: individual significance of coefficients, simultaneous significance of several coefficients, testing linear combinations of coefficients, computation of the standard error for the prediction interval. These analyses are performed into the Excel spreadsheet.

Thereafter, we perform the same analyses with the **R software**. We identify the objects provided by the ***lm(.)*** procedure that we can use in the same context.

2 Dataset

The dataset size is n = 28 instances. We want to explain the cars consumption (« Y - consumption », in l / 100 km) from the engine size (« X1 - eng.size », in cm^3), the horsepower (« X2 - horsepower », in ch) and the weight (« X3 - weight », in kg).

eng.size	horsepower	weight	consumption
846	32	650	5.7
993	39	790	5.8
899	29	730	6.1
1390	44	955	6.5
1195	33	895	6.8
658	32	740	6.8
1331	55	1010	7.1
1597	74	1080	7.4
1761	74	1100	9
2165	101	1500	11.7
1983	85	1075	9.5
1984	85	1155	9.5
1998	89	1140	8.8
1580	65	1080	9.3
1390	54	1110	8.6
1396	66	1140	7.7
2435	106	1370	10.8
1242	55	940	6.6
2972	107	1400	11.7
2958	150	1550	11.9
2497	122	1330	10.8
1998	66	1300	7.6
2496	125	1670	11.3
1998	89	1560	10.8
1997	92	1240	9.2
1984	85	1635	11.6
2438	97	1800	12.8
2473	125	1570	12.7

Using a matrix representation, the vector of dependent values is (vector size is n = 28 x 1)

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 5.7 \\ 5.8 \\ \vdots \\ 12.7 \end{pmatrix}$$

About the independent variables, we have the following matrix. We observe the last column which represents the constant (the intercept) into the regression (matrix size is n x p+1 =4).

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} & 1 \\ x_{21} & & x_{2p} & 1 \\ \vdots & & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} & 1 \end{pmatrix} = \begin{pmatrix} 846 & \cdots & 650 & 1 \\ 993 & & 790 & 1 \\ \vdots & & \vdots & \vdots \\ 2473 & \cdots & 1570 & 1 \end{pmatrix}$$

The OLS (ordinary least squares) estimators of the regression are

$$\boxed{\hat{a} = (X'X)^{-1}X'Y}$$

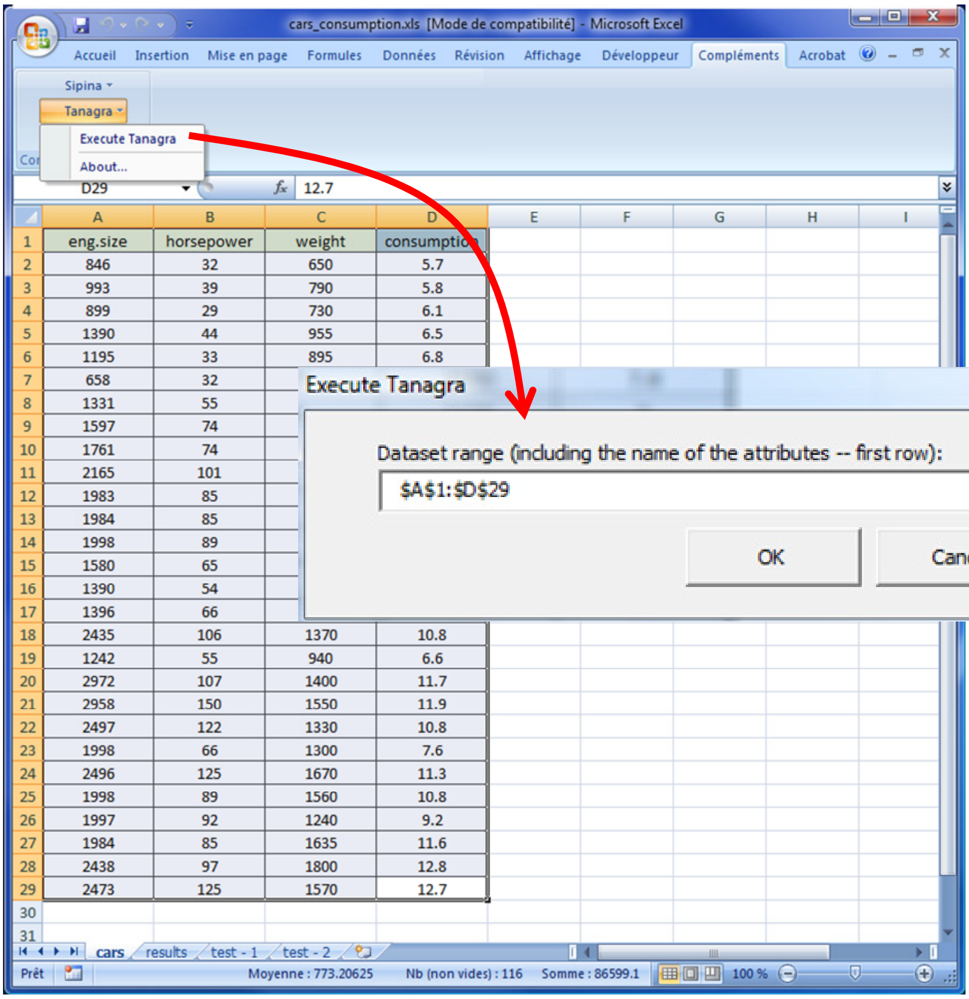
X' is the transposition of X; $(X'X)^{-1}$ is the inverse of (X'X). The fitted values are obtained as follows.

$$\hat{Y} = X'\hat{a} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix}$$

3 Linear regression with Tanagra

3.1 Importing the data file

We open the « cars_consumption.xls » into Excel. We select the range of values,including the name of variables, then we activate the TANAGRA / EXECUTE TANAGRA menu[1] to launch TANAGRA.



	A	B	C	D
1	eng.size	horsepower	weight	consumptio
2	846	32	650	5.7
3	993	39	790	5.8
4	899	29	730	6.1
5	1390	44	955	6.5
6	1195	33	895	6.8
7	658	32		
8	1331	55		
9	1597	74		
10	1761	74		
11	2165	101		
12	1983	85		
13	1984	85		
14	1998	89		
15	1580	65		
16	1390	54		
17	1396	66		
18	2435	106	1370	10.8
19	1242	55	940	6.6
20	2972	107	1400	11.7
21	2958	150	1550	11.9
22	2497	122	1330	10.8
23	1998	66	1300	7.6
24	2496	125	1670	11.3
25	1998	89	1560	10.8
26	1997	92	1240	9.2
27	1984	85	1635	11.6
28	2438	97	1800	12.8
29	2473	125	1570	12.7

Dataset range (including the name of the attributes -- first row): $A$1:$D$29

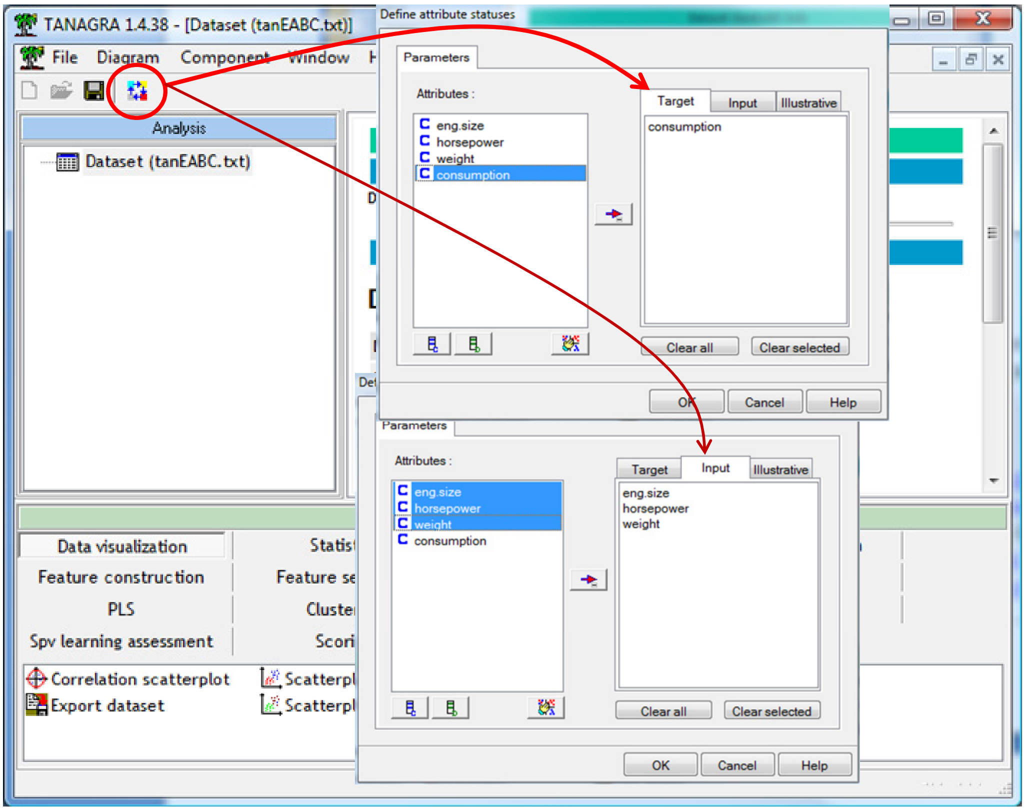
Moyenne : 773.20625 Nb (non vides) : 116 Somme : 86599.1

TANAGRA is automatically launched. The dataset is loaded. We check that the dataset (n = 28 instances and 4 variables) is correctly imported.

3.2 Linear regression

First of all, we must specify the TARGET attribute (Y) and the INPUT ones (X1, X2, X3). To do that, we insert the DEFINE STATUS component into the diagram.

[1] See http://data-mining-tutorials.blogspot.com/2010/08/tanagra-add-in-for-office-2007-and.html for the incorporation of the Tanagra.xla add-in under Excel 2007 and Excel 2010 ; for the previous version of Excel, see http://data-mining-tutorials.blogspot.com/2008/10/excel-file-handling-using-add-in.html. Tanagra can be incorporated also into Open Office Calc under Windows (http://data-mining-tutorials.blogspot.com/2008/10/ooocalc-file-handling-using-add-in.html) or Linux (http://data-mining-tutorials.blogspot.com/2009/04/launching-tanagra-from-oocalc-under.html).

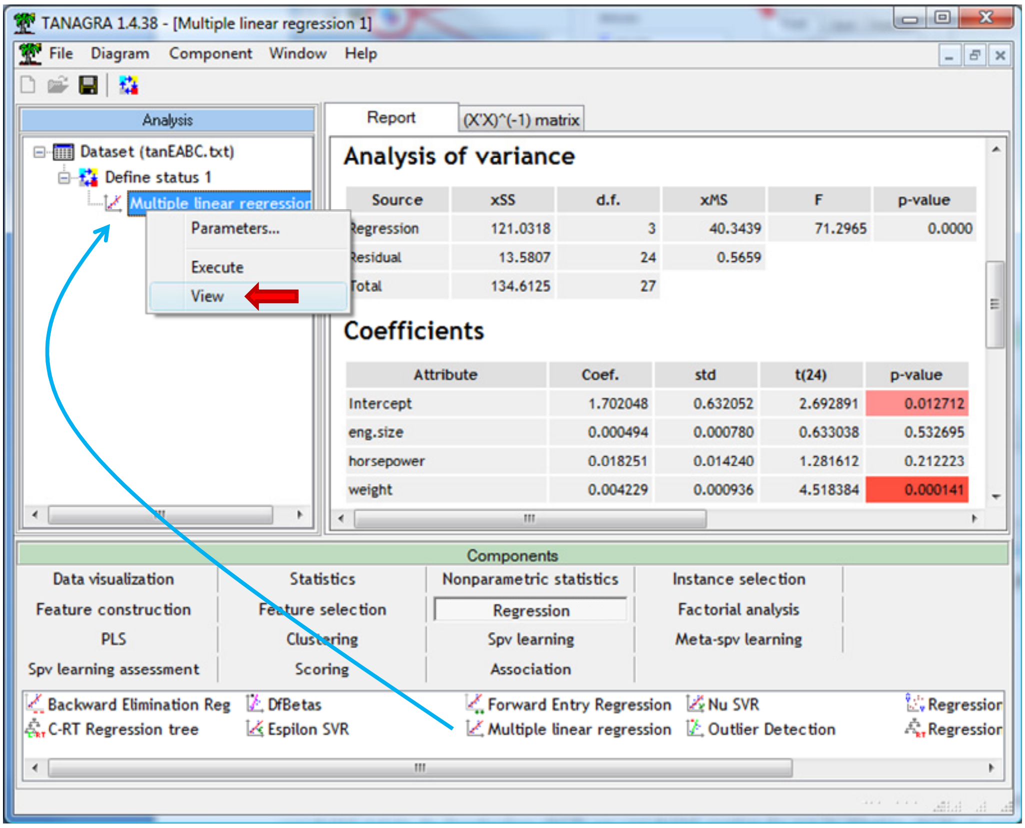


Attributes : eng.size, horsepower, weight, consumption

Target: consumption

Input: eng.size, horsepower, weight

Then we add the MULTIPLE LINEAR REGRESSION component (REGRESSION tab) into the diagram.



Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	121.0318	3	40.3439	71.2965	0.0000
Residual	13.5807	24	0.5659		
Total	134.6125	27			

Coefficients

Attribute	Coef.	std	t(24)	p-value
Intercept	1.702048	0.632052	2.692891	0.012712
eng.size	0.000494	0.000780	0.633038	0.532695
horsepower	0.018251	0.014240	1.281612	0.212223
weight	0.004229	0.000936	4.518384	0.000141

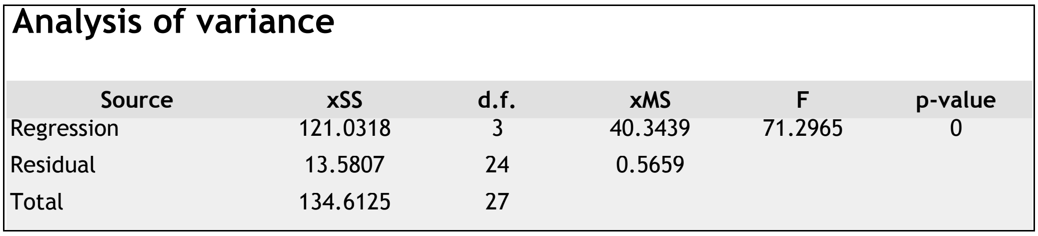
We activate the contextual VIEW menu. We obtain the results of the linear regression process.

3.3 Global significance of the regression

3.3.1 ANOVA table and the coefficient of determination R²

The ANOVA table partitions the total sum of squares of the dependent variables (SCT: somme des carrés totaux in French) in the sum of squares explained by the model (SCE in French) and the residual sum of squares (CRS in French) not explained by the model. The ratio between SCE and SCT is called the coefficient of determination. It reflects the proportion of the variance of the dependent variable explained by the model.

TANAGRA provides the following table.



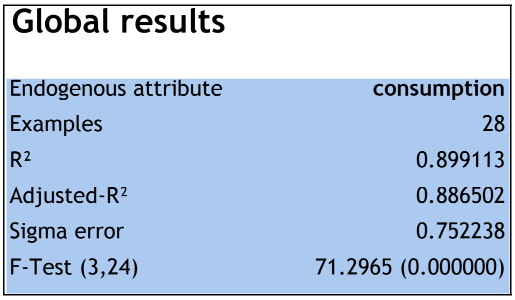
Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	121.0318	3	40.3439	71.2965	0
Residual	13.5807	24	0.5659		
Total	134.6125	27			

R² is defined as follows

$$R^2 = \frac{SCE}{SCT} = \frac{121.0318}{134.6125} = 0.899113 = 1 - \frac{SCR}{SCT} = 1 - \frac{13.5807}{134.6125}$$

We observe the R² into the « GLOBAL RESULTS » table.



Global results

Endogenous attribute	**consumption**
Examples	28
R²	0.899113
Adjusted-R²	0.886502
Sigma error	0.752238
F-Test (3,24)	71.2965 (0.000000)

Tanagra provides also the Adjusted-R² which adjusts the R² with the degree of freedom. It is preferable to use the adjusted value for the comparison of models with different number of independent variables. For instance, when we compare nested models, we want to detect if the additional independent variables are relevant.

$$\begin{aligned}\bar{R}^2 &= 1 - \frac{SCR/_{n-p-1}}{SCT/_{n-1}} = 1 - \frac{13.5807/_{24}}{134.6125/_{27}} = 0.886502 \\ &= 1 - \frac{n-1}{n-p-1}\left(1 - R^2\right) = 1 - \frac{27}{24}\left(1 - 0.899113\right)\end{aligned}$$

3.3.2 F-test – Testing the global significance

We want to test the significance of the model as a whole. Formally, we specify the following hypothesis testing.

$$\begin{cases} H0 : a_1 = a_2 = \cdots = a_p = 0 \\ H1 : \exists \mathrm{j}, \mathrm{a}_\mathrm{j} \neq 0 \end{cases}$$

The used test statistic is,

$$F = \frac{CME}{CMR} = \frac{SCE/p}{SCR/n-p-1} = \frac{40.3439}{0.5659} = 71.2965$$

We find the result provided by the table above. Under the null hypothesis, F follows a Fisher distribution with (p= 3, n-p-1 = 24) degrees of freedom. Tanagra provides directly the p-value i.e.

$$p-value = \Pr(Fisher \geq F)$$

If the p-value is lower than the significance level of test α, we decide that the model is significant. Here, we have p-value#0, the model is highly significant.

3.4 T-test to assess the significance of individual coefficients

The next step is to assess the influence of the independent variables in the model. For each coefficient associated to an independent variable, we test the following null hypothesis:

$$\begin{cases} H0: a_j = 0 \\ H1: a_j \neq 0 \end{cases}$$

The statistical test is

$$t_j = \frac{\hat{a}_j}{\hat{\sigma}_{\hat{a}_j}}$$

$\hat{\sigma}_{\hat{a}_j}$ is the standard error of the estimated coefficient. Its squared value is obtained on the diagonal of the covariance matrix of the estimated coefficients.

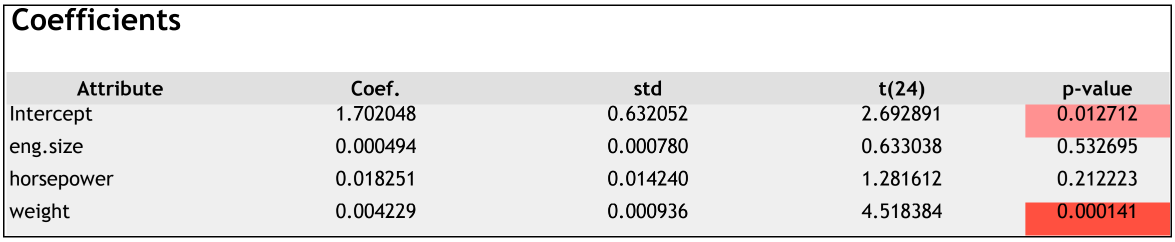
$$\hat{\Omega}_{\hat{a}} = \hat{\sigma}_{\varepsilon}^2 (X'X)^{-1}$$

$\hat{\sigma}_{\varepsilon}^2$ is the squared of the standard error of regression. The standard error of the regression is obtained with the following formula (« GLOBAL RESULTS » - Sigma error).

$$\hat{\sigma}_{\varepsilon} = \sqrt{\frac{SCR}{n-p-1}} = \sqrt{\frac{13.5807}{24}} = 0.752238$$

Into the "COEFFICIENTS" table, Tanagra provides both $\hat{\sigma}_{\hat{a}_j}$ and t_j. Under the null hypothesis, the statistical test follows a Student distribution with (n – p – 1 = 24) degrees of freedom. Into the last column, we obtained the p-value of the test.

We reject the null hypothesis if the p-value is lower than the significance level α (often α = 5%). Here, WEIGHT seems the only relevant variable in the regression, with a p-value equal to 0.000141.

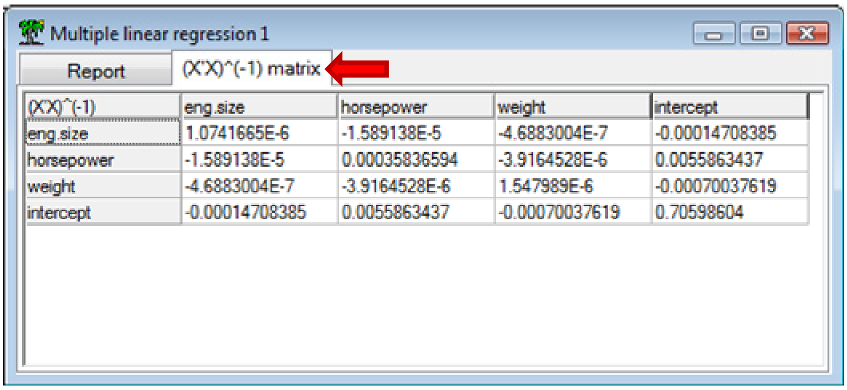


Coefficients

Attribute	Coef.	std	t(24)	p-value
Intercept	1.702048	0.632052	2.692891	0.012712
eng.size	0.000494	0.000780	0.633038	0.532695
horsepower	0.018251	0.014240	1.281612	0.212223
weight	0.004229	0.000936	4.518384	0.000141

3.5 Covariance matrix of the parameter estimates

The whole covariance matrix of the estimated parameters can be obtained from the (X'X)$^{-1}$ matrix provided by Tanagra into the second tab of the visualization window.



(X'X)^(-1)	eng.size	horsepower	weight	intercept
eng.size	1.0741665E-6	-1.589138E-5	-4.6883004E-7	-0.00014708385
horsepower	-1.589138E-5	0.00035836594	-3.9164528E-6	0.0055863437
weight	-4.6883004E-7	-3.9164528E-6	1.547989E-6	-0.00070037619
intercept	-0.00014708385	0.0055863437	-0.00070037619	0.70598604

We multiply this one by the square of the standard error of the regression $\hat{\Omega}_{\hat{a}} = \hat{\sigma}_{\varepsilon}^{2}\left(X'X\right)^{-1}$

MVCV(a^)	eng.size	horsepower	weight	intercept
eng.size	6.07830E-07	-8.99233E-06	-2.65293E-07	-8.32291E-05
horsepower	-8.99233E-06	2.02786E-04	-2.21617E-06	3.16110E-03
weight	-2.65293E-07	-2.21617E-06	8.75948E-07	-3.96316E-04
intercept	-8.32291E-05	3.16110E-03	-3.96316E-04	3.99491E-01

The variance of the estimated parameters are on the main diagonal of the matrix. For the coefficient of HORSEPOWER (X2) for instance, we obtain the standard error with

$$\hat{\sigma}_{\hat{a}_2} = \sqrt{2.02786 \times 10^{-4}} = 1.42403 \times 10^{-2} = 0.01424$$

We observe the same value into the COEFFICIENTS table.

3.6 Comparison of coefficients to hypothetical values

The test of significance for coefficients is a particular case of generalized tests for simultaneous comparisons. The hypothesis testing can be written as follows.

$$\begin{cases} H_0 : \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_q \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_q \end{pmatrix} \Leftrightarrow a_{(q)} = c_{(q)} \\ H_1 : \exists j / a_j \neq c_j \end{cases}$$

The statistical test F follows a Fisher distribution with (q, n-p-1) degrees of freedom under Ho.

$$F = \frac{1}{q}\left[\hat{a}_{(q)} - c_{(q)}\right]'\hat{\Omega}^{-1}_{\hat{a}_{(q)}}\left[\hat{a}_{(q)} - c_{(q)}\right]$$

The critical region corresponds to the unusual high values of F.

We want to test the simultaneous significance of the coefficients related to ENG.SIZE and HORSEPOWER. These coefficients seem not individually significant at 5% level (see section 3.4). But these tests do not consider the possible interaction between the coefficients (their covariance). Into the covariance matrix $\hat{\Omega}_{\hat{a}}$, we observe that $\hat{\text{cov}}(\hat{a}_{eng.size}, \hat{a}_{horsepower}) = -8.9923 \times 10^{-6}$. The simultaneous test incorporates this information. Into the covariance matrix, we highlight below the values related to the variance and the covariance for the two coefficients.

MVCV(a^)	eng.size	horsepower	weight	intercept
eng.size	6.0783E-07	-8.9923E-06	-2.6529E-07	-8.3229E-05
horsepower	-8.9923E-06	2.0279E-04	-2.2162E-06	3.1611E-03
weight	-2.6529E-07	-2.2162E-06	8.7595E-07	-3.9632E-04
intercept	-8.3229E-05	3.1611E-03	-3.9632E-04	3.9949E-01

Then, we have the following sub matrix

$$\hat{\Omega}_{\hat{a}_{(q)}} = \begin{pmatrix} 6.0783 \times 10^{-7} & -8.9923 \times 10^{-6} \\ -8.9923 \times 10^{-6} & 2.0279 \times 10^{-4} \end{pmatrix}$$

We compute the inverse of this matrix

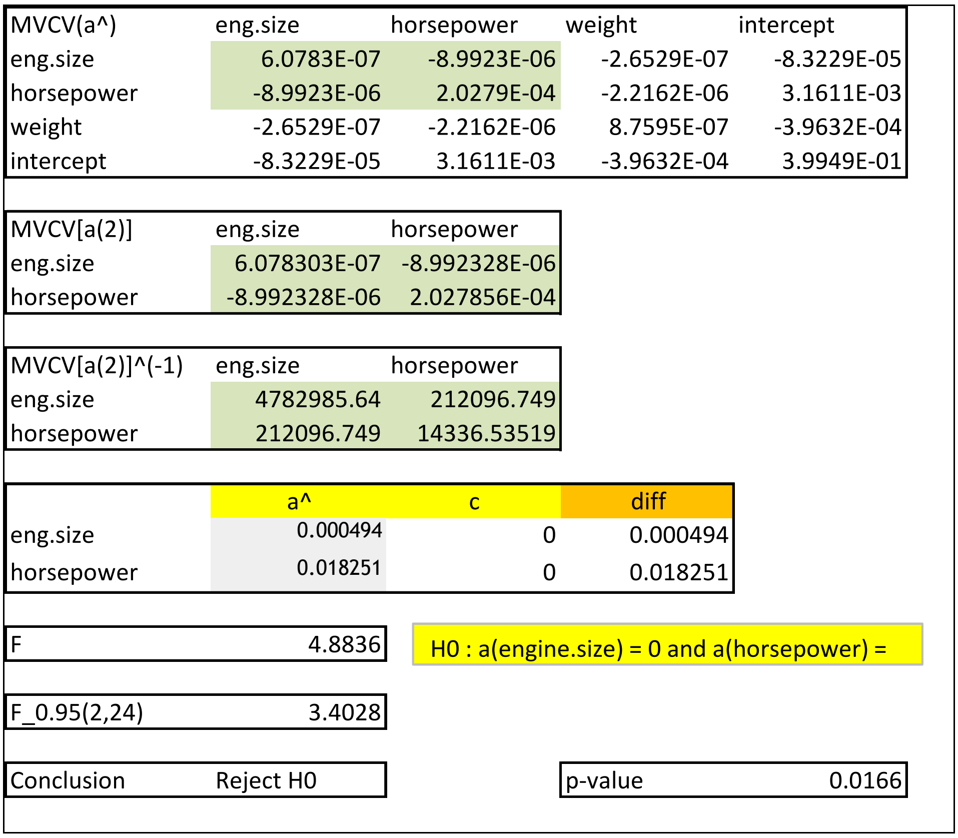
$$\hat{\Omega}^{-1}_{\hat{a}_{(q)}} = \begin{pmatrix} 4782985.64 & 212096.749 \\ 212096.749 & 14336.53519 \end{pmatrix}$$

We obtain the statistical test F:

$$\begin{aligned} F &= \frac{1}{q}\left[\hat{a}_{(q)} - c_{(q)}\right]'\hat{\Omega}^{-1}_{\hat{a}_{(q)}}\left[\hat{a}_{(q)} - c_{(q)}\right] \\ &= \frac{1}{2}\left[\begin{pmatrix} 0.000494 \\ 0.018251 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right]'\hat{\Omega}^{-1}_{\hat{a}_{(q)}}\left[\begin{pmatrix} 0.000494 \\ 0.018251 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right] = 4.8836 \end{aligned}$$

The p-value of the test is 0.0166. It is lower than the significance level 5%. Unlike the individual tests (the two coefficients did not appear significant individually), we reject the nullity of the two parameters simultaneously. Considering the relationship between the coefficients through their covariance leads us to take a different decision. This suggests that these variables are highly correlated. They interfere each other into the model.

All these calculations are performed into an Excel sheet that we detail below.



MVCV(a^)	eng.size	horsepower	weight	intercept
eng.size	6.0783E-07	-8.9923E-06	-2.6529E-07	-8.3229E-05
horsepower	-8.9923E-06	2.0279E-04	-2.2162E-06	3.1611E-03
weight	-2.6529E-07	-2.2162E-06	8.7595E-07	-3.9632E-04
intercept	-8.3229E-05	3.1611E-03	-3.9632E-04	3.9949E-01

MVCV[a(2)]	eng.size	horsepower
eng.size	6.078303E-07	-8.992328E-06
horsepower	-8.992328E-06	2.027856E-04

MVCV[a(2)]^(-1)	eng.size	horsepower
eng.size	4782985.64	212096.749
horsepower	212096.749	14336.53519

	a^	c	diff
eng.size	0.000494	0	0.000494
horsepower	0.018251	0	0.018251

F	4.8836
F_0.95(2,24)	3.4028
Conclusion	Reject H0
p-value	0.0166

H0 : a(engine.size) = 0 and a(horsepower) =

3.7 Testing linear combinations of coefficients

For the comparison to hypothetical values. The test above can be written using the following generic formulation

$$\begin{cases} H_0 : Ra = r \\ H_1 : Ra \neq r \end{cases}$$

The R matrix has q rows (the number of constraints) and (p+1) columns (the number of coefficients of the model). The r vector has q rows (and 1 column of course). The main difficulty is to write correctly the R matrix and the r vector. For the simultaneous comparison above, we have

$$H_0 : \begin{pmatrix} a_{eng.size} \\ a_{horsepower} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

R and r are defined as follows:

$$R = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \text{ et } r = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Thus

$$Ra = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} a_{eng.size} \\ a_{horsepower} \\ a_{weight} \\ a_{cons\tan te} \end{pmatrix} = \begin{pmatrix} a_{eng.size} \\ a_{horsepower} \end{pmatrix}$$

Comparison of coefficients. But this new formulation has a wider scope. For instance, it allows to compare the values of coefficients into the model.

We note that the ratio between the horsepower and the engine size is about 40 in our dataset. We want to check if this ratio is preserved into the model. Thus, we want to check if

$$\begin{cases} H_0 : 1000 \times a_{eng.size} = 40 \times a_{horsepower} \\ H_1 : 1000 \times a_{eng.size} \neq 40 \times a_{horsepower} \end{cases}$$

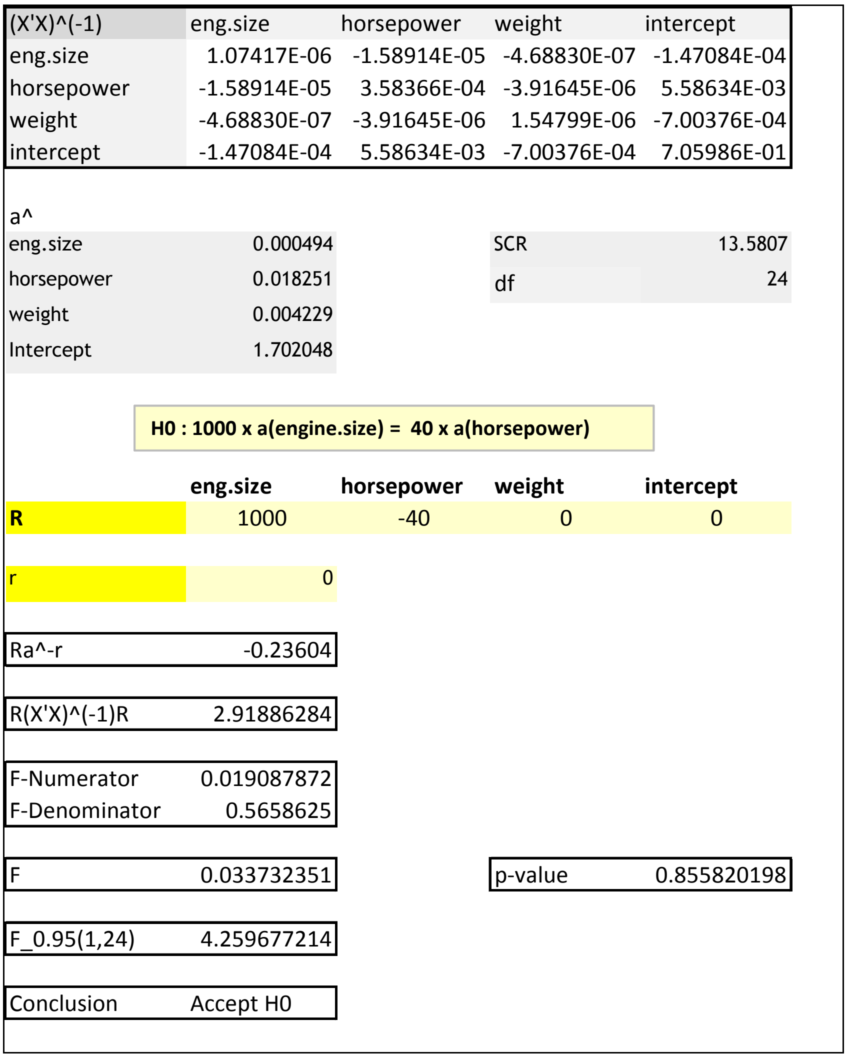
The matrix R and the vector r are written as follows

$$R = \begin{pmatrix} 1000 & -40 & 0 & 0 \end{pmatrix} \text{ and } r = \begin{pmatrix} 0 \end{pmatrix}$$

The F-statistic follows a Fisher distribution with (q, n – p – 1) degrees of freedom under Ho.

$$F = \frac{\frac{1}{q}(R\hat{a} - r)'\left[R(X'X)^{-1}R'\right]^{-1}(R\hat{a} - r)}{SCR / n - p - 1}$$

In our problem, q = 1; n – p – 1 = 24 is the degree of freedom of the regression. We described here the computation into an Excel sheet.



(X'X)^(-1)	eng.size	horsepower	weight	intercept
eng.size	1.07417E-06	-1.58914E-05	-4.68830E-07	-1.47084E-04
horsepower	-1.58914E-05	3.58366E-04	-3.91645E-06	5.58634E-03
weight	-4.68830E-07	-3.91645E-06	1.54799E-06	-7.00376E-04
intercept	-1.47084E-04	5.58634E-03	-7.00376E-04	7.05986E-01

a^	
eng.size	0.000494
horsepower	0.018251
weight	0.004229
Intercept	1.702048

SCR	13.5807
df	24

H0 : 1000 x a(engine.size) = 40 x a(horsepower)

	eng.size	**horsepower**	**weight**	**intercept**
R	1000	-40	0	0

r	0
Ra^-r	-0.23604
R(X'X)^(-1)R	2.91886284
F-Numerator	0.019087872
F-Denominator	0.5658625
F	0.033732351
F_0.95(1,24)	4.259677214
Conclusion	Accept H0

p-value	0.855820198

We obtain F = 0.0337. The p-value is 0.8585. We cannot reject the null hypothesis.

3.8 Prediction and prediction interval

The (X'X)$^{-1}$ matrix is also useful when we want to compute the prediction interval of Y.

For a new instance (i*) which is not used during the construction of the model, we can obtain the prediction of the model by applying the computed coefficients on the description (values of the independent variables). For instance, for the instance where (eng.size = 1984 ; horsepower = 85 ; weight = 1155), its description is the following:

$$x_{i*} = \begin{pmatrix}1984 & 85 & 1155 & 1\end{pmatrix}$$

We observe that we set the constant term in the last column. The prediction of the model is:

$$\hat{y}_{i*} = x_{i*}\,.\,\hat{a} = \begin{pmatrix}1984 & 85 & 1155 & 1\end{pmatrix}.\begin{pmatrix}0.000494\\0.018251\\0.004229\\1.702048\end{pmatrix} = 9.12$$

This predicted value is interesting in itself. But it is often more interesting to obtain a prediction interval for which we can attach a confidence level. To do that, we must obtain an estimation of the standard error of the error of prediction and its distribution.

The error of prediction is written as follows

$$\hat{\varepsilon}_{i*} = \hat{y}_{i*} - y_{i*}$$

The square of its standard error is

$$\hat{\sigma}^2_{\hat{\varepsilon}_{i*}} = \hat{\sigma}^2_{\varepsilon}\left(1 + x_{i*}\left(X'X\right)^{-1}x_{i*}'\right)$$

It depends on the standard error of the regression $\hat{\sigma}^2_{\varepsilon}$ (which is directly related to the residual sum of squares i.e. the quality of the model) and the leverage which measures how far the independent variables deviates from its mean [$h_{i*} = x_{i*}\left(X'X\right)^{-1}x_{i*}'$].

The standardized error follows a Student distribution with (n – p – 1) degrees of freedom.

$$\frac{\hat{\varepsilon}_{i*}}{\hat{\sigma}_{\varepsilon_{i*}}} \equiv \Im(n - p - 1)$$

For the (1 - α) confidence level, the prediction is defined as follows

$$\hat{y}_{i*} \pm t_{1-\alpha/2} \times \hat{\sigma}_{\hat{\varepsilon}_{i*}}$$

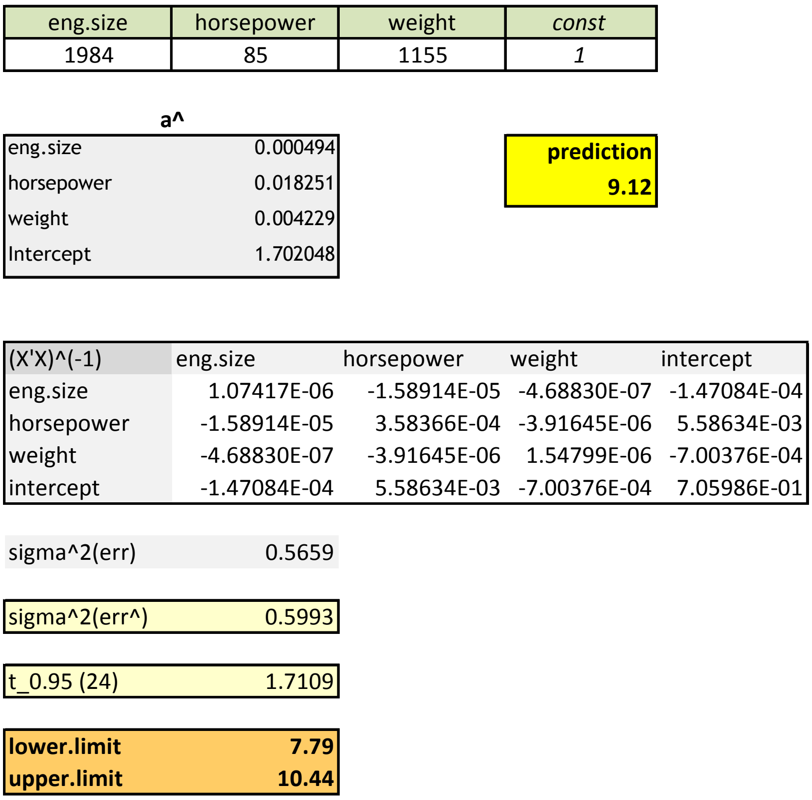
About the new car described above, we have:

$$\hat{\sigma}^2_{\hat{\varepsilon}_{i*}} = 0.752^2\left(1 + \begin{pmatrix}1984 & 85 & 1155 & 1\end{pmatrix}\left(X'X\right)^{-1}\begin{pmatrix}1984\\85\\115\\1\end{pmatrix}\right) = 0.5993$$

For the 90% confidence limits, the Student distribution table provides $t_{0.95} = 1.7109$, we obtain the following limits:

$$\begin{cases} y_{bb} = 9.12 - 1.7109 \times \sqrt{0.5993} = 7.79 \\ y_{bh} = 9.12 + 1.7109 \times \sqrt{0.5993} = 10.44 \end{cases}$$

All these calculations are described in an Excel sheet.



eng.size	horsepower	weight	*const*
1984	85	1155	*1*

a^

eng.size	0.000494
horsepower	0.018251
weight	0.004229
Intercept	1.702048

prediction
9.12

(X'X)^(-1)	eng.size	horsepower	weight	intercept
eng.size	1.07417E-06	-1.58914E-05	-4.68830E-07	-1.47084E-04
horsepower	-1.58914E-05	3.58366E-04	-3.91645E-06	5.58634E-03
weight	-4.68830E-07	-3.91645E-06	1.54799E-06	-7.00376E-04
intercept	-1.47084E-04	5.58634E-03	-7.00376E-04	7.05986E-01

sigma^2(err)	0.5659
sigma^2(err^)	0.5993
t_0.95 (24)	1.7109
lower.limit	**7.79**
upper.limit	**10.44**

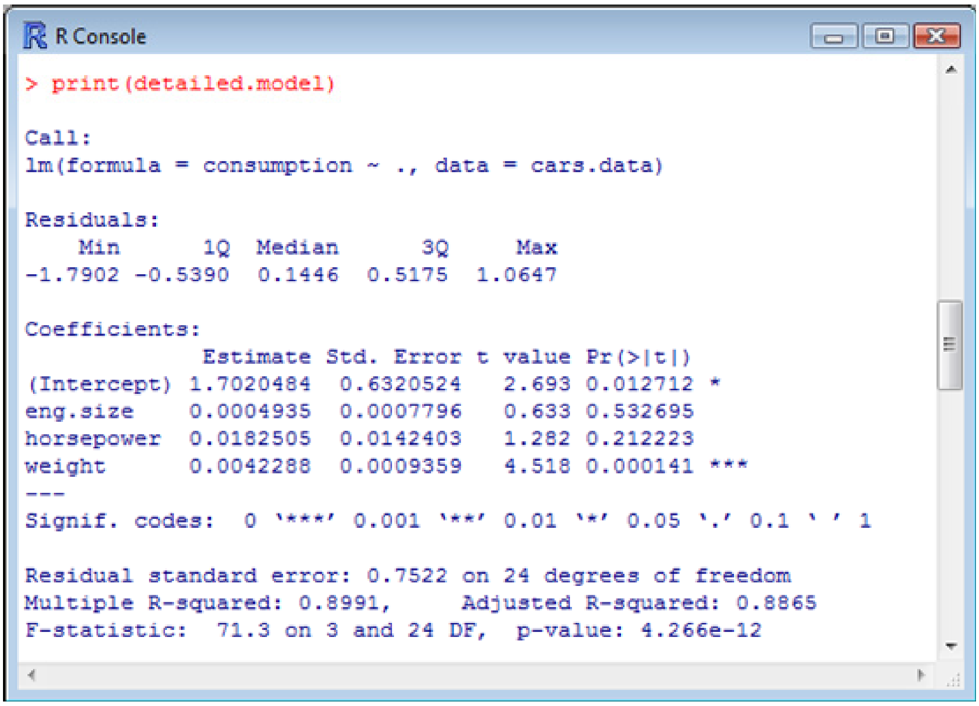
4 Multiple regression with R

4.1 Regression with the *lm(.)* procedure

We use the ***lm(.)*** procedure under R. We make use mainly the object provided by the associated ***summary(.)*** object. Here is the source code for the regression under R.

```
rm (list=ls())
#modifying the current directory
setwd("…")
#loading the dataset
library(xlsReadWrite)
cars.data <- read.xls(file="cars_consumption.xls",colNames=T,sheet=1)
#launching the regression
cars.model <-lm(consumption ~ ., data = cars.data)
print(cars.model)
#obtaining the summary.lm object
detailed.model <- summary(cars.model)
print(detailed.model)
```

The results are identical to those of Tanagra.

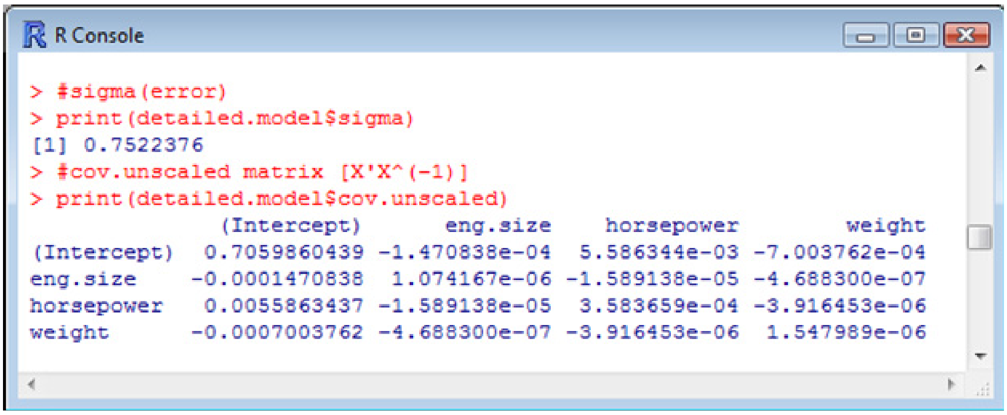


```
> print(detailed.model)

Call:
lm(formula = consumption ~ ., data = cars.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7902 -0.5390  0.1446  0.5175  1.0647 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.7020484  0.6320524   2.693 0.012712 *  
eng.size    0.0004935  0.0007796   0.633 0.532695    
horsepower  0.0182505  0.0142403   1.282 0.212223    
weight      0.0042288  0.0009359   4.518 0.000141 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.7522 on 24 degrees of freedom
Multiple R-squared: 0.8991,     Adjusted R-squared: 0.8865 
F-statistic:  71.3 on 3 and 24 DF,  p-value: 4.266e-12
```

The **summary.lm** provides various tools for the subsequent analyses. We can obtain, among others, the famous (X'X)$^{-1}$ matrix.

```
#standard error of regression
print(detailed.model$sigma)
#cov.unscaled is the [X'X^(-1)] matrix
print(detailed.model$cov.unscaled)
```

So, we have:



```
> #sigma(error)
> print(detailed.model$sigma)
[1] 0.7522376
> #cov.unscaled matrix [X'X^(-1)]
> print(detailed.model$cov.unscaled)
              (Intercept)      eng.size    horsepower        weight
(Intercept)  0.7059860439 -1.470838e-04  5.586344e-03 -7.003762e-04
eng.size    -0.0001470838  1.074167e-06 -1.589138e-05 -4.688300e-07
horsepower   0.0055863437 -1.589138e-05  3.583659e-04 -3.916453e-06
weight      -0.0007003762 -4.688300e-07 -3.916453e-06  1.547989e-06
```

The intercept is in the last position, unlike Tanagra. We must consider this information for the subsequent calculations.

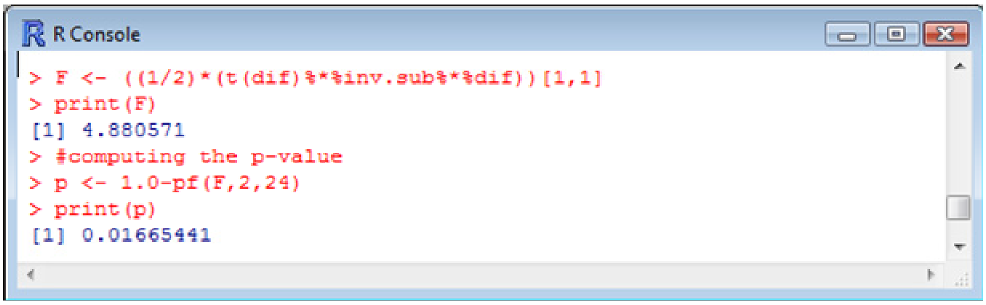
4.2 Comparison to hypothetical values

The main advantage of R is that we can perform all calculations into the same environment, with the associated programming language. For the test above (section 3.6), we set:

```
#getting the coefficients for eng.size and horsepower
a.test <- matrix(detailed.model$coefficients[2:3,1],nrow=2,ncol=1)
#the vector of the hypothetical values (the reference)
ref.test <- matrix(c(0,0),nrow=2,ncol=1)
```

```
#difference between the estimated coefficient and the reference
dif <- a.test - ref.test
#the whole covariance matrix of the parameter estimates
vcv.mat <- detailed.model$sigma^2 * detailed.model$cov.unscaled
#covariance sub matrix for eng.size and horsepower
sub.vcv.mat <- vcv.mat[2:3,2:3]
print(sub.vcv.mat)
#inversion of the sub matrix
inv.sub <- solve(sub.vcv.mat)
print(inv.sub)
#calculation of the Statistical test - it is a scalar
F <- ((1/2)*(t(dif)%*%inv.sub%*%dif))[1,1]
print(F)
#the associated p-value
p <- 1.0-pf(F,2,24)
print(p)
```

We obtain F = 4.880571, with p-value = 0.01665441. The very small difference from the results above is due to the lost of precision during the copying the estimated coefficients into the spreadsheet.



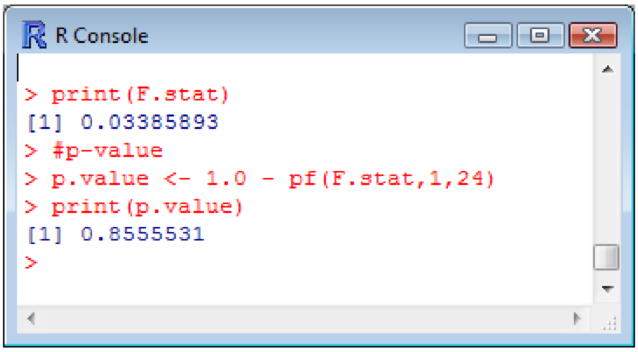
```
> F <- ((1/2)*(t(dif)%*%inv.sub%*%dif))[1,1]
> print(F)
[1] 4.880571
> #computing the p-value
> p <- 1.0-pf(F,2,24)
> print(p)
[1] 0.01665441
```

4.3 Testing linear combination of coefficients

In a similar fashion, we can implement the comparison of coefficients (section 3.7):

```
#estimated coefficients
a.hat <- matrix(detailed.model$coefficients[,1],nrow=4,ncol=1)
#the R matrix!!! Caution: the intercept is in the first column
R <- matrix(c(0,1000,-40,0),nrow=1,ncol=4)
#the vector r
r <- matrix(c(0),nrow=1,ncol=1)
#R(X'X)^(-1)R'
B <- R%*%detailed.model$cov.unscaled%*%t(R)
#B^(-1)
inv.B <- solve(B)
#numerator the statistical test
F_num <- (1/1)*(t(R%*%a.hat-r)%*%inv.B%*%(R%*%a.hat-r))[1,1]
#denominator of the statistical test
F_denom <- detailed.model$sigma^2
#F statistical test
F.stat <- F_num/F_denom
print(F.stat)
#p-value
p.value <- 1.0 - pf(F.stat,1,24)
print(p.value)
```

The constant is in the first column when we define the R matrix. We obtain the same results as above (section 3.7).



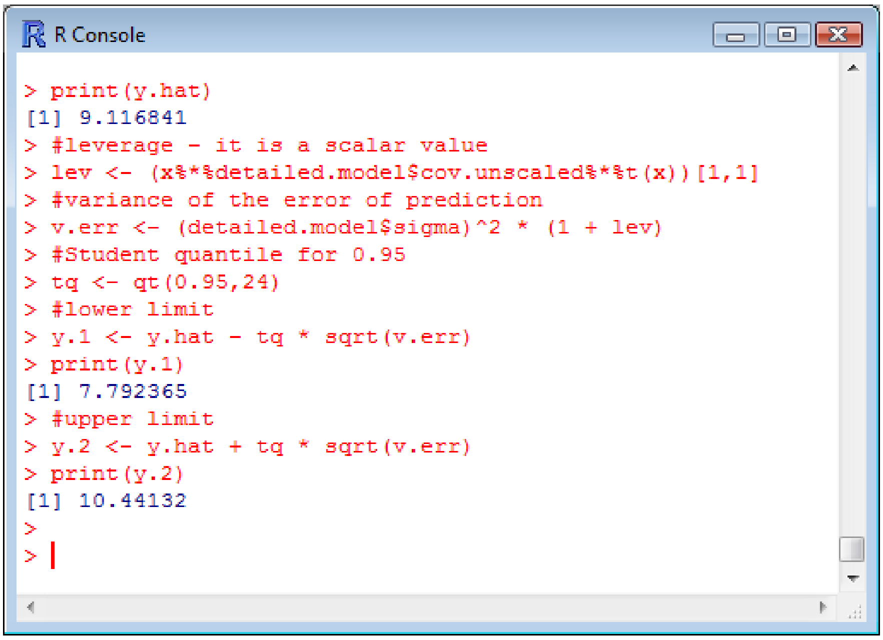
```
> print(F.stat)
[1] 0.03385893
> #p-value
> p.value <- 1.0 - pf(F.stat,1,24)
> print(p.value)
[1] 0.8555531
>
```

4.4 Prediction and prediction interval

The equivalent code for calculations in Excel is as follows.

```
#values for the independent variables (including the constant 1)
x <- matrix(c(1,1984,85,1155),nrow=1,ncol=4)
#prediction - it is a scalar
y.hat <- (x%*%a.hat)[1,1]
print(y.hat)
#the leverage - it is a scalar
lev <- (x%*%detailed.model$cov.unscaled%*%t(x))[1,1]
#square of the standard error of the prediction error
v.err <- (detailed.model$sigma)^2 * (1 + lev)
#quantile for the Student distribution
tq <- qt(0.95,24)
#lower limit
y.1 <- y.hat - tq * sqrt(v.err)
print(y.1)
#upper limit
y.2 <- y.hat + tq * sqrt(v.err)
print(y.2)
```

We obtain the following limits.



```
> print(y.hat)
[1] 9.116841
> #leverage - it is a scalar value
> lev <- (x%*%detailed.model$cov.unscaled%*%t(x))[1,1]
> #variance of the error of prediction
> v.err <- (detailed.model$sigma)^2 * (1 + lev)
> #Student quantile for 0.95
> tq <- qt(0.95,24)
> #lower limit
> y.1 <- y.hat - tq * sqrt(v.err)
> print(y.1)
[1] 7.792365
> #upper limit
> y.2 <- y.hat + tq * sqrt(v.err)
> print(y.2)
[1] 10.44132
>
> |
```

5 Conclusion

In this tutorial, we show the implementation of generalized testing of the coefficients using Excel from the results provided by Tanagra. In my point of view, the advantage of the spreadsheet is primarily educational. This tool is relatively easy to use; everyone knows how to use a spreadsheet. We show thereafter that the same calculations can be realized under the R software. We get the same results! This is the most important. Whatever the tool used, the essential is in understanding the techniques.