1 Objectif

Lecture des résultats de la régression linéaire multiple. Tests sur les coefficients.

La régression linéaire multiple est une technique de modélisation statistique. Elle vise à prédire et expliquer les valeurs prises par une variable endogène quantitative Y à partir de *p* variables exogènes X1, …, Xp, quantitatives ou qualitatives rendues binaires par recodage.

Nous disposons d'un échantillon de taille *n*. Formellement, le modèle s'écrit de la manière suivante

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

Le terme d'erreur ε_i résume les insuffisances du modèle (problème de spécifications, absence de certaines variables, etc.). Toute la partie inférentielle (propriétés des estimateurs, tests d'hypothèses, estimation par intervalle) repose sur les hypothèses édictées à son propos.

L'objectif de la régression est d'estimer les paramètres $(a_1, a_2, \ldots, a_p, a_{p+1})$ à partir d'un échantillon, de manière à ce que les valeurs prédites $(\hat{y}_i)$ par le modèle soient le plus proche possible des valeurs observées de l'endogène (y_i). Cette idée est résumée par un indicateur synthétique appelé « **somme des carrés des résidus** »

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

Le résidu $\hat{\varepsilon}_i = \hat{y}_i - y_i$ est l'erreur observée (le résidu) pour l'observation n°i.

Comme dans tout processus prédictif, l'évaluation de la régression intègre plusieurs niveaux : (1) Quelle est la qualité globale du modèle ? (2) Est-ce qu'il est globalement intéressant c.-à-d. est-ce que les X emmènent significativement de l'information dans l'explication de Y ? (3) Quel est l'impact individuel des exogènes sur l'endogène ? (4) Peut-on mettre en place des procédures spécifiques pour évaluer le rôle simultané de plusieurs variables exogènes ?

Dans ce tutoriel, à travers un exemple de prédiction de la consommation des véhicules à partir de leur poids, de leur cylindrée et de leur puissance, nous décrirons les sorties de TANAGRA en leur associant les formules utilisées. Nous mettrons en avant le rôle de la matrice $(X'X)^{-1}$ fournie depuis la version **1.4.38**. Elle est importante car elle tient une place centrale dans les tests généralisés sur les coefficients. Nous en accomplirons quelques uns manuellement avec le tableur Excel.

Dans un deuxième temps, nous réaliserons la régression à l'aide du logiciel R. Nous mettrons en parallèle ses résultats avec ceux de TANAGRA. Nous identifierons les objets qui fournissent les informations nécessaires aux différents post-traitements, notamment les tests généralisés. Nous effectuerons alors les calculs réalisés précédemment dans Excel directement dans R.

2 Données

Notre fichier comporte n = 28 observations. Il s'agit d'expliquer la consommation (« Y - consumption », en l / 100 km) à partir de la cylindrée (« X1 - eng.size », en cm^3), la puissance (« X2 - horsepower », en ch) et le poids (« X3 - weight », en 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

En passant à une notation matricielle, le vecteur de l'endogène Y est de dimension (n = 28, 1)

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

La matrice des exogènes, avec la constante dans la dernière colonne, est de dimension (n, p+1 =4)

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

L'estimateur des moindres carrés ordinaires (MCO) de la régression s'écrit

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

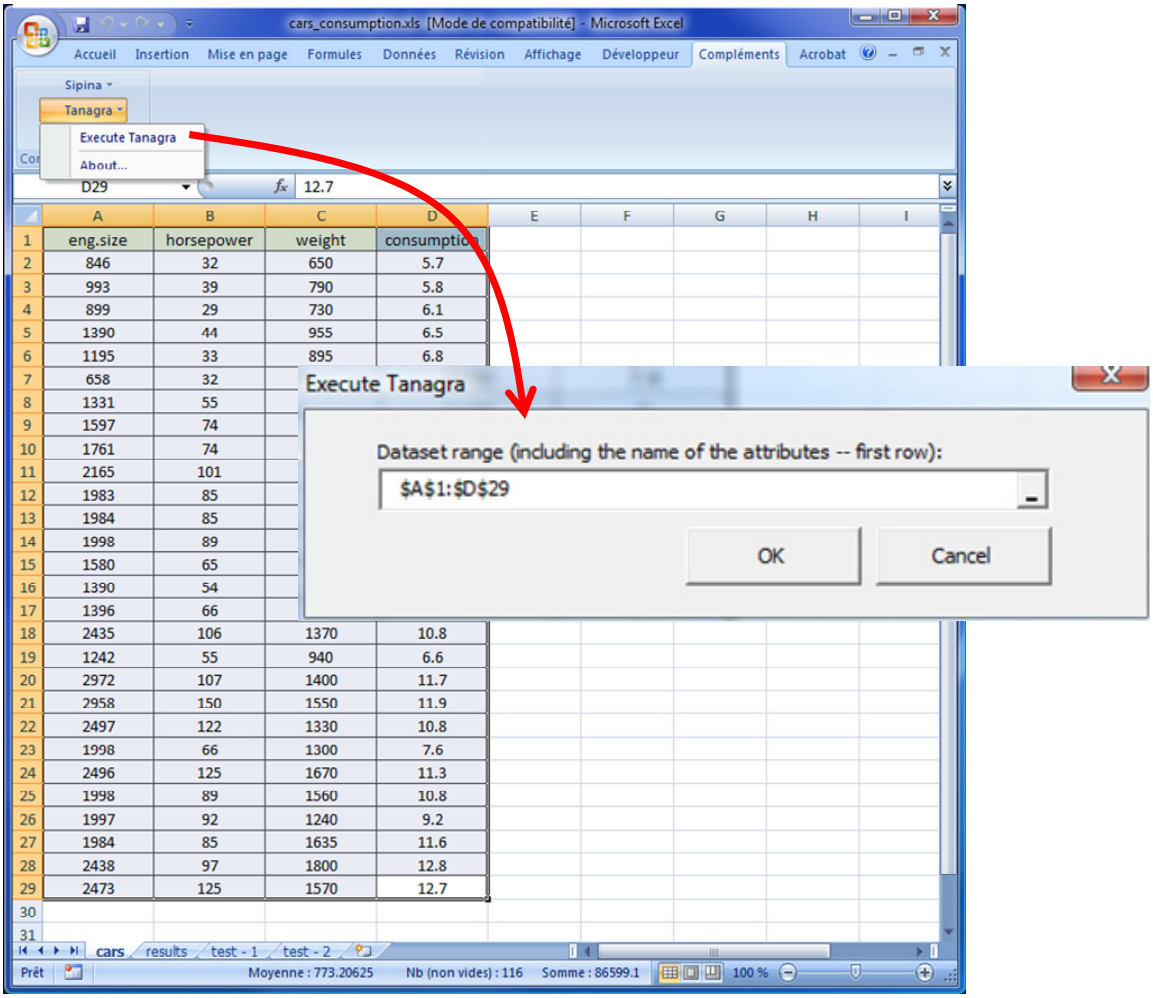
Où X' est la transposée de X, $(X'X)^{-1}$ est l'inverse de (X'X). Les valeurs prédites par le modèle s'obtiennent avec

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

3 Régression linéaire multiple avec Tanagra

3.1 Importation des données

Nous chargeons le fichier « cars_consumption.xls » dans Excel. Nous sélectionnons la plage de valeurs, y compris l'en-tête de colonne représentant les noms de variables, puis nous actionnons le menu TANAGRA / EXECUTE TANAGRA intégré dans le tableur[1].

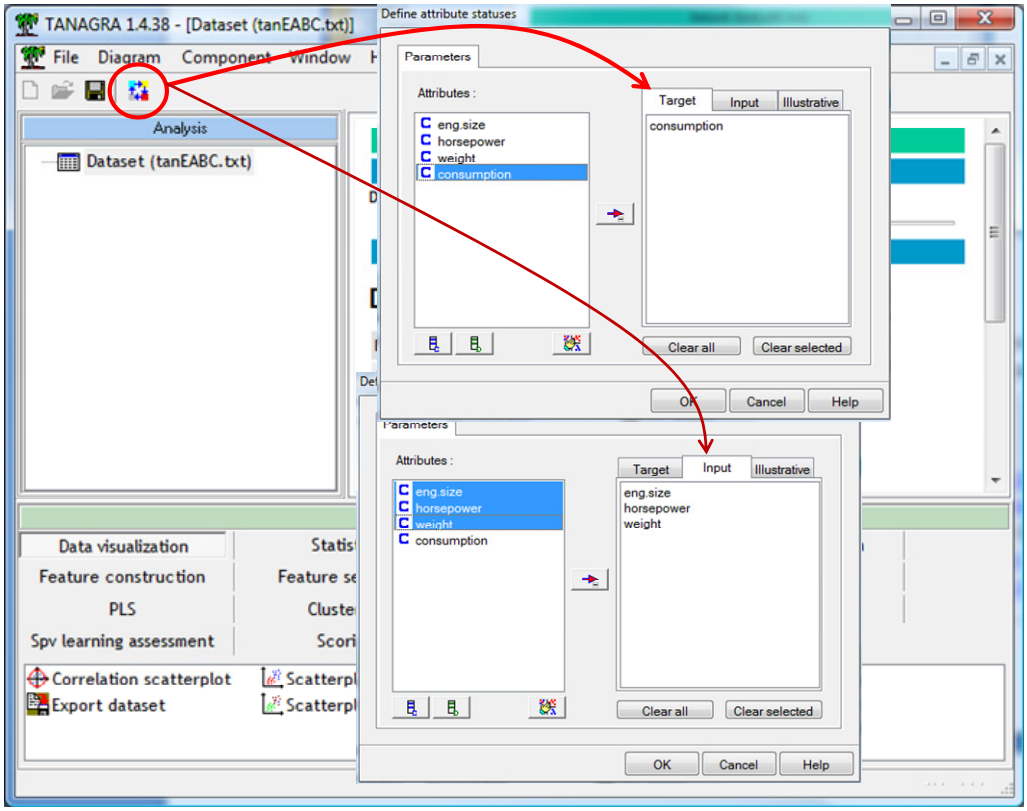


TANAGRA est automatiquement démarré et les données chargées. Nous vérifions que nous disposons de 4 variables et n = 28 observations.

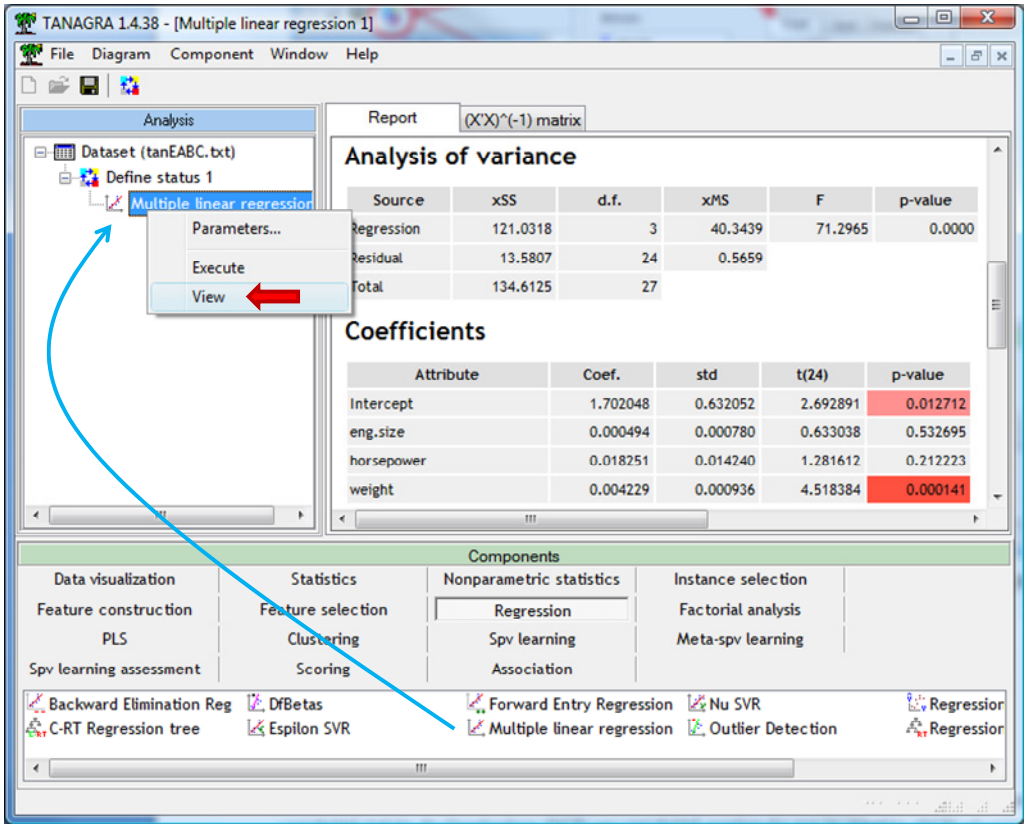
3.2 Régression linéaire multiple

Nous devons spécifier dans un premier temps la variable endogène (Y) et les exogènes (X1, X2, X3). Nous utilisons le composant DEFINE STATUS pour cela. Nous plaçons CONSUMPTION en TARGET, les autres en INPUT.

[1] Voir http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-macro.html pour Excel 2003 et versions antérieures ; et http://tutoriels-data-mining.blogspot.com/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html pour Excel 2007 et 2010. Un dispositif a aussi été prévu pour Open Office Calc, sous Windows http://tutoriels-data-mining.blogspot.com/2008/03/connexion-open-office-calc.html et sous Linux http://tutoriels-data-mining.blogspot.com/2009/04/connexion-open-office-calc-sous-linux.html



Dans un deuxième temps, nous insérons le composant de régression MULTIPLE LINEAR REGRESSION (onglet REGRESSION) dans le diagramme.



Nous actionnons le menu contextuel VIEW pour accéder aux résultats. Voyons en le détail.

3.3 Evaluation globale de la régression

3.3.1 Tableau d'analyse de variance et coefficient de détermination

Premier outil pour l'évaluation de la régression, le tableau d'analyse de variance décompose la variabilité totale de l'endogène (SCT) en variabilité expliquée par le modèle (SCE) et variabilité résiduelle (SCR), non prise en compte par le modèle. Le ratio (SCE / SCT) indique la part de variabilité expliquée, c'est le coefficient de détermination R^2, il varie entre 0 et 1. Lorsque le R^2 est égal à 1, le modèle permet de prévoir avec exactitude les valeurs prises par l'endogène Y c.-à-d. SCR = 0.

TANAGRA résume tout cela dans le tableau « ANALYSIS OF VARIANCE ».

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			

Nous en déduisons le R^2

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

Que nous retrouvons dans le tableau « GLOBAL RESULTS ».

Global results

Endogenous attribute	**consumption**
Examples	28
R^2	0.899113
Adjusted-R^2	0.886502
Sigma error	0.752238
F-Test (3,24)	71.2965 (0.000000)

TANAGRA fournit également le R^2 ajusté (Ajusted-R^2) qui est tout simplement le R^2 corrigé par les degrés de liberté. Nous privilégions cet indicateur lorsqu'il s'agit de comparer des modèles avec des complexités différentes (pour évaluer – rapidement – par exemple l'impact de variables supplémentaires dans le modèle).

$$\overline{R}^2 = 1 - \frac{SCR/n-p-1}{SCT/n-1} = 1 - \frac{13.5807/24}{134.6125/27} = 0.886502$$

$$= 1 - \frac{n-1}{n-p-1}\left(1-R^2\right) = 1 - \frac{27}{24}\left(1-0.899113\right)$$

3.3.2 Test de significativité globale

La régression est-elle globalement statistiquement pertinente ? Pour répondre à cette question, nous introduisons le test de significativité globale. Il s'agit de vérifier si aucune des exogènes, prises simultanément, n'apporte d'information sur l'endogène. Formellement, il s'écrit :

$$\begin{cases} H0 : a_1 = a_2 = \cdots = a_p = 0 \\ H1 : \text{un des coefficients au moins est non nul} \end{cases}$$

Il repose sur la statistique F,

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

Nous retrouvons les valeurs idoines dans les deux tableaux mis en avant dans la section précédente. Sous l'hypothèse nulle Ho, elle suit une loi de Fisher à (p = 3, n-p-1 = 24) degrés de liberté. Tanagra, à l'instar de la grande majorité des logiciels de statistique, fournit directement la probabilité critique (p-value)

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

Lorsque la (p-value) est inférieure au seuil de significativité α (généralement 5%), on décide le rejet de l'hypothèse nulle c.-à-d. on conclut à la significativité globale du modèle.

Dans notre exemple, p-value ≈ 0 (très largement inférieur à 5%), le modèle est très significatif.

3.4 Evaluation des coefficients

L'étape suivante consiste à évaluer l'apport individuel des variables dans l'explication de Y. Nous utilisons pour ce faire le test de significativité individuelle des coefficients. Pour la variable Xj, il s'écrit :

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

La statistique de test est obtenue avec

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

$\hat{\sigma}_{\hat{a}_j}$ est l'écart type estimé du coefficient estimé. Son carré, la variance estimée, est lu sur la diagonale principale de la matrice de variance covariance des coefficients estimés, soit

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

$\hat{\sigma}_\varepsilon^2$ est l'estimation de la variance de l'erreur du modèle, elle correspond au rapport entre la somme des carrés des résidus et les degrés de liberté. Nous observons l'écart-type estimé dans le tableau « GLOBAL RESULTS » (Sigma error).

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

Dans le tableau « COEFFICIENTS », Tanagra fournit directement $\hat{\sigma}_{\hat{a}_j}$ et t_j. Sous l'hypothèse nulle Ho, la statistique suit une loi de Student à (n – p – 1 = 24) degrés de libertés. Dans la dernière colonne, nous obtenons la p-value du test. La variable est significative si elle est inférieure au risque α = 5%. Dans notre cas, la variable WEIGHT semble être la seule significative avec une p-value égale à 0.000141.

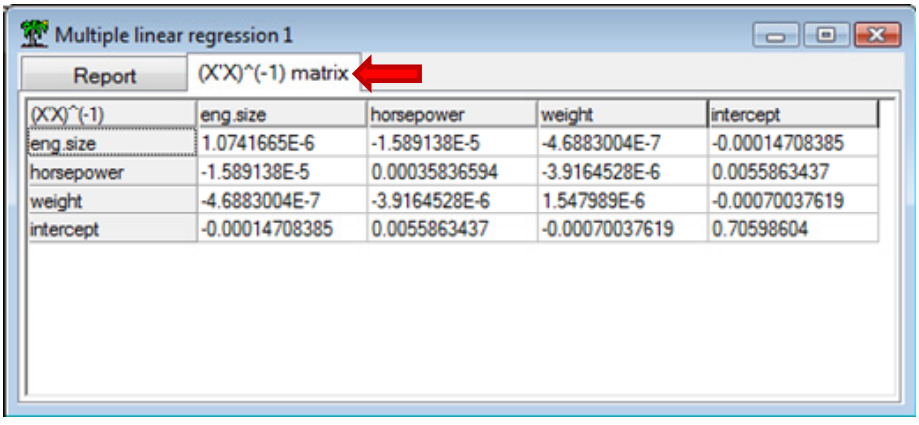
Coefficients

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 Matrice de variance covariance des coefficients

Nous pouvons reconstituer la matrice de variance covariance complète à partir des résultats fournis par Tanagra. Nous pouvons ainsi retrouver les écarts-type du tableau « COEFFICIENTS ». Mais cela dépasse la simple anecdote. Dans certains tests que nous présenterons plus loin, nous avons besoin des covariances entre les coefficients estimés.

Dans le second onglet de la fenêtre de visualisation, Tanagra propose la matrice $(X'X)^{-1}$.



(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

En la pré-multipliant par la variance de l'erreur, nous retrouvons la matrice $\hat{\Omega}_{\hat{a}} = \hat{\sigma}_\varepsilon^2 \left(X'X\right)^{-1}$

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

Nous avons les variances des estimateurs sur la diagonale principale. Pour la variable HORSEPOWER (X2) par exemple, nous avons

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

Ce qui correspond à l'écart-type fourni par Tanagra dans le tableau COEFFICIENTS.

3.6 Test de conformité d'un groupe de coefficients à un standard

La matrice $(X'X)^{-1}$ est importante car elle participe à toute une série de tests, notamment à travers la matrice de variance covariance des coefficients. Nous abordons dans cette section le test de conformité d'un groupe de q coefficients à un standard. Formellement, nous confrontons les deux hypothèses :

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

La statistique de test F suit, sous Ho, une loi de Fisher à (q, n - p – 1) degrés de liberté.

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

La région critique correspond aux grandes valeurs de F.

Explicitons cela sur un exemple. Nous souhaitons tester la nullité simultanée des coefficients de ENG.SIZE et HORSEPOWER. Rappelons-nous, les deux n'étaient pas significatifs à 5% lorsqu'ils étaient testés individuellement (section 3.4). Mais ces tests ne tiennent pas compte de la possible interaction entre les variables, la covariance entre les coefficients n'est pas nulle, nous lisons dans la matrice $\hat{\Omega}_{\hat{a}}$ la valeur $\hat{\text{cov}}(\hat{a}_{eng.size,}\hat{a}_{horsepower}) = -8.9923 \times 10^{-6}$. Le test simultané, lui, va tenir compte de cette information. La sous-matrice de variance covariance des coefficients à tester est restreinte à la partie colorisée dans la matrice globale

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

Soit

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

Nous l'inversons

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

Nous pouvons dès lors obtenir la valeur de la statistique de test :

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

La p-value du test est 0.0166, inférieure au risque 5%. A la différence des tests individuels (les deux coefficients ne semblaient pas individuellement significatifs), nous rejetons la nullité simultanée des deux coefficients. La prise en compte du lien entre les coefficients via leur covariance nous amène à prendre une décision différente. Cela laisse à penser que ces deux variables sont très fortement corrélées et, du fait de leur redondance, elles se « gênent » mutuellement dans la régression.

Tous ces calculs ont été réalisés à l'aide d'une feuille de calcul Excel que nous reproduisons ici.

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

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

p-value	0.0166

3.7 Test de combinaisons linéaires des coefficients

Pour les tests de comparaisons à un standard. Le test ci-dessus peut s'écrire d'une manière plus générique encore à travers les tests de combinaisons linéaires de variables. Nous testons :

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

La matrice R est de taille (q, p+1), où q est le nombre de contraintes que l'on souhaite mettre en place, (p+1) est le nombre de paramètres du modèle. Le vecteur r est de taille (q, 1).

Tout l'enjeu est d'écrire correctement la matrice R et le vecteur r. Pour le test de la section précédente, avec

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

Nous aurons :

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

Ainsi

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

Comparaisons de coefficients. Mais les tests de combinaisons linéaires ont une portée plus large. Ils permettent la mise en place de comparaisons de coefficients. Prenons un exemple simple pour illustrer cela.

La puissance spécifique est une notion clé pour qualifier les moteurs des automobiles. Il s'agit du nombre de chevaux développés par unité de cylindrée (le plus souvent en litres, soit 1000 cm^3). On sait que les véhicules sportifs font plus de 100 ch/L. Mais, plus raisonnablement, le rapport usuel pour les loutres sur roues serait plutôt autour de 40 ch/L (à peu près la moyenne constatée sur notre fichier). Voyons si, dans la régression, les coefficients conservent le même rapport dans leur impact sur la consommation c.-à-d. nous souhaitons tester :

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

Ecrivons la matrice R et le vecteur r correspondants. Nous avons

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

La statistique de test F suit une loi de Fisher à (q, n – p – 1) degrés de liberté sous l'hypothèse nulle.

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

q = 1 dans notre exemple ; n – p – 1 = 24 est le degré de liberté de la régression.

Détailler les opérations serait fastidieux, nous nous contentons de présenter la feuille de calcul Excel que nous avons utilisée.

(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

Au final, nous avons F = 0.0337, avec une p-value = 0.8585. Les données ne contredisent pas l'hypothèse nulle. Pour nous cela veut dire que les deux variables ont le même impact sur la consommation.

3.8 Prédiction ponctuelle et intervalle de prédiction

Un autre usage de la matrice $(X'X)^{-1}$ est la construction des intervalles de prédiction.

Pour un nouvel individu (i*) n'appartenant pas à l'échantillon ayant servi à la construction du modèle, nous pouvons prévoir la valeur de l'endogène en fournissant celles des exogènes. On parle

de prédiction ponctuelle. Concrètement, il s'agit de réaliser un produit scalaire entre le vecteur de description de l'individu et le vecteur des coefficients estimés.

Pour l'individu décrit par les valeurs (eng.size = 1984 ; horsepower = 85 ; weight = 1155), le vecteur de description serait la suivante :

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

On notera l'adjonction de la constante à la 4ème cellule pour assurer la correspondance avec la taille du vecteur des paramètres estimés où, justement, la constante est en dernière position. Nous obtenons en appliquant les coefficients de la régression sur la description de l'individu :

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

Une prédiction ponctuelle est toujours intéressante pour se donner un ordre d'idée sur le positionnement d'un individu par rapport aux valeurs de l'endogène. Mais nous savons très bien que la valeur fournie par le modèle est entachée d'incertitude, tout simplement parce qu'aux coefficients estimés à partir d'un échantillon sont associés une certaine variabilité.

Il est donc plus intéressant de fournir un intervalle de prédiction, à laquelle nous attribuons un niveau de confiance. Ainsi, nous pourrons dire : « l'intervalle a tant de chances de contenir la vraie valeur de Y pour l'individu étudié ».

Deux informations concernant l'erreur de prévision ($\hat{\varepsilon}_{i*} = \hat{y}_{i*} - y_{i*}$) sont nécessaires pour mener à bien l'opération : il nous faut connaître sa variance et sa loi de distribution.

Sa variance estimée s'écrit

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

Grosso modo, la variance sera d'autant plus petite (la prédiction sera précise) que la régression est de bonne qualité (la variance de l'erreur du modèle $\hat{\sigma}^2_{\varepsilon}$, qui dépend directement de la SCR, est faible) et que l'observation est proche du barycentre du nuage de points dans l'espace des exogènes (le levier $h_{i*} = x_{i*} \left(X' X \right)^{-1} x_{i*}'$ est faible).

L'erreur normalisée est distribuée suivant une loi de Student à (n – p – 1) degrés de liberté

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

Ainsi, l'intervalle de confiance au niveau (1 - α) est définie comme suit

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

Concernant nos données, nous avons :

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

Pour un niveau de confiance de 90%, nous lisons dans la table de la loi de Student $t_{0.95} = 1.7109$, nous obtenons alors les bornes de l'intervalle de prévision :

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

Tous ces calculs sont résumés dans la feuille Excel ci-jointe.

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 Régression linéaire multiple avec R

4.1 Régression avec lm(.)

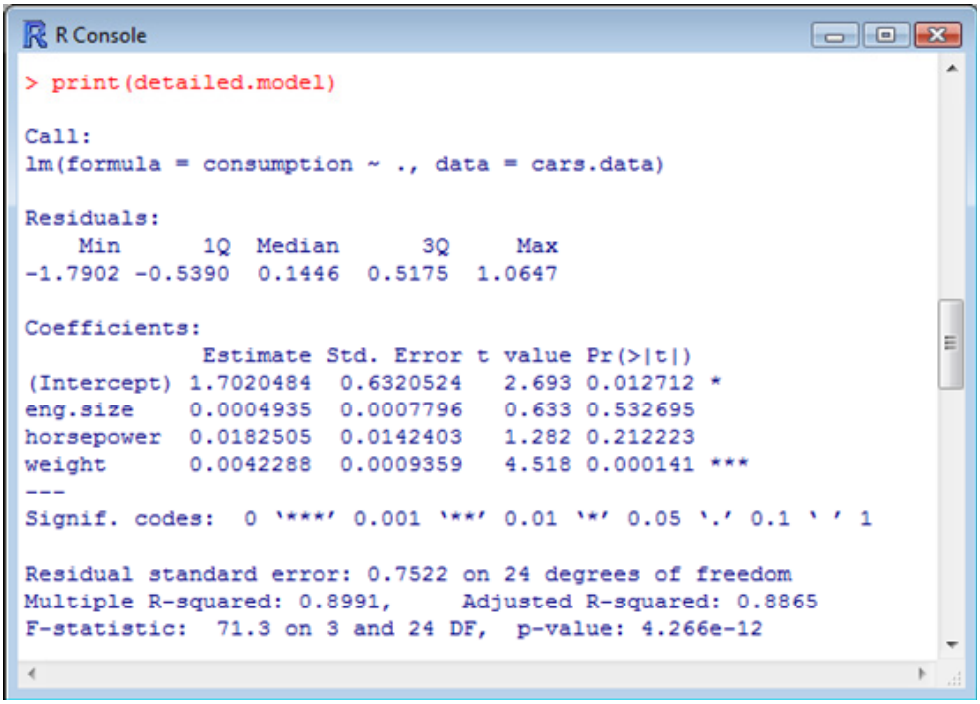
Nous utilisons la méthode ***lm(.)*** sous R. Nous exploitons plus particulièrement l'objet fourni par la procédure *summary(.)* associée. Cette dernière affiche en standard les principales informations attendues de la régression linéaire multiple.

Voici le code source R :

```
rm (list=ls())
#modifier le répertoire courant
setwd("…")
#charger les données
library(xlsReadWrite)
```

```
cars.data <- read.xls(file="cars_consumption.xls",colNames=T,sheet=1)
#lancer la régression
cars.model <-lm(consumption ~ ., data = cars.data)
print(cars.model)
#objet résumé pour obtenir des résultats détaillés
detailed.model <- summary(cars.model)
print(detailed.model)
```

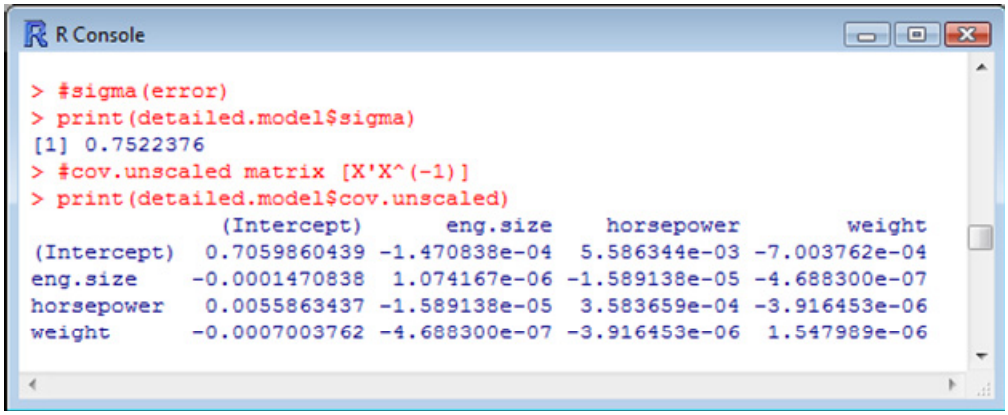
Nous obtenons les résultats en tout points identiques à ceux de Tanagra.



L'objet **summary.lm** est très intéressant car il fournit toute une série d'informations nécessaires aux calculs ultérieurs. Nous pouvons par exemple obtenir l'estimation de l'écart-type de l'erreur, ou encore la fameuse matrice $(X'X)^{-1}$.

```
#écart-type estimé de l'erreur
print(detailed.model$sigma)
#cov.unscaled correspond à la matrice [X'X^(-1)]
print(detailed.model$cov.unscaled)
```

Nous avons ainsi :



R place la constante (intercept) en première position. Qu'importe. Les valeurs fournies sont cohérentes avec celles de Tanagra.

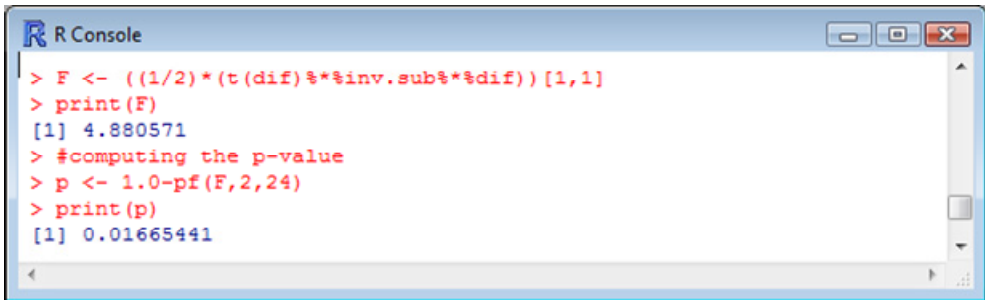
4.2 Test de conformité simultanée pour plusieurs coefficients

L'énorme avantage de R est que nous pouvons quasiment tout calculer avec son langage de programmation. Bien sûr, il faut faire l'apprentissage des commandes. Mais le bénéfice qui en est retiré par la suite justifie pleinement l'investissement initial.

Pour le test de conformité ci-dessus (section 3.6), nous écrivons le programme suivant :

```
#récupérer les coefficients estimés pour eng.size et horsepower
a.test <- matrix(detailed.model$coefficients[2:3,1],nrow=2,ncol=1)
#le vecteur des valeurs de référence
ref.test <- matrix(c(0,0),nrow=2,ncol=1)
#différence entre coefficients estimés et références
dif <- a.test - ref.test
#matrice de variance covariance complète des estimateurs
vcv.mat <- detailed.model$sigma^2 * detailed.model$cov.unscaled
#sous-matrice de var.covar pour eng.size et horsepower
sub.vcv.mat <- vcv.mat[2:3,2:3]
print(sub.vcv.mat)
#inversion de la sous-matrice
inv.sub <- solve(sub.vcv.mat)
print(inv.sub)
#calcul de la statistique de test
#c'est un scalaire
F <- ((1/2)*(t(dif)%*%inv.sub%*%dif))[1,1]
print(F)
#calculer la p-value du test
p <- 1.0-pf(F,2,24)
print(p)
```

A la sortie, nous obtenons F = 4.880571 avec un p-value de 0.01665441 ; très similaire avec ce que l'on obtenu via Excel. La très faible différence (à partir de la 3ème décimale pour F) provient de la perte de précision lors de la recopie des paramètres estimés dans le tableur.



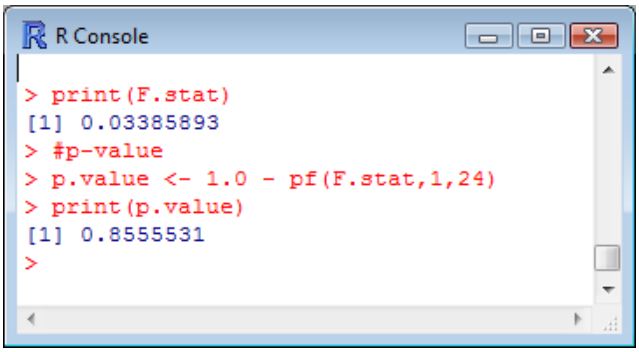
4.3 Test de combinaisons linéaires de coefficients

De la même manière, concernant le test de combinaisons linéaires de coefficients (section 3.7), nous pouvons réaliser tous les calculs directement dans R. Le code est le suivant :

```
#vecteur des coefficients estimés
a.hat <- matrix(detailed.model$coefficients[,1],nrow=4,ncol=1)
#la matrice R !!! attention la constante est en première position
R <- matrix(c(0,1000,-40,0),nrow=1,ncol=4)
#le vecteur r
```

```
r <- matrix(c(0),nrow=1,ncol=1)
#R(X'X)^(-1)R'
B <- R%*%detailed.model$cov.unscaled%*%t(R)
#B^(-1)
inv.B <- solve(B)
#numerateur de la statistique de test
F_num <- (1/1)*(t(R%*%a.hat-r)%*%inv.B%*%(R%*%a.hat-r))[1,1]
#dénominateur de la statistique de test
F_denom <- detailed.model$sigma^2
#statistique de test F
F.stat <- F_num/F_denom
print(F.stat)
#probabilité critique du test
p.value <- 1.0 - pf(F.stat,1,24)
print(p.value)
```

Attention, la constante étant placée en première position, il faut en tenir lors de l'élaboration de la matrice R. Le reste est une simple transposition des opérations que nous avons réalisées dans Excel. Nous obtenons les résultats suivants, identiques à ceux de la section 3.7.



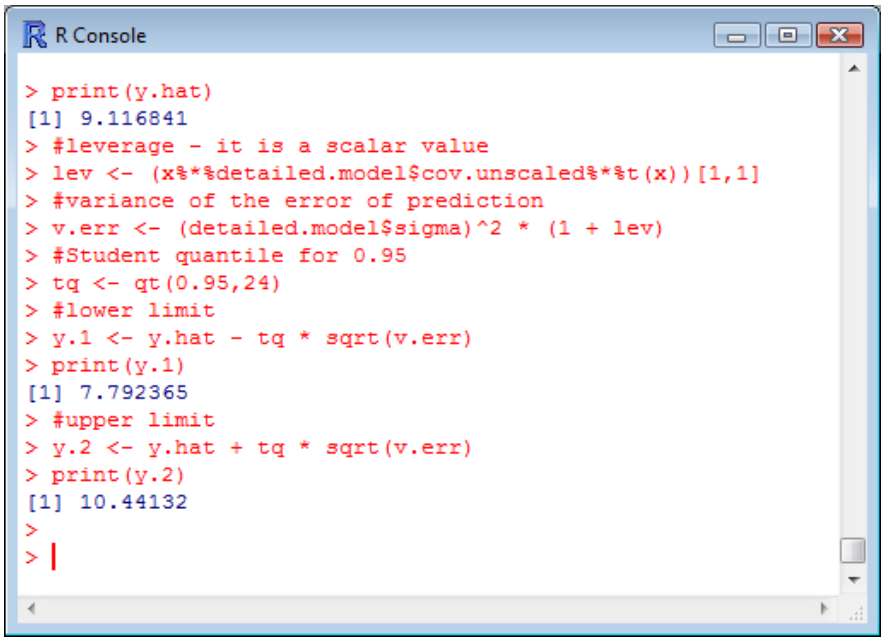
4.4 Prédiction et intervalle de prédiction

Le code équivalent aux calculs réalisés sous Excel est le suivant.

```
#description - valeurs des exogènes (y compris la constante)
x <- matrix(c(1,1984,85,1155),nrow=1,ncol=4)
#prédiction ponctuelle - c'est un scalaire
y.hat <- (x%*%a.hat)[1,1]
print(y.hat)
#le levier - c'est un scalaire
lev <- (x%*%detailed.model$cov.unscaled%*%t(x))[1,1]
#variance de l'erreur de prédiction
v.err <- (detailed.model$sigma)^2 * (1 + lev)
#quantile d'ordre 0.95 de la loi de Student
tq <- qt(0.95,24)
#borne basse
y.1 <- y.hat - tq * sqrt(v.err)
print(y.1)
#borne haute
y.2 <- y.hat + tq * sqrt(v.err)
print(y.2)
```

De nouveau, en accord avec les conventions adoptées par le logiciel R, la constante est placée en première position dans le vecteur x_{i*}

A la sortie, nous obtenons les mêmes bornes que sous Excel.



5 Conclusion

Dans ce didacticiel, nous montrons la mise en œuvre des tests généralisés des coefficients à l'aide d'Excel à partir des résultats fournis par Tanagra. A mon sens, l'avantage du tableur est essentiellement pédagogique. C'est un outil relativement accessible, tout le monde sait s'en servir (peut être un peu moins en ce qui concerne les opérations matricielles…). Nous montrons par la suite que les mêmes opérations peuvent être menées sous R. Tout aussi facilement si l'on connaît un peu le langage de commande. Au final, nous obtenons les mêmes résultats ! C'est le plus important. Qu'importe l'outil, l'essentiel est dans la compréhension des techniques.