1 Objectif

Réalisation des différentes étapes de la régression linéaire multiple sous Excel.

Il y a deux écueils à éviter lors des travaux dirigés (TD) sur machine. Le premier est de ne pas assez guider les étudiants. On entend très vite fuser du fond de la salle la question fatidique « Qu'est-ce qu'il faut faire là ? ». Ce n'est jamais très bon signe. Le second est de trop les guider en donnant des solutions toutes faites. Cela peut arriver lorsqu'on utilise des outils tel que R ou Python où programmer est nécessaire. On est tenté de fournir des indications sous forme de code source. Les étudiants ont alors le sentiment de recopier bêtement sans en comprendre réellement la teneur, zappant totalement les notions clés des méthodes étudiées. Là également, ils se désintéressent rapidement de la séance et commencent à tripoter leurs smartphones.

La vérité est entre ces deux extrêmes. Mais elle n'est pas toujours facile à trouver. Pour ma part, au moins dans la première partie de mes cours d'initiation, que ce soit de statistique ou de data mining, je travaille énormément avec le tableur Excel. Entendons-nous bien, il ne s'agit pas de substituer Excel aux logiciels de statistiques qui sont autrement plus puissants et précis. Ses limites en la matière sont connues (Keeling & Pavur, 2011 ; Yalta, 2008). Quoique les fonctions statistiques d'Excel aient favorablement évolué ces dernières années semble-t-il (Mélard, 2017). Il s'agit avant tout d'une démarche pédagogique qui permet de détailler les étapes de calcul. L'idéal, c'est quand même un peu l'objectif des travaux dirigés, serait que les exercices sur machine permettent aux étudiants de mieux appréhender les thèmes et formules présentés durant le cours magistral.

Dans ce tutoriel, nous reprenons à partir d'un exemple traité sous Excel les principaux concepts présentés dans les documents accessibles sur ma page de cours d'Econométrie (http://eric.univ-lyon2.fr/~ricco/cours/cours_econometrie.html). Deux ouvrages en particulier seront mis à contribution (voir Références, section 7). Les principales formules seront explicitées pour chaque calcul. Elles seront mises en relation avec les opérations sous Excel. Puisque nous traitons de la régression linéaire multiple, nous ferons un usage intensif des fonctions matricielles du tableur.

2 Données

Nous traitons les données « cigarettes » (« cigarettes_tutorial_regression.xlsx »). On souhaite expliquer la nocivité des cigarettes (teneur en monoxyde de carbone – CO) (y) à partir de leur composition : TAR (goudron) (x_1), NICOTINE (x_2) et WEIGHT (poids) (x_3), soit p = 3 variables explicatives. Nous disposons de n = 24 observations. Les observations sont visibles dans la feuille « cigarettes » du fichier Excel (Figure 1).

Cigarette	TAR (mg)	NICOTINE (m	WEIGHT (g)	CO (mg)
Alpine	14.1	0.86	0.9853	13.6
Benson&Hedges	16	1.06	1.0938	16.6
CamelLights	8	0.67	0.928	10.2
Carlton	4.1	0.4	0.9462	5.4
Chesterfield	15	1.04	0.8885	15
GoldenLights	8.8	0.76	1.0267	9
Kent	12.4	0.95	0.9225	12.3
Kool	16.6	1.12	0.9372	16.3
L&M	14.9	1.02	0.8858	15.4
LarkLights	13.7	1.01	0.9643	13
Marlboro	15.1	0.9	0.9316	14.4
Merit	7.8	0.57	0.9705	10
MultiFilter	11.4	0.78	1.124	10.2
NewportLights	9	0.74	0.8517	9.5
Now	1	0.13	0.7851	1.5
OldGold	17	1.26	0.9186	18.5
PallMallLight	12.8	1.08	1.0395	12.6
Raleigh	15.8	0.96	0.9573	17.5
SalemUltra	4.5	0.42	0.9106	4.9
Tareyton	14.5	1.01	1.007	15.9
TrueLight	7.3	0.61	0.9806	8.5
ViceroyRichLight	8.6	0.69	0.9693	10.6
VirginiaSlims	15.2	1.02	0.9496	13.9
WinstonLights	12	0.82	1.1184	14.9

Figure 1 - Données "cigarettes" (Feuille "cigarettes")

3 Régression par calcul matriciel

L'équation de régression s'écrit :

$$y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + a_3 x_{i3} + \varepsilon_i \; ; \; i = 1, \ldots, n$$

Où a = (a_0, a_1, a_2, a_3) est le vecteur des paramètres à estimer à partir des données disponibles.

En passant à l'écriture matricielle, nous avons :

$$y = Xa + \varepsilon$$

Où X est la matrice des variables explicatives avec, dans la première colonne, la valeur 1 pour matérialiser la constante de la régression a_0. X est de dimension (n, p +1), soit (24, 4) pour nos données. Notre *dataset*, sous sa forme matricielle, a été copiée dans la feuille « matrices ». La matrice X est située aux coordonnées (A2:D25), le vecteur y en (E2:E25) (Figure 2).

	A	B	C	D	E
1		X			Y
2	1	14.1	0.86	0.9853	13.6
3	1	16	1.06	1.0938	16.6
4	1	8	0.67	0.928	10.2
5	1	4.1	0.4	0.9462	5.4
6	1	15	1.04	0.8885	15
7	1	8.8	0.76	1.0267	9
8	1	12.4	0.95	0.9225	12.3
9	1	16.6	1.12	0.9372	16.3
10	1	14.9	1.02	0.8858	15.4
11	1	13.7	1.01	0.9643	13
12	1	15.1	0.9	0.9316	14.4
13	1	7.8	0.57	0.9705	10
14	1	11.4	0.78	1.124	10.2
15	1	9	0.74	0.8517	9.5
16	1	1	0.13	0.7851	1.5
17	1	17	1.26	0.9186	18.5
18	1	12.8	1.08	1.0395	12.6
19	1	15.8	0.96	0.9573	17.5
20	1	4.5	0.42	0.9106	4.9
21	1	14.5	1.01	1.007	15.9
22	1	7.3	0.61	0.9806	8.5
23	1	8.6	0.69	0.9693	10.6
24	1	15.2	1.02	0.9496	13.9
25	1	12	0.82	1.1184	14.9

Figure 2 - Données "cigarettes", forme matricielle (Feuille "matrices")

3.1 Estimation des paramètres de la régression

L'estimateur des moindres carrés ordinaires s'écrit :

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

Où X' est la transposée de X.

Nous procédons en plusieurs étapes sous Excel :

- Tout d'abord nous calculons la matrice (X'X) de dimension (4, 4) en (I3:L6) avec l'instruction {=PRODUITMAT(TRANSPOSE(A2:D25);A2:D25)}. Nous avons introduit une fonction matricielle qui complète simultanément plusieurs cellules (de I3 à L6). Elle doit être validée avec la combinaison de touches CTRL + SHIFT + ENTREE. Excel ajoute automatiquement les accolades { }.
- En (I9:L12), nous inversons la matrice avec l'instruction {=INVERSEMAT(G3:J6)} pour obtenir $(X'X)^{-1}$, toujours de dimension (4, 4).
- Nous calculons (X'y) en (I16:I19) avec {=PRODUITMAT(TRANSPOSE(A2:D25);E2:E25)}.
- Il ne reste plus qu'à former â en (K16:K19) avec {=PRODUITMAT(I9:L12;I16:I19)}.

Voici la feuille de calcul à ce stade (Figure 3).

	A	B	C	D	E
1		X			Y
2	1	14.1	0.86	0.9853	13.6
3	1	16	1.06	1.0938	16.6
4	1	8	0.67	0.928	10.2
5	1	4.1	0.4	0.9462	5.4
6	1	15	1.04	0.8885	15
7	1	8.8	0.76	1.0267	9
8	1	12.4	0.95	0.9225	12.3
9	1	16.6	1.12	0.9372	16.3
10	1	14.9	1.02	0.8858	15.4
11	1	13.7	1.01	0.9643	13
12	1	15.1	0.9	0.9316	14.4
13	1	7.8	0.57	0.9705	10
14	1	11.4	0.78	1.124	10.2
15	1	9	0.74	0.8517	9.5
16	1	1	0.13	0.7851	1.5
17	1	17	1.26	0.9186	18.5
18	1	12.8	1.08	1.0395	12.6
19	1	15.8	0.96	0.9573	17.5
20	1	4.5	0.42	0.9106	4.9
21	1	14.5	1.01	1.007	15.9
22	1	7.3	0.61	0.9806	8.5
23	1	8.6	0.69	0.9693	10.6
24	1	15.2	1.02	0.9496	13.9
25	1	12	0.82	1.1184	14.9

(X'X)			
24	275.60	19.88	23.09
275.60	3613.16	254.18	267.46
19.88	254.18	18.09	19.27
23.09	267.46	19.27	22.36

(X'X)^-1			
6.563	0.063	-0.939	-6.720
0.063	0.028	-0.452	-0.015
-0.939	-0.452	7.863	-0.399
-6.720	-0.015	-0.399	7.510

X'y
289.70
3742.85
264.08
281.15

a^
-0.5517
0.8876
0.5185
2.0793

Figure 3 - Estimation matricielle des paramètres de la régression (Feuille "matrices")

Les coefficients estimés sont $\hat{a}_0 = -0.5517$, $\hat{a}_1 = 0.8876$, $\hat{a}_2 = 0.5185$ et $\hat{a}_3 = 2.0793$.

3.2 Prédiction des valeurs de la variable cible

Avec les coefficients estimés de la régression, nous pouvons calculer les valeurs estimées de la variable dépendante :

$$\hat{y} = X\hat{a}$$

En (F2:F25), nous insérons la commande {=PRODUITMAT(A2:D25;K16:K19)} (Figure 4).

Remarque : Nous aurions pu également réaliser la prédiction pour la première observation en F2 par {=PRODUITMAT(A2:D2;$K$16:$K$19)} (même si le résultat est un scalaire, il s'agit quand même d'une opération matricielle, il faut valider avec CTRL + SHIFT + ENTREE), et compléter la colonne par « copier – coller » vers le bas.

	A	B	C	D	E	F
1		X			Y	Y^
2	1	14.1	0.86	0.9853	13.6	14.46
3	1	16	1.06	1.0938	16.6	16.47
4	1	8	0.67	0.928	10.2	8.83
5	1	4.1	0.4	0.9462	5.4	5.26
6	1	15	1.04	0.8885	15	15.15
7	1	8.8	0.76	1.0267	9	9.79
8	1	12.4	0.95	0.9225	12.3	12.87
9	1	16.6	1.12	0.9372	16.3	16.71
10	1	14.9	1.02	0.8858	15.4	15.04
11	1	13.7	1.01	0.9643	13	14.14
12	1	15.1	0.9	0.9316	14.4	15.25
13	1	7.8	0.57	0.9705	10	8.68
14	1	11.4	0.78	1.124	10.2	12.31
15	1	9	0.74	0.8517	9.5	9.59
16	1	1	0.13	0.7851	1.5	2.04
17	1	17	1.26	0.9186	18.5	17.10
18	1	12.8	1.08	1.0395	12.6	13.53
19	1	15.8	0.96	0.9573	17.5	15.96
20	1	4.5	0.42	0.9106	4.9	5.55
21	1	14.5	1.01	1.007	15.9	14.94
22	1	7.3	0.61	0.9806	8.5	8.28
23	1	8.6	0.69	0.9693	10.6	9.45
24	1	15.2	1.02	0.9496	13.9	15.44
25	1	12	0.82	1.1184	14.9	12.85

	I	J	K	L
		(X'X)		
	24	275.60	19.88	23.09
	275.60	3613.16	254.18	267.46
	19.88	254.18	18.09	19.27
	23.09	267.46	19.27	22.36
		(X'X)^-1		
	6.563	0.063	-0.939	-6.720
	0.063	0.028	-0.452	-0.015
	-0.939	-0.452	7.863	-0.399
	-6.720	-0.015	-0.399	7.510

X'y	a^
289.70	-0.5517
3742.85	0.8876
264.08	0.5185
281.15	2.0793

Figure 4 - Valeurs estimées de la variable dépendante (Feuille "matrices")

3.3 Calcul des résidus

Les résidus, les erreurs observées, de la régression s'obtiennent par la différence entre les valeurs observées et les valeurs prédites de la variable dépendante.

$$\hat{\varepsilon} = y - \hat{y}$$

Dans notre feuille, nous plaçons le calcul pour la première observation en (G2) avec =E2-F2, puis nous copions vers le bas (Figure 5).

3.4 Tableau d'analyse de variance et coefficient de détermination R^2

Nous pouvons maintenant construire le tableau d'analyse de variance qui est fondamental pour rendre compte de la qualité de la régression (Figure 5). Plusieurs grandeurs doivent être calculées :

- La variabilité totale de la variable dépendante, $SCT = \sum_i (y_i - \bar{y})^2$ [somme des carrés totaux] en (J25) avec =SOMME.CARRES.ECARTS(E2:E25), avec pour degrés de liberté $\mathrm{DDL} = \mathrm{n} - 1 = 23$;
- La variabilité non expliquée par la régression, $SCR = \sum_i (y_i - \bar{y})^2$ [somme des carrés résiduels] en (J24) avec =SOMME.CARRES(G2:G25), $\mathrm{DDL} = \mathrm{n} - (\mathrm{p} + 1) = 20$;
- La variabilité expliquée (J23) est obtenue par différence SCE = SCT – SCR = J25-J24, DDL = p = 3.
- Les carrés moyens sont obtenus par le rapport entre les sommes des carrés et les degrés de liberté : $CME = \frac{SCE}{p} = 128.949$ et $CMR = \frac{SCR}{n-p-1} = 1.345$

	A	B	C	D	E	F	G
1		X			Y	Y^	Résidus
2	1	14.1	0.86	0.9853	13.6	14.46	-0.86
3	1	16	1.06	1.0938	16.6	16.47	0.13
4	1	8	0.67	0.928	10.2	8.83	1.37
5	1	4.1	0.4	0.9462	5.4	5.26	0.14
6	1	15	1.04	0.8885	15	15.15	-0.15
7	1	8.8	0.76	1.0267	9	9.79	-0.79
8	1	12.4	0.95	0.9225	12.3	12.87	-0.57
9	1	16.6	1.12	0.9372	16.3	16.71	-0.41
10	1	14.9	1.02	0.8858	15.4	15.04	0.36
11	1	13.7	1.01	0.9643	13	14.14	-1.14
12	1	15.1	0.9	0.9316	14.4	15.25	-0.85
13	1	7.8	0.57	0.9705	10	8.68	1.32
14	1	11.4	0.78	1.124	10.2	12.31	-2.11
15	1	9	0.74	0.8517	9.5	9.59	-0.09
16	1	1	0.13	0.7851	1.5	2.04	-0.54
17	1	17	1.26	0.9186	18.5	17.10	1.40
18	1	12.8	1.08	1.0395	12.6	13.53	-0.93
19	1	15.8	0.96	0.9573	17.5	15.96	1.54
20	1	4.5	0.42	0.9106	4.9	5.55	-0.65
21	1	14.5	1.01	1.007	15.9	14.94	0.96
22	1	7.3	0.61	0.9806	8.5	8.28	0.22
23	1	8.6	0.69	0.9693	10.6	9.45	1.15
24	1	15.2	1.02	0.9496	13.9	15.44	-1.54
25	1	12	0.82	1.1184	14.9	12.85	2.05

(X'X)

24	275.60	19.88	23.09
275.60	3613.16	254.18	267.46
19.88	254.18	18.09	19.27
23.09	267.46	19.27	22.36

(X'X)^-1

6.563	0.063	-0.939	-6.720
0.063	0.028	-0.452	-0.015
-0.939	-0.452	7.863	-0.399
-6.720	-0.015	-0.399	7.510

X'y
289.70
3742.85
264.08
281.15

a^
-0.5517
0.8876
0.5185
2.0793

Tableau d'analyse de variance

Source	SC	DDL	CM
Expliquée	386.846	3	128.949
Résiduelle	26.904	20	1.345
Totale	413.750	23	

Figure 5 - Tableau d'analyse de variance (Feuille "matrices")

Nous pouvons en déduire le coefficient de détermination $R^2 = \frac{SCE}{SCT} = \frac{386.846}{413.750} = 0.935$

3.5 Calcul de la matrice de variance covariance des coefficients

La matrice de variance covariance estimée des coefficients estimés est égale à

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

Où la variance estimée de l'erreur (N9) s'écrit :

$$\hat{\sigma}_{\varepsilon}^2 = \frac{SCR}{n-p-1} = \frac{26.904}{20} = 1.34520$$

En (P9), nous plaçons la formule =$N$9*I9 que nous étirons par la suite à droite et vers le bas pour compléter la matrice de variance covariance (Figure 6).

	I	J	K	L	M	N	O	P	Q	R	S
1											
2		(X'X)									
3	24	275.60	19.88	23.09							
4	275.60	3613.16	254.18	267.46							
5	19.88	254.18	18.09	19.27							
6	23.09	267.46	19.27	22.36							
7											
8		(X'X)^-1				sigma²(espilon)		Mat.VAR.COVAR Coef.Estimés			
9	6.563	0.063	-0.939	-6.720		1.34520		8.829	0.085	-1.263	-9.040
10	0.063	0.028	-0.452	-0.015				0.085	0.038	-0.608	-0.021
11	-0.939	-0.452	7.863	-0.399				-1.263	-0.608	10.578	-0.537
12	-6.720	-0.015	-0.399	7.510				-9.040	-0.021	-0.537	10.102
13											
14											
15	X'y		a^					sigma(a0)	sigma(a1)	sigma(a2)	sigma(a3)
16	289.70		-0.5517					2.971	0.195	3.252	3.178
17	3742.85		0.8876								
18	264.08		0.5185								
19	281.15		2.0793								

Figure 6 - Matrice de variance covariance des coefficients estimés (Feuille "matrices")

Les écarts-type des coefficients estimés se lisent sur la diagonale principale de cette matrice, en passant à la racine carrée, soit $\hat{\sigma}_{\hat{a}_0} = \sqrt{8.829} = 2.971$; $\hat{\sigma}_{\hat{a}_1} = 0.195$; $\hat{\sigma}_{\hat{a}_2} = 3.252$; $\hat{\sigma}_{\hat{a}_3} = 3.178$

4 Utilisation de la fonction DROITEREG

Excel propose la fonction DROITEREG pour la régression linéaire multiple. Voyons si les résultats fournis concordent avec ceux obtenus dans la section précédente.

4.1 Utilisation de DROITEREG – Lecture des résultats

Nous copions les données dans la nouvelle feuille de calcul « droitereg ». Nous insérons la fonction DROITEREG en (G2:J6) avec : {=DROITEREG(E2:E25;B2:D25;1;1)]}. C'est une fonction matricielle puisqu'elle complète simultanément un bloc de cellules.

Attention ! Il faut sélectionner autant de colonnes qu'il y a de paramètres à estimer, 4 en l'occurrence pour nos données ; et systématiquement 5 lignes si l'on souhaite disposer des statistiques de la régression (Remarque : on ne sélectionnera qu'une ligne si l'on veut uniquement le vecteur des paramètres estimés â).

Les paramètres de la fonction sont respectivement :

- E2:E25, le vecteur des valeurs de la variable dépendante, sans l'en-tête ;
- B2:D25, la matrice des variables indépendantes, elles doivent être forcément contigües, la sélection multiple n'est pas possible ;
- 1 veut dire que l'on souhaite réaliser une régression avec constante ;
- 1 pour obtenir les résultats additionnels, situés à partir de la seconde ligne du tableau.

G2 {=DROITEREG(E2:E25;B2:D25;1;1)}

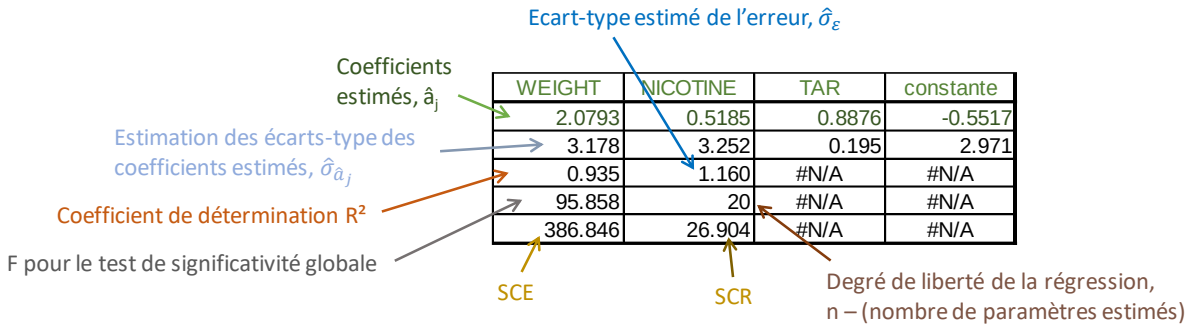
	A	B	C	D	E	F	G	H	I	J
1	Cigarette	TAR (mg)	NICOTINE (m	WEIGHT (g)	CO (mg)		WEIGHT	NICOTINE	TAR	constante
2	Alpine	14.1	0.86	0.9853	13.6		2.0793	0.5185	0.8876	-0.5517
3	Benson&Hedges	16	1.06	1.0938	16.6		3.178	3.252	0.195	2.971
4	CamelLights	8	0.67	0.928	10.2		0.935	1.160	#N/A	#N/A
5	Carlton	4.1	0.4	0.9462	5.4		95.858	20	#N/A	#N/A
6	Chesterfield	15	1.04	0.8885	15		386.846	26.904	#N/A	#N/A
7	GoldenLights	8.8	0.76	1.0267	9					
8	Kent	12.4	0.95	0.9225	12.3					
9	Kool	16.6	1.12	0.9372	16.3					
10	L&M	14.9	1.02	0.8858	15.4					
11	LarkLights	13.7	1.01	0.9643	13					
12	Marlboro	15.1	0.9	0.9316	14.4					
13	Merit	7.8	0.57	0.9705	10					
14	MultiFilter	11.4	0.78	1.124	10.2					
15	NewportLights	9	0.74	0.8517	9.5					
16	Now	1	0.13	0.7851	1.5					
17	OldGold	17	1.26	0.9186	18.5					
18	PallMallLight	12.8	1.08	1.0395	12.6					
19	Raleigh	15.8	0.96	0.9573	17.5					
20	SalemUltra	4.5	0.42	0.9106	4.9					
21	Tareyton	14.5	1.01	1.007	15.9					
22	TrueLight	7.3	0.61	0.9806	8.5					
23	ViceroyRichLight	8.6	0.69	0.9693	10.6					
24	VirginiaSlims	15.2	1.02	0.9496	13.9					
25	WinstonLights	12	0.82	1.1184	14.9					

Figure 7 - Sorties de DROITEREG (Feuille "droitereg")

La constante $\hat{a}_0$ est toujours en dernière colonne (la plus à droite) dans le tableau, quel que soit le nombre de variables indépendantes (Figure 7). Les autres coefficients sont dans l'ordre inverse de la matrice initiale des données. Je n'ai jamais su pourquoi. Il faut en tenir compte tout simplement lors de l'analyse des résultats. Nous avons, dans la première ligne du tableau, les coefficients estimés avec, de gauche à droite : $\hat{a}_3$=2.0793, $\hat{a}_2$=0.5185, $\hat{a}_1$=0.8876 et $\hat{a}_0$=-0.5517. Soit les mêmes résultats obtenus via le calcul matriciel (Figure 3).

4.2 Tableau d'analyse de variance

Les autres éléments fournis permettent de diagnostiquer la régression (Figure 8).



WEIGHT	NICOTINE	TAR	constante
2.0793	0.5185	0.8876	-0.5517
3.178	3.252	0.195	2.971
0.935	1.160	#N/A	#N/A
95.858	20	#N/A	#N/A
386.846	26.904	#N/A	#N/A

Figure 8 – Description des sorties de la fonction DROITEREG (Feuille "droitereg")

Il est ainsi possible, à partir des sorties de DROITEREG, de reconstituer le tableau d'analyse de variance en (G9:J12) (Figure 9) sans avoir à calculer explicitement les résidus de la régression.

	G	H	I	J
1	WEIGHT	NICOTINE	TAR	constante
2	2.0793	0.5185	0.8876	-0.5517
3	3.178	3.252	0.195	2.971
4	0.935	1.160	#N/A	#N/A
5	95.858	20	#N/A	#N/A
6	386.846	26.904	#N/A	#N/A
7				
8	Tableau d'analyse de variance			
9	Source	SC	DDL	CM
10	Expliquée	386.846	3	128.949
11	Résiduelle	26.904	20	1.345
12	Totale	413.750	23	

Figure 9 - Tableau d'analyse de variance via DROITREG (Feuille "droitereg")

Le coefficient de détermination est fourni directement par DROITEREG, R^2 = 0.935 (Figure 8).

5 Pratique de la régression

5.1 Test de significativité globale de la régression

Le test de significativité globale consiste à vérifier qu'il existe au moins une variable pertinente parmi les explicatives. La statistique de test F peut être obtenue par le rapport entre les carrés moyens expliqués et résiduels lus dans le tableau d'analyse de variance $F = \frac{CME}{CMR} = \frac{128.949}{1.345} = 95.858$; elle est également directement produite par la fonction DROITEREG. Les degrés de liberté sont lus dans le tableau d'analyse de variance. Nous calculons la p-value du test (en H18) avec la fonction de répartition de la loi de Fisher =LOI.F.DROITE(H14;H15;H16)

	G	H	I	J
1	WEIGHT	NICOTINE	TAR	constante
2	2.0793	0.5185	0.8876	-0.5517
3	3.178	3.252	0.195	2.971
4	0.935	1.160	#N/A	#N/A
5	95.858	20	#N/A	#N/A
6	386.846	26.904	#N/A	#N/A
7				
8	**Tableau d'analyse de variance**			
9	**Source**	**SC**	**DDL**	**CM**
10	Expliquée	386.846	3	128.949
11	Résiduelle	26.904	20	1.345
12	Totale	413.750	23	
13				
14	**Test de significativité globale**			
15	F	95.858		
16	DDL 1	3		
17	DDL 2	20		
18	p-value	4.85029E-12		

Figure 10 - Test de significativité globale (Feuille "droitereg")

Puisque la p-value est inférieure au risque 5% que l'on s'est choisi, nous concluons que la régression est globalement significative. Une des variables explicatives au moins est pertinente pour expliquer la variable dépendante.

5.2 Tests de significativité des coefficients

5.2.1 Test de significativité chaque coefficient pris individuellement

Nous passons maintenant à la significativité individuelle des coefficients pour évaluer la contribution de chaque variable. Nous testons uniquement les coefficients associés aux variables (a_1, a_2 et a_3), avec $H_0 : a_j = 0$ vs. $H_1 : a_j \neq 0$

Pour nous, il s'agit de former la statistique de test à l'aide du rapport entre le coefficient estimé et l'estimation de son écart-type fourni par DROITEREG (Figure 8).

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

Nous rentrons la première formule en G22 avec =G2/G3. Nous complétons le tableau par copier-coller à droite. Comme il s'agit d'un test bilatéral, nous utilisons la LOI.STUDENT.BILATERALE d'Excel pour calculer la p-value. Pour la première variable WEIGHT en G23, nous insérons =LOI.STUDENT.BILATERALE(ABS(G22);$I$11). Nous prenons la valeur absolue ABS() de la statistique de test parce qu'elle est susceptible d'être négative. Le degré de liberté correspond à celui de la régression.

	WEIGHT	NICOTINE	TAR	constante
	2.0793	0.5185	0.8876	-0.5517
	3.178	3.252	0.195	2.971
	0.935	1.160	#N/A	#N/A
	95.858	20	#N/A	#N/A
	386.846	26.904	#N/A	#N/A

Tableau d'analyse de variance

Source	SC	DDL	CM
Expliquée	386.846	3	128.949
Résiduelle	26.904	20	1.345
Totale	413.750	23	

Test de significativité globale

F	95.858
DDL 1	3
DDL 2	20
p-value	4.85029E-12

Test de significativité individuelle

	WEIGHT	NICOTINE	TAR
t-calculé	0.6542	0.1594	4.5405
p-value	0.5204	0.8749	0.0002

Figure 11 - Test de significativité individuelle des coefficients (Feuille "droitereg")

Au risque 5%, seul le coefficient de la variable TAR est significatif c.-à-d. significativement différent de 0 avec une p-value inférieure à 0.05.

5.2.2 Test de significativité d'un bloc de coefficients

Dans la section précédente, les coefficients de WEIGHT et NICOTINE, pris individuellement, ne semblent pas significativement différent de 0. Cela ne veut pas dire qu'ils sont simultanément nuls. Il faudrait passer par un test spécifique pour le vérifier, où H_0 : $a_{weight}=a_{nicotine}=0$ vs. H_1 : un des deux coefficients au moins est non nul.

Nous copions les données et les résultats de DROITEREG avec l'ensemble des variables dans la feuille « test bloc » (Figure 12). Le coefficient de détermination sous l'hypothèse alternative est $R_1^2 = 0.93498$ (cf. Figure 8 ; la précision a été augmentée pour rendre les calculs plus lisibles).

Nous lui opposons la régression sous hypothèse nulle c.-à-d. avec uniquement la variable TAR, ($q = 2$) variables ont été retirées. Le coefficient de détermination va mécaniquement diminuer puisque nous avons deux modèles imbriqués, et que ce dernier a moins de variables que le premier. Toute la question est de savoir si cette diminution est substantielle, indiquant une contribution notoire d'au moins une des deux variables qui ont été exclues. Les données contrediraient l'hypothèse nulle dans ce cas.

Pour la régression avec TAR seule (G8:H13), nous obtenons $R_0^2 = 0.93346$, plus faible comme prévu (Figure 12). Nous formons la statistique de test en H16 :

$$F = \frac{(R_1^2 - R_0^2)/q}{(1 - R_1^2)/(n - p - 1)} = \frac{(0.93498 - 0.93346)/2}{(1 - 0.93498)/20} = 0.23274$$

	A	B	C	D	E
1	Cigarette	TAR (mg)	NICOTINE	WEIGHT (g)	CO (mg)
2	Alpine	14.1	0.86	0.9853	13.6
3	Benson&Hed	16	1.06	1.0938	16.6
4	CamelLights	8	0.67	0.928	10.2
5	Carlton	4.1	0.4	0.9462	5.4
6	Chesterfield	15	1.04	0.8885	15
7	GoldenLights	8.8	0.76	1.0267	9
8	Kent	12.4	0.95	0.9225	12.3
9	Kool	16.6	1.12	0.9372	16.3
10	L&M	14.9	1.02	0.8858	15.4
11	LarkLights	13.7	1.01	0.9643	13
12	Marlboro	15.1	0.9	0.9316	14.4
13	Merit	7.8	0.57	0.9705	10
14	MultiFilter	11.4	0.78	1.124	10.2
15	NewportLight	9	0.74	0.8517	9.5
16	Now	1	0.13	0.7851	1.5
17	OldGold	17	1.26	0.9186	18.5
18	PallMallLight	12.8	1.08	1.0395	12.6
19	Raleigh	15.8	0.96	0.9573	17.5
20	SalemUltra	4.5	0.42	0.9106	4.9
21	Tareyton	14.5	1.01	1.007	15.9
22	TrueLight	7.3	0.61	0.9806	8.5
23	ViceroyRichL	8.6	0.69	0.9693	10.6
24	VirginiaSlims	15.2	1.02	0.9496	13.9
25	WinstonLight	12	0.82	1.1184	14.9

WEIGHT	NICOTINE	TAR	constante
2.0793	0.5185	0.8876	-0.5517
3.178	3.252	0.195	2.971
0.93498	1.160	#N/A	#N/A
95.858	20	#N/A	#N/A
386.846	26.904	#N/A	#N/A

TAR	constante
0.9281	1.4129
0.0528	0.6482
0.93346	1.1186
308.6377	22
386.2195	27.5301

Test (a_weight = a_nicotine = 0)

F calculé	0.23274
ddl 1 (q)	2
ddl 2 (n - p -1)	20
p-value	0.79447

Figure 12 - Test de significativité d'un bloc de coefficients (Feuille "test bloc")

La probabilité critique (en H19) est fournie par =LOI.F.DROITE(H16;H17;H18). Elle est égale 0.79447. Les données ne contredisent pas l'hypothèse nulle.

Remarque : Néanmoins, pour des raisons pédagogiques, nous conserverons l'ensemble des variables explicatives dans la suite de ce document.

5.3 Intervalle de confiance des coefficients

Nous souhaitons calculer les intervalles de confiance au niveau 1-α = 95% des coefficients de la régression impliquant l'ensemble des variables explicatives. Nous copions les données et le tableau de DROITEREG dans une nouvelle feuille « étude coefs ».

Pour rappel, les bornes s'écrivent :

$$\hat{a}_j \pm t \times \hat{\sigma}_{\hat{a}_j}$$

Les coefficients estimés et les écarts-types sont produites par DROITEREG (Figure 8), t est le quantile de la loi de Student en G8 avec =LOI.STUDENT.INVERSE.BILATERALE(0.05;$H$5), le premier paramètre correspond à α, le second au degré de liberté de la régression.

Pour le coefficient de la variable WEIGHT par exemple, nous avons respectivement en G12 et G13 pour les bornes basses et hautes : =G2-$G$8*G3 et =G2+$G$8*G3

	A	B	C	D	E
1	Cigarette	TAR (mg)	NICOTINE (m	WEIGHT (g)	CO (mg)
2	Alpine	14.1	0.86	0.9853	13.6
3	Benson&Hed	16	1.06	1.0938	16.6
4	CamelLights	8	0.67	0.928	10.2
5	Carlton	4.1	0.4	0.9462	5.4
6	Chesterfield	15	1.04	0.8885	15
7	GoldenLights	8.8	0.76	1.0267	9
8	Kent	12.4	0.95	0.9225	12.3
9	Kool	16.6	1.12	0.9372	16.3
10	L&M	14.9	1.02	0.8858	15.4
11	LarkLights	13.7	1.01	0.9643	13
12	Marlboro	15.1	0.9	0.9316	14.4
13	Merit	7.8	0.57	0.9705	10
14	MultiFilter	11.4	0.78	1.124	10.2
15	NewportLight:	9	0.74	0.8517	9.5
16	Now	1	0.13	0.7851	1.5
17	OldGold	17	1.26	0.9186	18.5
18	PallMallLight	12.8	1.08	1.0395	12.6
19	Raleigh	15.8	0.96	0.9573	17.5
20	SalemUltra	4.5	0.42	0.9106	4.9
21	Tareyton	14.5	1.01	1.007	15.9
22	TrueLight	7.3	0.61	0.9806	8.5
23	ViceroyRichL	8.6	0.69	0.9693	10.6
24	VirginiaSlims	15.2	1.02	0.9496	13.9
25	WinstonLight	12	0.82	1.1184	14.9

	F	G	H	I	J
		WEIGHT	NICOTINE	TAR	constante
	coef	2.0793	0.5185	0.8876	-0.5517
	ecart.type	3.178	3.252	0.195	2.971
		0.93498	1.160	#N/A	#N/A
		95.858	**20**	#N/A	#N/A
		386.846	26.904	#N/A	#N/A
	quantile	2.08596			

Intervalles de confiance des coefficients

	WEIGHT	NICOTINE	TAR	constante
B.basse	-4.5507	-6.2658	0.4798	-6.7497
B.Haute	8.7094	7.3027	1.2953	5.6463

Figure 13 - Intervalle de confiance des coefficients à 95% (Feuille "etude coefs")

TAR est bien la seule variable pertinente puisque son intervalle ne couvre pas la valeur 0. Ces résultats sont cohérents avec ceux des tests de significativité individuelle (Figure 11).

5.4 Influence comparée des coefficients – Coefficients standardisés

Comparer les influences des variables dans la régression à travers les valeurs absolues des coefficients n'est pas une bonne idée parce qu'elles sont généralement définies sur des échelles différentes. Une solution serait de les centrer et réduire avant de lancer la régression pour évacuer les problèmes d'unités. Les coefficients de la régression sont alors dits « standardisés » (ou « coefficients β »). Ils sont directement comparables car ils expriment des variations en écarts-type.

Plutôt que relancer les calculs sur les variables transformées, on peut post-traiter les coefficients de la régression avec les écart-types des variables. Le coefficient β pour la variable x_j s'écrit :

$$\hat{\beta}_j = \hat{a}_j \times \frac{\sigma_{x_j}}{\sigma_y}$$

Nous calculons les écarts-type de l'ensemble des variables en (B27:E27). Pour la variable **TAR** par exemple, nous avons =ECARTYPE.STANDARD(B2:B25). Puis nous appliquons la formule ci-dessous pour chaque coefficient. Attention, les variables ne sont pas dans le même ordre dans les tableaux de données et de DROITEREG (merci Excel !). Pour simplifier, je les ai saisies unes à unes en ce qui me concerne (G17:I17). Pour WEIGHT par exemple, nous avons en G17 : =G2*D27/$E$27

	A	B	C	D	E
1	Cigarette	TAR (mg)	NICOTINE (m	WEIGHT (g)	CO (mg)
2	Alpine	14.1	0.86	0.9853	13.6
3	Benson&Hed	16	1.06	1.0938	16.6
4	CamelLights	8	0.67	0.928	10.2
5	Carlton	4.1	0.4	0.9462	5.4
6	Chesterfield	15	1.04	0.8885	15
7	GoldenLights	8.8	0.76	1.0267	9
8	Kent	12.4	0.95	0.9225	12.3
9	Kool	16.6	1.12	0.9372	16.3
10	L&M	14.9	1.02	0.8858	15.4
11	LarkLights	13.7	1.01	0.9643	13
12	Marlboro	15.1	0.9	0.9316	14.4
13	Merit	7.8	0.57	0.9705	10
14	MultiFilter	11.4	0.78	1.124	10.2
15	NewportLight	9	0.74	0.8517	9.5
16	Now	1	0.13	0.7851	1.5
17	OldGold	17	1.26	0.9186	18.5
18	PallMallLight	12.8	1.08	1.0395	12.6
19	Raleigh	15.8	0.96	0.9573	17.5
20	SalemUltra	4.5	0.42	0.9106	4.9
21	Tareyton	14.5	1.01	1.007	15.9
22	TrueLight	7.3	0.61	0.9806	8.5
23	ViceroyRichL	8.6	0.69	0.9693	10.6
24	VirginiaSlims	15.2	1.02	0.9496	13.9
25	WinstonLight	12	0.82	1.1184	14.9
26					
27	**Ecart-type**	4.4152	0.2656	0.0795	4.2414

	F	G	H	I	J
		WEIGHT	NICOTINE	TAR	constante
	coef	2.0793	0.5185	0.8876	-0.5517
	ecart.type	3.178	3.252	0.195	2.971
		0.93498	1.160	#N/A	#N/A
		95.858	**20**	#N/A	#N/A
		386.846	26.904	#N/A	#N/A
	quantile	2.08596			

Intervalles de confiance des coefficients

	WEIGHT	NICOTINE	TAR	constante
B.basse	-4.5507	-6.2658	0.4798	-6.7497
B.Haute	8.7094	7.3027	1.2953	5.6463

	WEIGHT	NICOTINE	TAR
Coef.Stand	0.0390	0.0325	0.9240

Figure 14 - Coefficients standardisés (Feuille "etude coefs")

La variable TAR est la plus influente avec $\hat{\beta}_{TAR} = 0.9240$. Ce n'est pas vraiment étonnant, son coefficient était le seul significatif dans la régression. Au moins, les résultats sont cohérents.

5.5 Etude des résidus

Pour étudier les résidus, nous créons la nouvelle feuille « résidus » où nous copions la variable dépendante, les valeurs prédites, et les résidus disponibles dans la feuille « matrices ». Nous effectuons un copier – collage spécial « valeurs » pour éviter les problèmes de références.

5.5.1 Graphique des résidus

Le graphique des résidus (graphique nuage de points) est un puissant outil de diagnostic. Nous plaçons en ordonnée les résidus, en abscisse, selon l'analyse que nous souhaitons mener, nous insérons la variable adaptée. Dans notre cas, nous plaçons la variable dépendante CO (y). La

régression est suspecte dès que l'on observe des formes de régularité ou des exceptions parmi les points. Ce qui ne semble pas trop être le cas pour nous même si l'on constate une sous-estimation de la variable dépendante lorsqu'elle prend des valeurs élevées (les résidus ont tendance à prendre des valeurs positives sur la partie droite de l'abscisse).

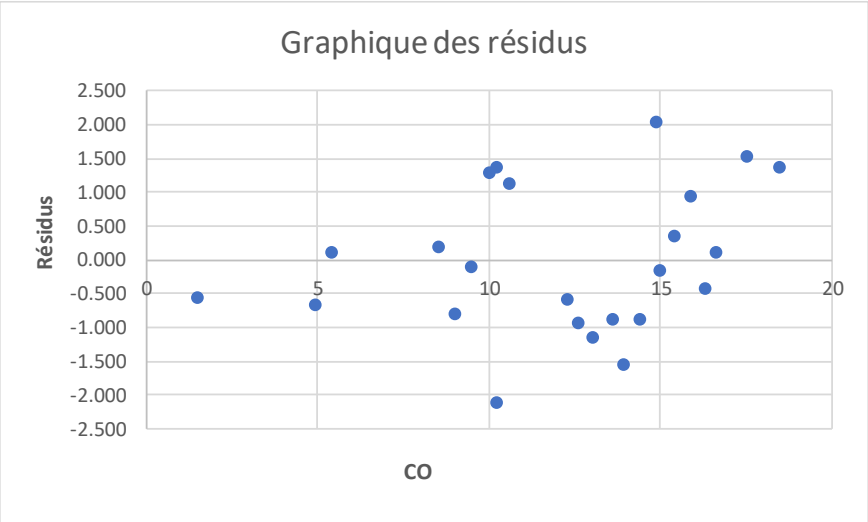


Figure 15 - Graphique des résidus (Feuille "résidus")

5.5.2 Test de normalité des résidus – Droite de Henry

L'hypothèse de normalité des erreurs joue un rôle très important dans la partie inférentielle de la régression. Il est tout à fait légitime que nous la vérifiions pour nos données.

La droite de Henry est une diagramme quantile-quantile (Q-Q plot) confrontant les quantiles observés avec ceux que l'on aurait sous hypothèse de normalité (quantiles théoriques). S'ils sont cohérents, c.-à-d. forment une droite, l'hypothèse est crédible.

Nous travaillons en plusieurs étapes sous Excel (Figure 16) :

- Trier les données par ordre croissant des résidus observés ;
- Former la fréquence espérée pour la loi normale $F_i = \frac{i-0.375}{n+0.25}$, où i est le numéro d'individu dans la suite triée, n est le nombre d'observations[1] ;

[1] En réalité, la formule de F_i peut être un peu différente selon que n est > 10 ou ≤ 10, mais on ne va pas s'attarder là-dessus. D'autant plus que les différentes références divergent à ce sujet.

- Calculer le quantile de la loi normale centrée et réduite en utilisant l'inverse de la fonction de répartition, LOI.NORMALE.STANDARD.INVERSE.N() sous Excel ;
- Calculer les résidus théoriques en revenant sur l'échelle initiale des résidus en multipliant ces quantiles par l'estimation de l'écart-type de l'erreur $\hat{\sigma}_\varepsilon = 1.160$
- La droite de Henry confronte les résidus observés et théoriques.

	A	B	C	D	E	F	G
1	Y	Y^	Rés.Obs.	i	F_i	Quantiles	Rés.Théoriques
2	10.2	12.308	-2.108	1	0.026	-1.947	-2.258
3	13.9	15.443	-1.543	2	0.067	-1.498	-1.738
4	13	14.137	-1.137	3	0.108	-1.236	-1.433
5	12.6	13.531	-0.931	4	0.149	-1.039	-1.205
6	13.6	14.458	-0.858	5	0.191	-0.875	-1.015
7	14.4	15.255	-0.855	6	0.232	-0.732	-0.849
8	9	9.788	-0.788	7	0.273	-0.603	-0.700
9	4.9	5.554	-0.654	8	0.314	-0.483	-0.561
10	12.3	12.865	-0.565	9	0.356	-0.370	-0.429
11	1.5	2.036	-0.536	10	0.397	-0.261	-0.303
12	16.3	16.712	-0.412	11	0.438	-0.156	-0.181
13	15	15.149	-0.149	12	0.479	-0.052	-0.060
14	9.5	9.591	-0.091	13	0.521	0.052	0.060
15	16.6	16.474	0.126	14	0.562	0.156	0.181
16	5.4	5.262	0.138	15	0.603	0.261	0.303
17	8.5	8.283	0.217	16	0.644	0.370	0.429
18	15.4	15.044	0.356	17	0.686	0.483	0.561
19	15.9	14.936	0.964	18	0.727	0.603	0.700
20	10.6	9.455	1.145	19	0.768	0.732	0.849
21	10	8.685	1.315	20	0.809	0.875	1.015
22	10.2	8.826	1.374	21	0.851	1.039	1.205
23	18.5	17.101	1.399	22	0.892	1.236	1.433
24	17.5	15.960	1.540	23	0.933	1.498	1.738
25	14.9	12.850	2.050	24	0.974	1.947	2.258

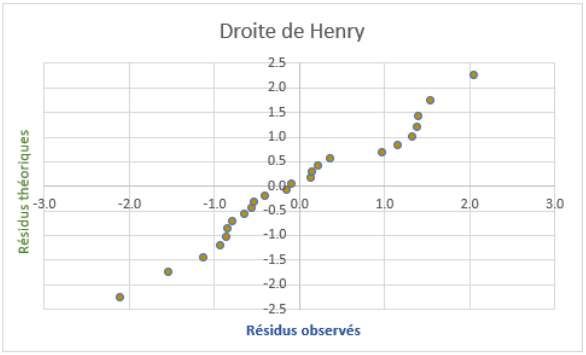


Figure 16 - Droite de Henry (Feuille "résidus")

L'hypothèse de normalité des résidus semble crédible concernant notre régression.

5.5.3 Test de normalité des résidus – Test de Jarque-Bera

Le test de Jarque-Bera est une alternative numérique à l'approche graphique ci-dessus. Elle est basée sur les coefficients d'asymétrie γ_1 et d'aplatissement γ_2 de la distribution. Les deux sont nuls lorsque nous avons affaire à la loi normale. L'hypothèse de normalité n'est pas crédible dans le cas contraire c.-à-d. lorsque ces coefficients diffèrent significativement de 0 sur nos données.

Le coefficient d'asymétrie est estimé par :

$$g_1 = \frac{\frac{1}{n}\sum_i \hat{\varepsilon}_i^3}{\left[\frac{1}{n}\sum_i \hat{\varepsilon}_i^2\right]^{3/2}} = 0.16084$$

Pour réaliser les calculs, nous créons une nouvelle feuille « Jarque-Bera » (Figure 17). Nous récupérons les résidus de la feuille « matrices ». Nous calculons les deux colonnes supplémentaires

$\hat{\varepsilon}^2$ (C2:C25) et $\hat{\varepsilon}^3$ (B2:B25). Nous calculons **g_1**=MOYENNE(C2:C25)/(MOYENNE(B2:B25)^(3/2)). Ce qui correspond à la fonction COEFFICIENT.ASYMETRIE.P() d'Excel.

Pour le coefficient d'aplatissement, nous utilisons :

$$g_2 = \frac{\frac{1}{n}\sum_i \hat{\varepsilon}_i^4}{\left[\frac{1}{n}\sum_i \hat{\varepsilon}_i^2\right]^2} - 3 = -0.81232$$

Nous rajoutons $\hat{\varepsilon}^4$ dans notre feuille (D2:D25) et nous formons **g_2**=MOYENNE(D2:D25)/(MOYENNE(B2:B25)^2)-3. Notre formule diffère de celle de la fonction KURTOSIS() d'Excel. Les deux sont cependant asymptotiquement équivalentes (pour n élevé).

	A	B	C	D	E	F	G
1	Résidus	Résidus^2	Résidus^3	Résidus^4			
2	-0.858	0.736	-0.631	0.542			
3	0.126	0.016	0.002	0.000		g1	0.16084
4	1.374	1.888	2.594	3.565			
5	0.138	0.019	0.003	0.000		g2	-0.81232
6	-0.149	0.022	-0.003	0.000			
7	-0.788	0.621	-0.489	0.385		T	0.63611
8	-0.565	0.319	-0.180	0.102			
9	-0.412	0.169	-0.070	0.029		p-value	0.72756
10	0.356	0.127	0.045	0.016			
11	-1.137	1.293	-1.470	1.671			
12	-0.855	0.730	-0.624	0.533			
13	1.315	1.729	2.274	2.991			
14	-2.108	4.445	-9.371	19.758			
15	-0.091	0.008	-0.001	0.000			
16	-0.536	0.287	-0.154	0.082			
17	1.399	1.959	2.741	3.836			
18	-0.931	0.866	-0.806	0.750			
19	1.540	2.370	3.650	5.619			
20	-0.654	0.427	-0.279	0.183			
21	0.964	0.930	0.896	0.864			
22	0.217	0.047	0.010	0.002			
23	1.145	1.312	1.502	1.720			
24	-1.543	2.381	-3.673	5.667			
25	2.050	4.203	8.616	17.663			

Figure 17 - Test de Jarque-Bera (Feuille "Jarque-Bera")

La statistique de test (G7) s'écrit :

$$T = \frac{n-p-1}{6}\left(g_1^2 + \frac{g_2^2}{4}\right) = \frac{24-3-1}{6}\left(0.16084^2 + \frac{-0.81232^2}{4}\right) = 0.63611$$

Sous l'hypothèse de normalité (H_0), elle suit une loi du χ^2 à 2 degrés de liberté. En G9, nous calculons la p-value =LOI.KHIDEUX.DROITE(G7;2). Elle est égale à 0.72756, l'hypothèse de normalité des

erreurs ne peut pas être rejetée. L'analyse graphique ci-dessus (Droite de Henry, Figure 16) est confortée.

5.6 Points atypiques et influents

Surtout sur de petits effectifs comme le notre ($n=24$), l'analyse des points atypiques et influents est importante. L'objectif est de détecter les observations qui pèsent indûment sur la régression et sont susceptibles de la fausser.

5.6.1 Levier

Le levier est une sorte de distance au barycentre des points dans l'espace des variables explicatives. Il identifie les observations anormalement éloignées des autres. Pour l'observation n°i,

$$h_i = x_i(X'X)^{-1}x'_i$$

Où x_i correspond à la description de l'observation n°i, incluant la constante. Par exemple, pour la première observation (Alpine), nous aurons :

$$x_1 = (1\ ;14.1\ ;0.86\ ;0.9853)$$

Son levier serait alors, en récupérant la matrice $(X'X)^{-1}$ dans la feuille « matrices » :

$$(1\ ;14.1\ ;0.86\ ;0.9853)\begin{pmatrix} 6.563 & 0.063 & -0.939 & -6.720 \\ 0.063 & 0.028 & -0.452 & -0.015 \\ -0.939 & -0.452 & 7.863 & -0.399 \\ -6.720 & -0.015 & -0.399 & 7.510 \end{pmatrix}\begin{pmatrix} 1 \\ 14.1 \\ 0.86 \\ 0.9853 \end{pmatrix} = 0.1707$$

Nous créons une nouvelle feuille « outliers » pour nos calculs. Nous y copions les labels des observations (cf. « cigarettes »), les descriptions (X) et les résidus $\hat{\varepsilon}$, la matrice $(X'X)^{-1}$ (cf. « matrices »). Il faut faire un copier – coller / collage spécial valeurs pour éviter les problèmes de référencement. La feuille, dans sa phase préparatoire devrait ressemblait à ceci (Figure 18).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Cigarette	X				Résidus									
2	Alpine	1	14.1	0.86	0.9853	-0.858									
3	Benson&Hedges	1	16	1.06	1.0938	0.126									
4	CamelLights	1	8	0.67	0.928	1.374									
5	Carlton	1	4.1	0.4	0.9462	0.138									
6	Chesterfield	1	15	1.04	0.8885	-0.149									
7	GoldenLights	1	8.8	0.76	1.0267	-0.788							(X'X)^-1		
8	Kent	1	12.4	0.95	0.9225	-0.565						6.563	0.063	-0.939	-6.720
9	Kool	1	16.6	1.12	0.9372	-0.412						0.063	0.028	-0.452	-0.015
10	L&M	1	14.9	1.02	0.8858	0.356						-0.939	-0.452	7.863	-0.399
11	LarkLights	1	13.7	1.01	0.9643	-1.137						-6.720	-0.015	-0.399	7.510
12	Marlboro	1	15.1	0.9	0.9316	-0.855									
13	Merit	1	7.8	0.57	0.9705	1.315									
14	MultiFilter	1	11.4	0.78	1.124	-2.108									
15	NewportLights	1	9	0.74	0.8517	-0.091									
16	Now	1	1	0.13	0.7851	-0.536									
17	OldGold	1	17	1.26	0.9186	1.399									
18	PallMallLight	1	12.8	1.08	1.0395	-0.931									
19	Raleigh	1	15.8	0.96	0.9573	1.540									
20	SalemUltra	1	4.5	0.42	0.9106	-0.654									
21	Tareyton	1	14.5	1.01	1.007	0.964									
22	TrueLight	1	7.3	0.61	0.9806	0.217									
23	ViceroyRichLight	1	8.6	0.69	0.9693	1.145									
24	VirginiaSlims	1	15.2	1.02	0.9496	-1.543									
25	WinstonLights	1	12	0.82	1.1184	2.050									

Figure 18 - Préparation de la feuille "outliers"

La formule matricielle sous Excel demande quelques contorsions pour rentrer la première valeur en G2. Nous avons : {=PRODUITMAT(PRODUITMAT(B2:E2;$L$8:$O$11);TRANSPOSE(B2:E2))}. Nous complétons la colonne par un copier-coller vers le bas.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Cigarette	X				Résidus	Levier								
2	Alpine	1	14.1	0.86	0.9853	-0.858	0.1707								
3	Benson&Hedges	1	16	1.06	1.0938	0.126	0.1849								
4	CamelLights	1	8	0.67	0.928	1.374	0.0857								
5	Carlton	1	4.1	0.4	0.9462	0.138	0.1669								
6	Chesterfield	1	15	1.04	0.8885	-0.149	0.1335								
7	GoldenLights	1	8.8	0.76	1.0267	-0.788	0.1572						(X'X)^-1		
8	Kent	1	12.4	0.95	0.9225	-0.565	0.0979					6.563	0.063	-0.939	-6.720
9	Kool	1	16.6	1.12	0.9372	-0.412	0.1196					0.063	0.028	-0.452	-0.015
10	L&M	1	14.9	1.02	0.8858	0.356	0.1336					-0.939	-0.452	7.863	-0.399
11	LarkLights	1	13.7	1.01	0.9643	-1.137	0.0763					-6.720	-0.015	-0.399	7.510
12	Marlboro	1	15.1	0.9	0.9316	-0.855	0.2315								
13	Merit	1	7.8	0.57	0.9705	1.315	0.0948								
14	MultiFilter	1	11.4	0.78	1.124	-2.108	0.2599								
15	NewportLights	1	9	0.74	0.8517	-0.091	0.1554								
16	Now	1	1	0.13	0.7851	-0.536	0.4603								
17	OldGold	1	17	1.26	0.9186	1.399	0.2553								
18	PallMallLight	1	12.8	1.08	1.0395	-0.931	0.3157								
19	Raleigh	1	15.8	0.96	0.9573	1.540	0.1948								
20	SalemUltra	1	4.5	0.42	0.9106	-0.654	0.1525								
21	Tareyton	1	14.5	1.01	1.007	0.964	0.0687								
22	TrueLight	1	7.3	0.61	0.9806	0.217	0.0961								
23	ViceroyRichLight	1	8.6	0.69	0.9693	1.145	0.0695								
24	VirginiaSlims	1	15.2	1.02	0.9496	-1.543	0.0835								
25	WinstonLights	1	12	0.82	1.1184	2.050	0.2356								
26															
27						Seuil	0.3333								

Figure 19 - Calcul du levier (Feuille "outliers")

On considère qu'une observation est suspecte lorsque

$$h_i \geq 2 \times \frac{p+1}{n} = 2 \times \frac{3+1}{24} = 0.3333$$

Au regard du levier, il faudrait se pencher un peu plus sur la cigarette « Now ». Nous constatons qu'elle se distingue par de très faibles teneurs en TAR et NICOTINE par rapport aux autres marques.

5.6.2 Résidus studentisés

Le résidu studentisé sert à mettre en évidence les observations mal modélisées c.-à-d. qui se démarquent par rapport à la relation entre les explicatives et l'expliquée mise en lumière par la régression. Nous passons d'abord par le calcul du résidu standardisé :

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_\varepsilon \sqrt{1-h_i}}$$

L'estimation de l'écart-type de l'erreur $\hat{\sigma}_\varepsilon$ peut être lue dans le tableau fourni par DROITEREG (Figure 8). Nous complétons la feuille « outliers » en insérant en H2 : =F2/($L$4*RACINE(1-G2)). La colonne est complétée par un copier-coller vers le bas ().

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Cigarette	X				Résidus	Levier	Rés.Stand.							
2	Alpine	1	14.1	0.86	0.9853	-0.858	0.1707	-0.8122							
3	Benson&Hedges	1	16	1.06	1.0938	0.126	0.1849	0.1208				sigma^(epsilon)			
4	CamelLights	1	8	0.67	0.928	1.374	0.0857	1.2390				1.1598			
5	Carlton	1	4.1	0.4	0.9462	0.138	0.1669	0.1301							
6	Chesterfield	1	15	1.04	0.8885	-0.149	0.1335	-0.1377							
7	GoldenLights	1	8.8	0.76	1.0267	-0.788	0.1572	-0.7400				(X'X)^-1			
8	Kent	1	12.4	0.95	0.9225	-0.565	0.0979	-0.5129				6.563	0.063	-0.939	-6.720
9	Kool	1	16.6	1.12	0.9372	-0.412	0.1196	-0.3782				0.063	0.028	-0.452	-0.015
10	L&M	1	14.9	1.02	0.8858	0.356	0.1336	0.3298				-0.939	-0.452	7.863	-0.399
11	LarkLights	1	13.7	1.01	0.9643	-1.137	0.0763	-1.0199				-6.720	-0.015	-0.399	7.510
12	Marlboro	1	15.1	0.9	0.9316	-0.855	0.2315	-0.8404							
13	Merit	1	7.8	0.57	0.9705	1.315	0.0948	1.1917							
14	MultiFilter	1	11.4	0.78	1.124	-2.108	0.2599	-2.1130							
15	NewportLights	1	9	0.74	0.8517	-0.091	0.1554	-0.0855							
16	Now	1	1	0.13	0.7851	-0.536	0.4603	-0.6288							
17	OldGold	1	17	1.26	0.9186	1.399	0.2553	1.3982							
18	PallMallLight	1	12.8	1.08	1.0395	-0.931	0.3157	-0.9701							
19	Raleigh	1	15.8	0.96	0.9573	1.540	0.1948	1.4794							
20	SalemUltra	1	4.5	0.42	0.9106	-0.654	0.1525	-0.6122							
21	Tareyton	1	14.5	1.01	1.007	0.964	0.0687	0.8615							
22	TrueLight	1	7.3	0.61	0.9806	0.217	0.0961	0.1969							
23	ViceroyRichLight	1	8.6	0.69	0.9693	1.145	0.0695	1.0237							
24	VirginiaSlims	1	15.2	1.02	0.9496	-1.543	0.0835	-1.3896							
25	WinstonLights	1	12	0.82	1.1184	2.050	0.2356	2.0216							
26															
27						Seuil	0.3333								

Figure 20 - Calcul du résidu standardisé (Feuille "outliers")

On pourrait définir une valeur seuil ici, mais la détection sera plus puissante avec le résidu studentisé. Ce dernier est dérivé du résidu standardisé :

$$t_i^* = t_i \sqrt{\frac{n-p-2}{n-p-1-t_i^2}}$$

Dans la feuille de calcul (Figure 21), deux marques se distinguent : MultiFilter dont la nocivité est surestimée par le modèle c.-à-d. par rapport aux autres cigarettes, sa teneur en CO est faible au regard de ses caractéristiques ; celle de WinstonLights est, à l'inverse, sous-estimée.

	A	B	C	D	E	F	G	H	I
1	Cigarette	X				Résidus	Levier	Rés.Stand.	Res.Stud
2	Alpine	1	14.1	0.86	0.9853	-0.858	0.1707	-0.8122	-0.8050
3	Benson&Hedges	1	16	1.06	1.0938	0.126	0.1849	0.1208	0.1177
4	CamelLights	1	8	0.67	0.928	1.374	0.0857	1.2390	1.2568
5	Carlton	1	4.1	0.4	0.9462	0.138	0.1669	0.1301	0.1269
6	Chesterfield	1	15	1.04	0.8885	-0.149	0.1335	-0.1377	-0.1343
7	GoldenLights	1	8.8	0.76	1.0267	-0.788	0.1572	-0.7400	-0.7313
8	Kent	1	12.4	0.95	0.9225	-0.565	0.0979	-0.5129	-0.5033
9	Kool	1	16.6	1.12	0.9372	-0.412	0.1196	-0.3782	-0.3700
10	L&M	1	14.9	1.02	0.8858	0.356	0.1336	0.3298	0.3223
11	LarkLights	1	13.7	1.01	0.9643	-1.137	0.0763	-1.0199	-1.0210
12	Marlboro	1	15.1	0.9	0.9316	-0.855	0.2315	-0.8404	-0.8340
13	Merit	1	7.8	0.57	0.9705	1.315	0.0948	1.1917	1.2051
14	MultiFilter	1	11.4	0.78	1.124	-2.108	0.2599	-2.1130	-2.3368
15	NewportLights	1	9	0.74	0.8517	-0.091	0.1554	-0.0855	-0.0834
16	Now	1	1	0.13	0.7851	-0.536	0.4603	-0.6288	-0.6191
17	OldGold	1	17	1.26	0.9186	1.399	0.2553	1.3982	1.4347
18	PallMallLight	1	12.8	1.08	1.0395	-0.931	0.3157	-0.9701	-0.9686
19	Raleigh	1	15.8	0.96	0.9573	1.540	0.1948	1.4794	1.5280
20	SalemUltra	1	4.5	0.42	0.9106	-0.654	0.1525	-0.6122	-0.6023
21	Tareyton	1	14.5	1.01	1.007	0.964	0.0687	0.8615	0.8557
22	TrueLight	1	7.3	0.61	0.9806	0.217	0.0961	0.1969	0.1921
23	ViceroyRichLight	1	8.6	0.69	0.9693	1.145	0.0695	1.0237	1.0250
24	VirginiaSlims	1	15.2	1.02	0.9496	-1.543	0.0835	-1.3896	-1.4249
25	WinstonLights	1	12	0.82	1.1184	2.050	0.2356	2.0216	2.2090
26									
27						Seuil	0.3333		2.0930

	L	M	N	O
	sigma^(epsilon)			
	1.1598			
	(X'X)^-1			
	6.563	0.063	-0.939	-6.720
	0.063	0.028	-0.452	-0.015
	-0.939	-0.452	7.863	-0.399
	-6.720	-0.015	-0.399	7.510
	n	24		
	p	3		

Figure 21 - Calcul du résidu studentisé (Feuille "outliers")

La formule insérée en I2 est =H2*RACINE(($M$14-$M$16-2)/($M$14-$M$16-1-H2^2)). Elle fait référence à n et p, placés respectivement en M14 et M16.

Le seuil est défini par le quantile bilatéral de la loi de Student à (n-p-2) degrés de liberté. Au risque 5%, voici la formule sous Excel : =LOI.STUDENT.INVERSE.BILATERALE(0.05;M14-M16-2)

5.6.3 Distance de Cook

La distance de Cook permet de mesurer l'impact des observations sur les paramètres estimés de la régression. Un point est considéré comme influent si son retrait de la base de données induit une

modification substantielle des coefficients calculés. Elle s'appuie sur le résidu standardisé et le levier :

$$D_i = \frac{t_i^2}{p+1} \times \frac{1-h_i}{h_i}$$

La formule insérée en J2 est =(H2^2/($M$16+1))*(G2/(1-G2)) (Figure 22).

Le seuil en J27 est défini par :

$$D_i > \frac{4}{n-p-1} = \frac{4}{24-3-1} = 0.2$$

On retrouve (Figure 22) les deux mêmes marques de cigarettes mises en évidence par le résidu studentisé : MultiFilter et WinstonLights

	A	B	C	D	E	F	G	H	I	J
1	Cigarette	X				Résidus	Levier	Rés.Stand.	Res.Stud	D.Cook
2	Alpine	1	14.1	0.86	0.9853	-0.858	0.1707	-0.8122	-0.8050	0.0340
3	Benson&Hedges	1	16	1.06	1.0938	0.126	0.1849	0.1208	0.1177	0.0008
4	CamelLights	1	8	0.67	0.928	1.374	0.0857	1.2390	1.2568	0.0360
5	Carlton	1	4.1	0.4	0.9462	0.138	0.1669	0.1301	0.1269	0.0008
6	Chesterfield	1	15	1.04	0.8885	-0.149	0.1335	-0.1377	-0.1343	0.0007
7	GoldenLights	1	8.8	0.76	1.0267	-0.788	0.1572	-0.7400	-0.7313	0.0255
8	Kent	1	12.4	0.95	0.9225	-0.565	0.0979	-0.5129	-0.5033	0.0071
9	Kool	1	16.6	1.12	0.9372	-0.412	0.1196	-0.3782	-0.3700	0.0049
10	L&M	1	14.9	1.02	0.8858	0.356	0.1336	0.3298	0.3223	0.0042
11	LarkLights	1	13.7	1.01	0.9643	-1.137	0.0763	-1.0199	-1.0210	0.0215
12	Marlboro	1	15.1	0.9	0.9316	-0.855	0.2315	-0.8404	-0.8340	0.0532
13	Merit	1	7.8	0.57	0.9705	1.315	0.0948	1.1917	1.2051	0.0372
14	MultiFilter	1	11.4	0.78	1.124	-2.108	0.2599	-2.1130	-2.3368	0.3920
15	NewportLights	1	9	0.74	0.8517	-0.091	0.1554	-0.0855	-0.0834	0.0003
16	Now	1	1	0.13	0.7851	-0.536	0.4603	-0.6288	-0.6191	0.0843
17	OldGold	1	17	1.26	0.9186	1.399	0.2553	1.3982	1.4347	0.1675
18	PallMallLight	1	12.8	1.08	1.0395	-0.931	0.3157	-0.9701	-0.9686	0.1085
19	Raleigh	1	15.8	0.96	0.9573	1.540	0.1948	1.4794	1.5280	0.1324
20	SalemUltra	1	4.5	0.42	0.9106	-0.654	0.1525	-0.6122	-0.6023	0.0169
21	Tareyton	1	14.5	1.01	1.007	0.964	0.0687	0.8615	0.8557	0.0137
22	TrueLight	1	7.3	0.61	0.9806	0.217	0.0961	0.1969	0.1921	0.0010
23	ViceroyRichLight	1	8.6	0.69	0.9693	1.145	0.0695	1.0237	1.0250	0.0196
24	VirginiaSlims	1	15.2	1.02	0.9496	-1.543	0.0835	-1.3896	-1.4249	0.0440
25	WinstonLights	1	12	0.82	1.1184	2.050	0.2356	2.0216	2.2090	0.3148
26										
27						Seuil	0.3333		2.0930	0.2000

	L	M	N	O
	sigma^(epsilon)			
	1.1598			
	(X'X)^-1			
	6.563	0.063	-0.939	-6.720
	0.063	0.028	-0.452	-0.015
	-0.939	-0.452	7.863	-0.399
	-6.720	-0.015	-0.399	7.510
	n	24		
	p	3		

Figure 22 - Distance de Cook (Feuille "outliers")

5.7 Prédiction ponctuelle et par intervalle

La prédiction est une des principales finalités de la régression. On s'appuie sur le modèle pour prédire (deviner) la valeur prise par la variable expliquée d'un individu supplémentaire dont on connaît la description c.-à-d. les valeurs des explicatives (x_j).

Mettons que l'on dispose d'une cigarette de marque « Mélia Bleue », avec les caractéristiques suivantes : TAR = 11.5 ; NICOTINE = 0.8 ; WEIGHT = 0.95. Quelle serait sa teneur en CO ?

Nous créons une feuille « prediction » pour cette nouvelle analyse (Figure 23). Nous y copions (collage spécial « valeurs ») le vecteur des coefficients â (A6:A9), la matrice $(X'X)^{-1}$ (C6:F9) et la variance estimée de l'erreur de la régression $\hat{\sigma}_\varepsilon^2$ (H6) que l'on aura récupéré dans la feuille « matrices ». Nous y portons la description de « Mélia Bleue » (B2:E2), sans oublier d'insérer la valeur 1 en rapport avec la constante $\hat{a}_0$.

	A	B	C	D	E	F	G	H	I	J
1	**Marque**	constante	TAR	NICOTINE	WEIGHT	PREDICTION	Var.Err.Pred	Quantile.t	B.Basse	B.Haute
2	Mélia Bleue	1	11.5	0.8	0.95					
3										
4										
5	**a^**			**(X'X)^-1**				**sigma²(espilon)**		
6	-0.5517		6.5630	0.0629	-0.9391	-6.7199		1.3452		
7	0.8876		0.0629	0.0284	-0.4520	-0.0153				
8	0.5185		-0.9391	-0.4520	7.8633	-0.3990				
9	2.0793		-6.7199	-0.0153	-0.3990	7.5099				

Figure 23 - Préparation de la feuille "prediction"

5.7.1 Prédiction ponctuelle

La description de l'individu à traiter est représenté par un vecteur :

$$x_* = (1\,;\,11.5\,;\,0.8\,;\,0.95)$$

La prédiction ponctuelle est un produit scalaire entre ce vecteur et l'estimation des paramètres de la régression â :

$$\hat{y}_* = x_*\hat{a} = (1 \quad 11.5 \quad 0.8 \quad 0.95)\begin{pmatrix} -0.5517 \\ 0.8876 \\ 0.5185 \\ 2.0793 \end{pmatrix} = 12.0456$$

En F2, nous insérons {=PRODUITMAT(B2:E2;A6:A9)}. Malgré que le résultat soit un scalaire, il s'agit bien d'une opération matricielle (Figure 24).

F2 : {=PRODUITMAT(B2:E2;A6:A9)}

	A	B	C	D	E	F	G	H	I	J
1	**Marque**	constante	TAR	NICOTINE	WEIGHT	PREDICTION	Var.Err.Pred	Quantile.t	B.Basse	B.Haute
2	Mélia Bleue	1	11.5	0.8	0.95	12.0456				
3										
4										
5	**a^**			**(X'X)^-1**				**sigma²(espilon)**		
6	-0.5517		6.5630	0.0629	-0.9391	-6.7199		1.3452		
7	0.8876		0.0629	0.0284	-0.4520	-0.0153				
8	0.5185		-0.9391	-0.4520	7.8633	-0.3990				
9	2.0793		-6.7199	-0.0153	-0.3990	7.5099				

Figure 24 - Prédiction ponctuelle (Feuille "prediction")

Remarque : Nous aurions pu également réaliser la prédiction en introduisant les coefficients dans l'équation de régression,

$$\hat{y}_* = -0.5517 + 0.8876 \times 11.5 + 0.5185 \times 0.8 + 2.0793 \times 0.95 = 12.0456$$

5.7.2 Variance de l'erreur de prédiction et intervalle de prédiction

Pour obtenir la fourchette de prédiction, nous devons tout d'abord calculer la variance de l'erreur de prédiction :

$$\hat{\sigma}^2_{\hat{\varepsilon}_*} = \hat{\sigma}^2_{\varepsilon}[1 + x_*(X'X)^{-1}x'_*]$$

Remarque : Notez bien la nuance selon qu'il y ait un ^ ou non sur ε.

Une lecture rapide de la formule montre que cette variance sera d'autant plus faible, et donc la prédiction d'autant plus précise, que : (1) la régression est de bonne qualité c.-à-d. la variance de l'erreur ($\hat{\sigma}^2_{\varepsilon}$) de la régression est faible ; (2) l'observation à traiter est proche du barycentre du nuage de points c.-à-d. le levier $x_*(X'X)^{-1}x'_*$ est faible.

En G2, nous insérons : =H6*(1+PRODUITMAT(PRODUITMAT(B2:E2;C6:F9);TRANSPOSE(B2:E2))) (Figure 25).

G2 : {=H6*(1+PRODUITMAT(PRODUITMAT(B2:E2;C6:F9);TRANSPOSE(B2:E2)))}

	A	B	C	D	E	F	G	H	I	J
1	**Marque**	constante	TAR	NICOTINE	WEIGHT	PREDICTION	Var.Err.Pred	Quantile.t	B.Basse	B.Haute
2	Mélia Bleue	1	11.5	0.8	0.95	12.0456	1.4115			
3										
4										
5	**a^**			**(X'X)^-1**				**sigma²(espilon)**		
6	-0.5517		6.5630	0.0629	-0.9391	-6.7199		1.3452		
7	0.8876		0.0629	0.0284	-0.4520	-0.0153				
8	0.5185		-0.9391	-0.4520	7.8633	-0.3990				
9	2.0793		-6.7199	-0.0153	-0.3990	7.5099				

Figure 25 - Variance de l'erreur de prédiction (Feuille "prediction")

Les bornes de la fourchette de prédiction s'écrivent :

$$\hat{y}_* \pm t \times \hat{\sigma}_{\hat{\varepsilon}_*}$$

Où est t le quantile de la loi de Student à $(n - p - 1 = 24 - 3 - 1 = 20)$ degrés de liberté. Pour un niveau de confiance à 95%, nous plaçons =LOI.STUDENT.INVERSE.BILATERALE(0.05;20) en H6.

Il ne nous reste plus qu'à former les bornes (Figure 26) :

- =F2-H2*RACINE(G2) en I6 ;
- =F2+H2*RACINE(G2) en J6.

	A	B	C	D	E	F	G	H	I	J
1	**Marque**	constante	TAR	NICOTINE	WEIGHT	PREDICTION	Var.Err.Pred	Quantile.t	B.Basse	B.Haute
2	Mélia Bleue	1	11.5	0.8	0.95	12.0456	1.4115	2.0860	9.5674	14.5239
3										
4										
5	**a^**			**(X'X)^-1**				**sigma²(espilon)**		
6	-0.5517		6.5630	0.0629	-0.9391	-6.7199		1.3452		
7	0.8876		0.0629	0.0284	-0.4520	-0.0153				
8	0.5185		-0.9391	-0.4520	7.8633	-0.3990				
9	2.0793		-6.7199	-0.0153	-0.3990	7.5099				

Figure 26 - Calcul des bornes de l'intervalle de prédiction (Feuille "prediction")

Il y a 95% de chances pour la fourchette [9.5674 ; 14.5239] couvre la vraie valeur de la teneur en CO de la cigarette « Mélia Bleue ».

6 Conclusion

Travailler sous Excel permet d'évacuer les situations où les étudiants cliquent ou lancent des commandes au petit bonheur la chance avec le fatidique « ça marche ! » (ou « ça marche pas » !) à la sortie, sans vraiment s'intéresser aux paramètres utilisés, sans véritable réflexion concernant la teneur des résultats. La démarche est pédagogiquement intéressante car elle oblige à regarder dans le détail les calculs pour chaque méthode étudiée. Pour moi, la page Excel'ense de la Revue Modulad, consacrée à l'enseignement de la statistique avec un tableur, est une référence incontournable de ce point de vue (https://www.rocq.inria.fr/axis/modulad/excel.htm).

Mais la démarche a ses limites. Rentrer à la main des formules sous Excel peut se révéler fastidieux. Cela prend du temps surtout, nous ralentissant dans notre progression. Manipuler des données un peu plus volumineuses est vite rédhibitoire (par ex. imaginons la manipulation de la matrice X'X lorsqu'on a 30 variables explicatives). A un certain stade, certaines formules sont particulièrement complexes, on peut se demander même si c'est pédagogiquement intéressant. C'est pour cette raison qu'après une première phase introductive où j'utilise Excel, je les fais passer sur d'autres outils statistiques spécialisés, en espérant que les concepts et les schémas ont été entre temps bien assimilés.

7 Références

R. Rakotomalala, « Econométrie – Régression linéaire simple et multiple », Fascicule de cours, Version 1.1, Janvier 2018.

R. Rakotomalala, « Pratique de la Régression Linéaire Multiple – Diagnostic et sélection de variables », Version 2.1, Mai 2015.