Master 1 SES
Enquêtes et méthodes d'analyse quantitative
C. Hardouin, A.-K. Fermin

Régression linéaire multiple

On observe $p+1$ variables Y, X_1, X_2, ..X_p et on cherche à "expliquer" Y par X_1, X_2.. et X_p. On propose le modèle linéaire (théorique)

$$\begin{aligned} Y &= a_0 + a_1 X_1 + a_2 X_2 + ... + a_p X_p + \varepsilon \\ &= \mathbf{X}A + \varepsilon \end{aligned}$$

$$\text{avec } Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & & x_{p,1} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{1,n} & & x_{p,n} \end{pmatrix}, A = \begin{pmatrix} a_0 \\ \vdots \\ a_p \end{pmatrix} \text{ et } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Hypothèses : ε est indépendante de $X_1, ...X_p$ et suit une loi normale centrée $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I)$.

Méthode des MCO :

$$\hat{A} = \left(\mathbf{X}^{\mathrm{T}}\ \mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}} Y$$

La variable estimée est :

$$\begin{aligned} \hat{Y} &= \hat{a}_0 + \hat{a}_1 X_1 + \hat{a}_2 X_2 + .. + \hat{a}_p X_p \\ &= \mathbf{X}\hat{A} = HY \end{aligned}$$

avec $H = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\ \mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}}$.

Le résidu de la régression est :

$$e = Y - \hat{Y}$$

1 Propriétés des résidus

1. Ils sont centrés

2. On définit les **résidus standardisés** par

$$r_i = \frac{e_i}{s_e^*}$$

$s_e^* = s_e\sqrt{1 - h_{ii}}$
Problème : le numérateur et le dénominateur ne sont pas indépendants.

Les **résidus studentisés** sont les résidus normés de façon à ce que le $i$ème résidu n'intervienne pas dans la variance s_e :

$$r_i' = \frac{e_i}{s_{(i)}\sqrt{1 - h_{ii}}} \sim T_{n-p-2}$$

3. Ils ne sont pas liés linéairement aux régresseurs ni à la variable estimée $\hat{Y}$; alors le graphe des résidus en fonction de $\hat{Y}$ ne doit pas avoir de forme particulière.

2 Equation fondamentale d'analyse de la variance :

$$SCT = SCE + SCR$$

Coefficient de détermination R^2.

$$R^2 = \frac{SCE}{SCT}$$

On appelle $R = \sqrt{R^2}$ le coefficient de corrélation multiple entre Y et les variables explicatives, c'est le coefficient de corrélation entre Y et $\hat{Y}$.

Coefficient R^2 ajusté ou corrigé :

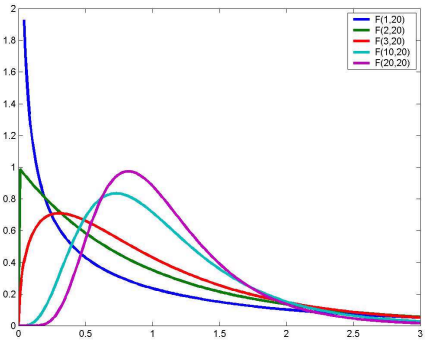
$$\bar{R}^2 = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

3 Tests sur la régression

Test de significativité globale de la régression (test de Fisher)

$(H_0) : a_1 = a_2 = ... = a_p = 0$ $\quad$ (H_1) : l'un au moins des coefficients est non nul,
à un seuil α fixé.
Sous (H_0), F^* suit approximativement la loi de Fisher à p et $n-p-1$ degrés de liberté.

$$F^* = \frac{SCE/p}{SCR/(n-p-1)} \sim F(p; n-p-1)$$



Tests sur les coefficients (tests de Student)
Pour chaque coefficient, le test est le suivant :
$(H_0) : a_i = 0$ $\quad$ $(H_1) : a_i \neq 0$ au seuil de signification α.

Sous (H_0), $\dfrac{\hat{a}_i}{s_{\hat{a}_i}} \sim T_{n-p-1}$

4 Validation

- On regarde le test sur le F et les tests sur les coefficients.

- On regarde la valeur du R^2 (ou du $\bar{R}^2$ si on compare plusieurs modèles).

- Examen des résidus :
Le nuage de points des résidus en fonction de $\hat{Y}$ ne doit pas avoir de forme particulière.
Les résidus studentisés doivent être compris entre -2 et +2
Normalité des résidus et autres tests..... Voir en M2

- On peut vérifier la matrice de corrélation des variables explicatives.

5 Variable indicatrice

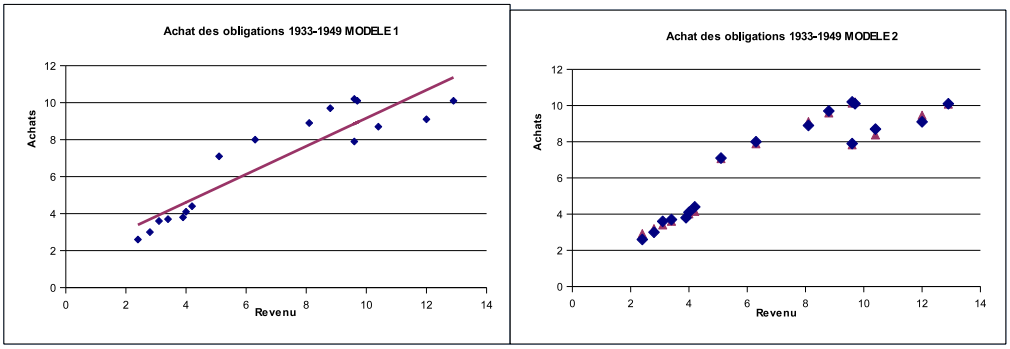
$$D = \begin{cases} 1 \text{ si le phénomène existe} \\ 0 \text{ sinon} \end{cases}$$

1. Prise en compte des effets individuels

2. Prise en compte d'effets temporels

   Exemple : On considère les deux variables suivantes : R est le revenu national, en billions de francs, et B est l'achat public d'obligations d'état, en centaines de millions. Les données couvrent les années 1933 à 1949.

   On considère le modèle : $B_t = a_0 + a_1 R_t + \varepsilon_t$

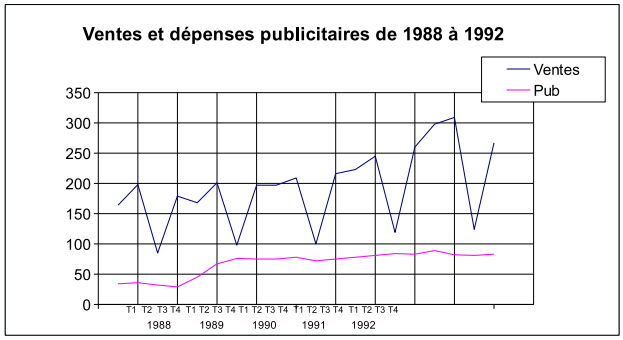


   On introduit la variable D qui vaut 1 pour les années de guerre (1940 à 1945 inclus), 0 sinon.

   Le modèle estimé est $B_t = 1.29 + 0.68\ R_t + 2.30\ D_t + e_t$

3. Prise en compte des effets saisonniers

Exemple : On s'intéresse au montant des ventes V en fonction de la publicité Pub.



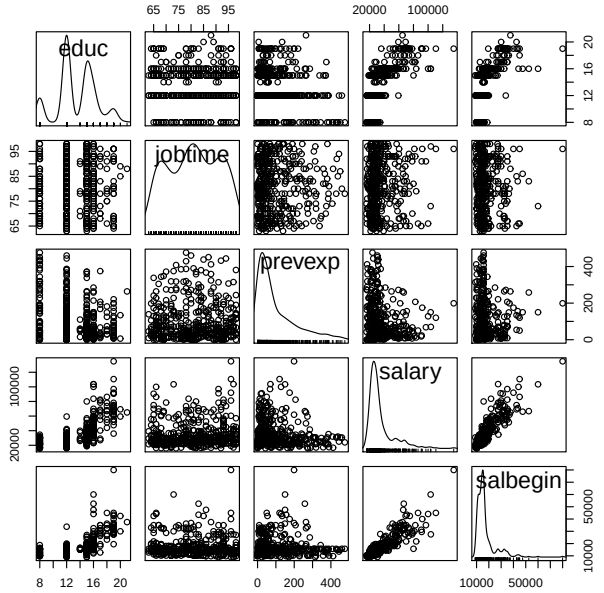
6 Exercices et exemples

Exemple 1 *Salary*

Les données sont observées sur un échantillon de 474 employés tirés au sort dans une entreprise canadienne. Les variables étudiées ici sont les suivantes :

— salary (salaire brut actuel en $ par an)
— salbegin (salaire de départ en $ par an)
— jobtime(nombre de mois depuis l'entrée dans l'entreprise)
— prevexp (nombre de mois de travail avant l'entrée dans l'entreprise)
— educ (nombre d'années d'étude)
— sex (sexe à deux modalités H = Homme et F = Femme)
— minority (appartenance à une minorité (Non, Oui))

On souhaite expliquer la variable salary en fonction de toutes les autres variables.



1.

```
> cor(Salaires[,c("educ","jobtime","prevexp","salary","salbegin")],
+   use="complete.obs")
                educ       jobtime      prevexp      salary    salbegin
educ       1.00000000  0.047378777 -0.252352521  0.66055891  0.63319565
jobtime    0.04737878  1.000000000  0.002978134  0.08409227 -0.01975347
prevexp   -0.25235252  0.002978134  1.000000000 -0.09746693  0.04513563
salary     0.66055891  0.084092267 -0.097466926  1.00000000  0.88011747
salbegin   0.63319565 -0.019753475  0.045135627  0.88011747  1.00000000
```

(a) Indiquer pour quels couples de variables la corrélation linéaire observée est la plus forte, la plus faible. Que peut-on dire de la corrélation linéaire entre le salaire de départ et le salaire actuel ?

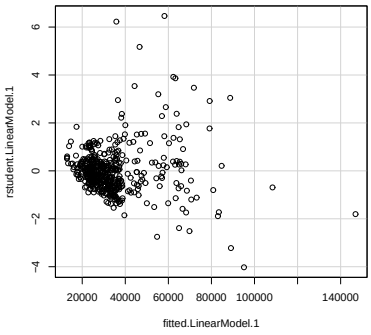
(b) Pourquoi n'y a-t-il pas les variables sex et minority dans la matrice de corrélations ?

2. On ajuste le modèle expliquant salary en fonction de toutes les autres variables. Commentez les résultats.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.206e+04  3.482e+03  -3.463 0.000582 ***
educ         5.893e+02  1.664e+02   3.542 0.000437 ***
jobtime      1.565e+02  3.405e+01   4.597 5.53e-06 ***
minority    -1.377e+03  8.513e+02  -1.618 0.106317
prevexp     -1.876e+01  3.601e+00  -5.210 2.83e-07 ***
salbegin     1.706e+00  6.124e-02  27.868  < 2e-16 ***
sex         -2.419e+03  7.990e+02  -3.027 0.002605 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7398 on 467 degrees of freedom
Multiple R-squared: 0.8147, Adjusted R-squared: 0.8123
F-statistic: 342.2 on 6 and 467 DF,  p-value: < 2.2e-16
```

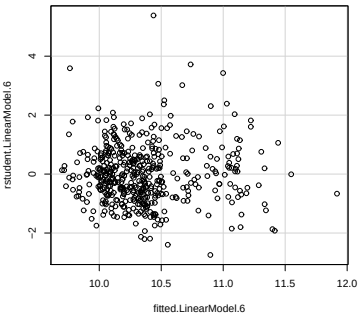
(a)



3. On applique une transformation logarithmique aux variables salary et salbegin et on ajuste le modèle de régression linéaire multiple en remplaçant ces variables par les variables transformées. Commentez les résultats.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.226e+00  3.204e-01   3.827 0.000147 ***
educ         1.091e-02  3.909e-03   2.792 0.005459 **
jobtime      4.541e-03  7.570e-04   5.999 3.98e-09 ***
logbegin     8.996e-01  3.457e-02  26.024  < 2e-16 ***
minority    -2.881e-02  1.901e-02  -1.516 0.130276
prevexp     -5.373e-04  7.986e-05  -6.729 5.03e-11 ***
sex         -5.509e-02  1.873e-02  -2.941 0.003430 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1636 on 467 degrees of freedom
Multiple R-squared: 0.8325, Adjusted R-squared: 0.8304
F-statistic: 386.9 on 6 and 467 DF,  p-value: < 2.2e-16
```



Exemple 2 *Taux de crimes à Detroit*

On dispose des données suivantes sur le nombre d'homicides dans la ville de Détroit sur une période de dix ans, de 1961 à 1970.

HOM - Number of homicides per 100,000 of population
FTP - Full-time police per 100,000 population
UEMP - % unemployed in the population
CLEAR - % homicides cleared by arrests
LIC - Number of handgun licences per 100,000 population
HE - Average hourly earnings

La matrice de corrélation est donnée ci-dessous.

Coefficients de corrélation de Pearson, N = 10 Proba > |r| sous H0: Rho=0						
	FTP	UEMP	LIC	CLEAR	HE	HOM
FTP	1.00000	-0.05832 0.8728	0.68957 0.0274	-0.91223 0.0002	0.73297 0.0159	0.89751 0.0004
UEMP	-0.05832 0.8728	1.00000	-0.35091 0.3201	0.10581 0.7711	-0.22497 0.5320	-0.20667 0.5667
LIC	0.68957 0.0274	-0.35091 0.3201	1.00000	-0.69464 0.0258	0.39521 0.2583	0.89856 0.0004
CLEAR	-0.91223 0.0002	0.10581 0.7711	-0.69464 0.0258	1.00000	-0.81675 0.0039	-0.91963 0.0002
HE	0.73297 0.0159	-0.22497 0.5320	0.39521 0.2583	-0.81675 0.0039	1.00000	0.71600 0.0199
HOM	0.89751 0.0004	-0.20667 0.5667	0.89856 0.0004	-0.91963 0.0002	0.71600 0.0199	1.00000

1. On propose le modèle de Hom sur les autres variables. Commentez les résultats.

Analyse de variance					
Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	5	1046.45117	209.29023	244.31	<.0001
Erreur	4	3.42659	0.85665		
Total sommes corrigées	9	1049.87776			

Root MSE	0.92555	R carré	0.9967
Moyenne dépendante	18.08200	R car. ajust.	0.9927
Coeff Var	5.11864		

Résultats estimés des paramètres						
Variable	Libellé	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	Intercept	1	-6.00007	22.20631	-0.27	0.8004
FTP	FTP	1	0.05464	0.02904	1.88	0.1331
UEMP	UEMP	1	0.42947	0.16148	2.66	0.0564
LIC	LIC	1	0.01967	0.00171	11.50	0.0003
CLEAR	CLEAR	1	-0.25648	0.15821	-1.62	0.1803
HE	HE	1	5.46077	1.29444	4.22	0.0135

2. On retire la variable FTP du modèle. Justifiez ce choix.

Analyse de variance					
Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	4	1043.41910	260.85478	201.94	<.0001
Erreur	5	6.45866	1.29173		
Total sommes corrigées	9	1049.87776			

Root MSE	1.13654	R carré	0.9938
Moyenne dépendante	18.08200	R car. ajust.	0.9889
Coeff Var	6.28550		

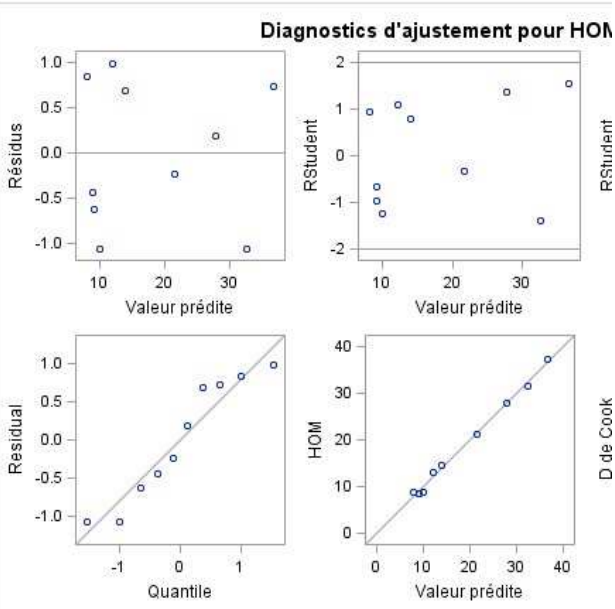
Résultats estimés des paramètres						
Variable	Libellé	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	Intercept	1	21.19073	20.70258	1.02	0.3530
UEMP	UEMP	1	0.49863	0.19308	2.58	0.0493
LIC	LIC	1	0.02057	0.00202	10.20	0.0002
CLEAR	CLEAR	1	-0.41306	0.16522	-2.50	0.0545
HE	HE	1	5.79815	1.57419	3.68	0.0142

3. On fait la même régression mais sans constante.

Analyse de variance					
Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	4	1044.19990	261.04998	229.88	<.0001
Erreur	5	5.67786	1.13557		
Total sommes corrigées	9	1049.87776			

Root MSE	1.06563	R carré	0.9946
Moyenne dépendante	18.08200	R car. ajust.	0.9903
Coeff Var	5.89333		

Résultats estimés des paramètres						
Variable	Libellé	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	Intercept	1	-41.33329	4.89411	-8.45	0.0004
FTP	FTP	1	0.07941	0.02844	2.79	0.0383
UEMP	UEMP	1	0.50756	0.17745	2.86	0.0354
LIC	LIC	1	0.02104	0.00171	12.28	<.0001
HE	HE	1	6.83428	1.12674	6.07	0.0018

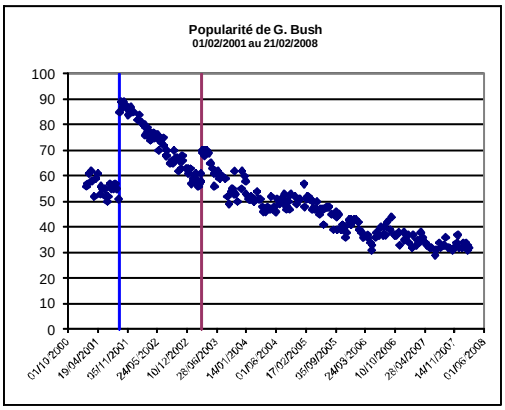


Exemple 3 *Popularité de G. Bush*

Les données sont composées des pourcentages de personnes satisfaites de l'action de Georges Bush lors de sondages réalisés entre le 1er février 2002 et le 24 février 2008.

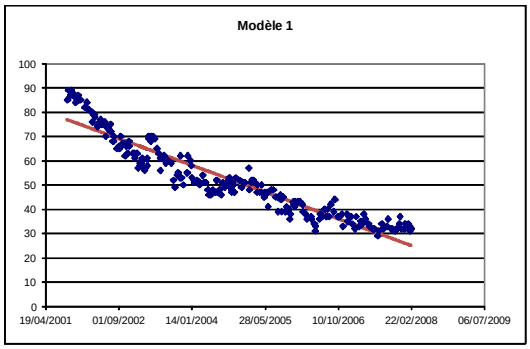
Plus précisément, la question posée est : "Do you approve or disapprove of the way president George Bush is handling his job as President?" Trois réponses possibles sont proposées : "Approving", "Disapproving" et "unsure" ; le refus de réponse a été regroupé avec la modalité "unsure".

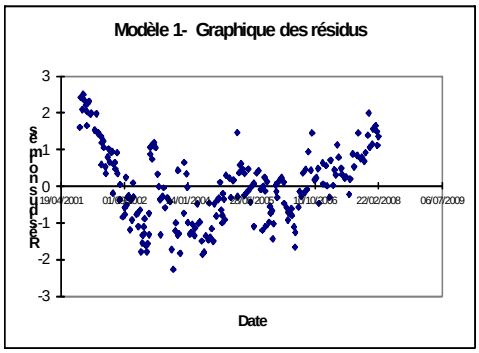
Le graphique suivant représente la cote de popularité de G. Bush au cours du temps, cette cote étant le pourcentage de réponse "Approving". Les deux lignes verticales étant tracées aux dates du 11 septembre 2001 et 20 mars 2003 (date de l'intervention en Irak).



On s'intéresse à la modélisation des données à partir du 11 septembre 2001. On pose Y la cote de popularité de G. Bush et X la date.

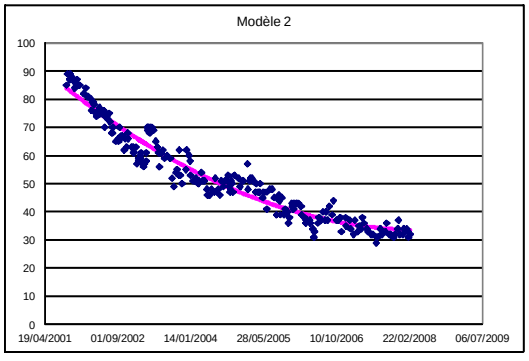
1. Modèle 1 : $Y = a_0 + a_1 X + \varepsilon$ $\quad R^2 = 0,8928$ $\quad \bar{R}^2 = 0,8923$

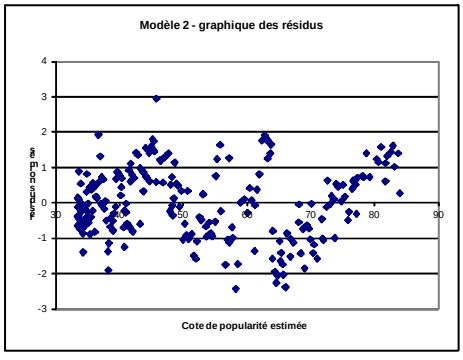




ANALYSE DE VARIANCE					
	Degré de libert	*mme des car*	*yenne des car*	*F*	*aleur critique de*
Régression	1	48852,2613	48852,2613	1856,75089	4,2602E-110
Résidus	223	5867,26756	26,3106169		
Total	224	54719,5289			
		Coefficients	*Erreur-type*	*Statistique t*	*Probabilité*
	Constante	893,61418	19,5174023	45,7855081	2,2257E-115
	Variable X 1	-0,02198909	0,00051031	-43,0900324	4,2602E-110

2. Modèle 2 : $Y = a_0 + a_1 X + a_2 X^2 + \varepsilon$ $\quad R^2 = 0,9411$ $\quad \bar{R}^2 = 0,9406$

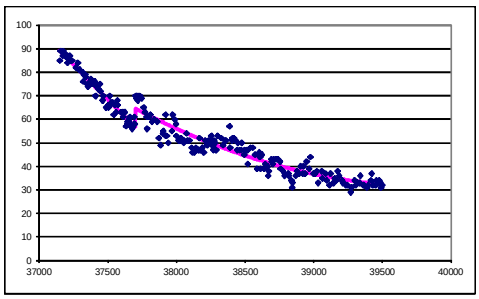


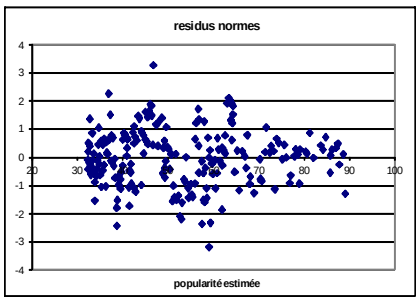


ANALYSE DE VARIANCE					
	Degré de libert	*mme des car*	*yenne des car*	*F*	*eur critique de*
Régression	2	51496,585	25748,2925	1773,57137	3,036E-137
Résidus	222	3222,94385	14,5177651		
Total	224	54719,5289			

	Coefficients	*Erreur-type*	*Statistique t*	*Probabilité*
Constante	13161,5381	909,115118	14,4773064	6,8404E-34
Variable X 1	8,3683E-06	6,2005E-07	13,4960732	1,0439E-30
Variable X 2	-0,66290497	0,04749058	-13,9586633	3,3061E-32

3. Modèle 3 : on introduit une variable indicatrice D qui vaut 1 à partir du 20 mars 2003, 0 sinon.
$Y = a_0 + a_1 X + a_2 D + a_3 DX + a_4 DX^2 + \varepsilon \qquad R^2 = 0,9596 \qquad \bar{R}^2 = 0,9588$

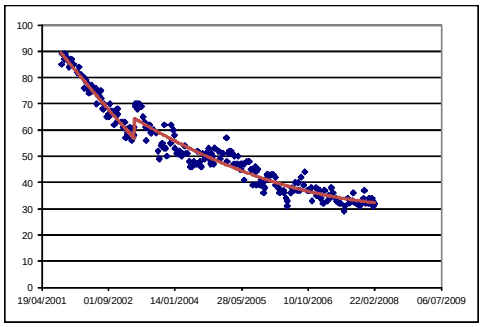


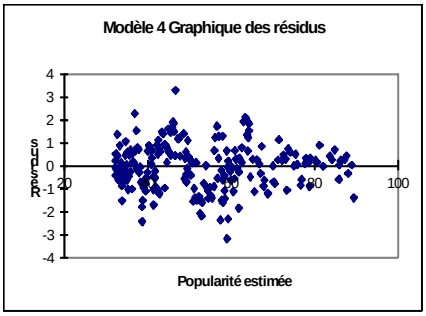


ANALYSE DE VARIANCE					
	Degré de libert	*mme des car*	*yenne des car*	*F*	*eur critique de*
Régression	4	52506,8987	13126,7247	1305,17944	5,914E-152
Résidus	220	2212,63018	10,0574099		
Total	224	54719,5289			

	Coefficients	*Erreur-type*	*Statistique t*	*Probabilité*
Constante	2289,40008	92,3707268	24,7849093	1,3108E-65
Indic *X^2	7,0858E-06	9,8145E-07	7,2196889	8,3822E-12
D	8998,9062	1463,56952	6,14860183	3,6372E-09
X	-0,05923234	0,00246731	-24,006869	2,2181E-63
Indic * X	-0,50561828	0,07577001	-6,67306639	2,0056E-10

4. Modèle 4 : $Y = a_0 + a_1 X + a_2 X^2 + a_3 D + a_4 DX + a_5 DX^2 + \varepsilon \qquad R^2 = 0,9596 \qquad \bar{R}^2 = 0,9587$



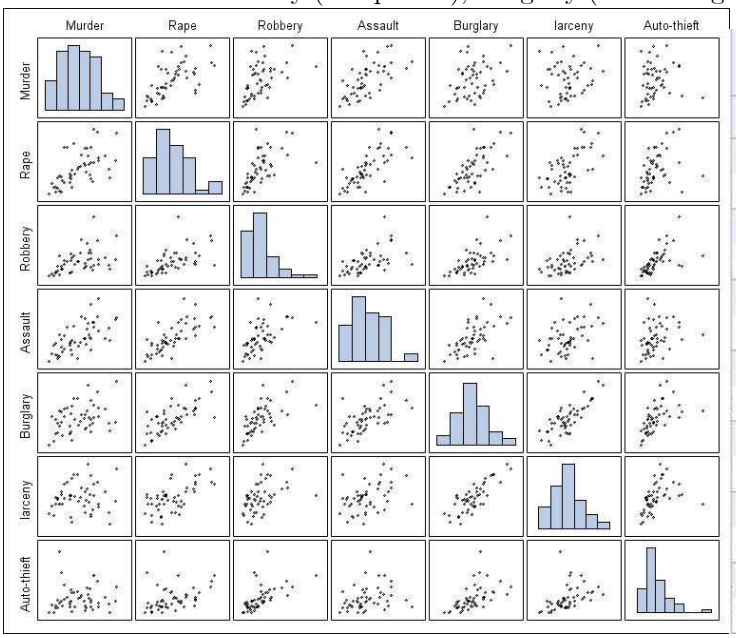


ANALYSE DE VARIANCE					
	Degré de libert	*mme des car*	*yenne des car*	*F*	*eur critique de*
Régression	5	52508,4304	10501,6861	1040,1478	2,11E-150
Résidus	219	2211,0985	10,0963402		
Total	224	54719,5289			

	Coefficients	*Erreur-type*	*Statistique t*	*Probabilité*
Constante	11335,2553	23224,7494	0,48806793	0,62599002
Indic	-46,9489944	23270,8132	-0,00201751	0,9983921
Date	-0,54260714	1,24103155	-0,43722268	0,6623807
Indic*date	-0,02224348	1,24334891	-0,01788997	0,98574292
Date^2	6,4573E-06	1,6579E-05	0,38949514	0,6972884
Indic*Date^2	6,2846E-07	1,6608E-05	0,03784144	0,96984858

Exemple 4 *Crimes US*

Les données sont les taux de crimes pour 100.000 personnes dans sept catégories et pour chacun des 50 états des US en 1977. Les variables sont les suivantes : Murder (meurtre), Rape (viol), Assault (agression), Robbery (vol qualifié), Burglary (cambriolage), Larceny (larcin) et Auto-thiefth (vol de voiture).



Coefficients de corrélation de Pearson, N = 50
Proba > |r| sous H0: Rho=0

	Murder	Rape	Robbery	Assault	Burglary	larceny	Auto_thieft
Murder	1.00000	0.60122	0.48371	0.64855	0.38582	0.10192	0.06881
Murder		<.0001	0.0004	<.0001	0.0057	0.4813	0.6349
Rape	0.60122	1.00000	0.59188	0.74026	0.71213	0.61399	0.34890
Rape	<.0001		<.0001	<.0001	<.0001	<.0001	0.0130
Robbery	0.48371	0.59188	1.00000	0.55708	0.63724	0.44674	0.59068
Robbery	0.0004	<.0001		<.0001	<.0001	0.0011	<.0001
Assault	0.64855	0.74026	0.55708	1.00000	0.62291	0.40436	0.27584
Assault	<.0001	<.0001	<.0001		<.0001	0.0036	0.0525
Burglary	0.38582	0.71213	0.63724	0.62291	1.00000	0.79212	0.55795
Burglary	0.0057	<.0001	<.0001	<.0001		<.0001	<.0001
larceny	0.10192	0.61399	0.44674	0.40436	0.79212	1.00000	0.44418
larceny	0.4813	<.0001	0.0011	0.0036	<.0001		0.0012
Auto_thieft	0.06881	0.34890	0.59068	0.27584	0.55795	0.44418	1.00000
Auto-thieft	0.6349	0.0130	<.0001	0.0525	<.0001	0.0012	

1. On fait la régression de la variable meurtre sur toutes les variables.

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	6	446.18724	74.36454	11.16	<.0001
Erreur	43	286.45596	6.66177		
Total sommes corrigées	49	732.64320			

Root MSE	2.58104	R carré	0.6090
Moyenne dépendante	7.44400	R car. ajust.	0.5545
Coeff Var	34.67275		

Résultats estimés des paramètres

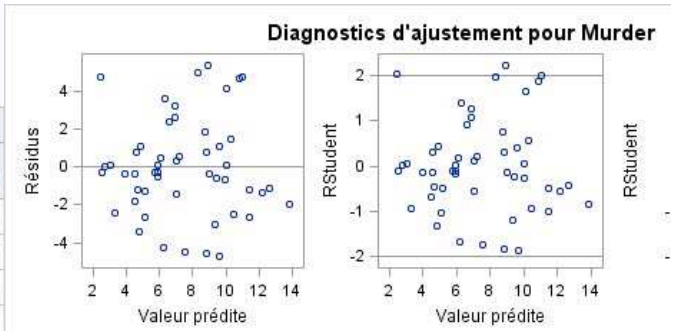
Variable	Libellé	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	Intercept	1	5.57034	1.47372	3.78	0.0005
Rape	Rape	1	0.15974	0.06127	2.61	0.0125
Robbery	Robbery	1	0.01177	0.00631	1.87	0.0688
Assault	Assault	1	0.01082	0.00595	1.82	0.0759
Burglary	Burglary	1	0.00225	0.00185	1.21	0.2316
larceny	larceny	1	-0.00264	0.00088513	-2.98	0.0047
Auto_thieft	Auto-thieft	1	-0.00485	0.00255	-1.90	0.0644

2. On enlève petit à petit les variables les moins explicatives : Burglary, puis Auto_thieft, puis Robbery. On obtient alors le résultat suivant. Commentez.

Analyse de variance					
Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	3	402.96000	134.32000	18.74	<.0001
Erreur	46	329.68320	7.16703		
Total sommes corrigées	49	732.64320			

Root MSE	2.67713	R carré	0.5500
Moyenne dépendante	7.44400	R car. ajust.	0.5207
Coeff Var	35.96360		

Résultats estimés des paramètres						
Variable	Libellé	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	Intercept	1	4.74225	1.48822	3.19	0.0026
Rape	Rape	1	0.19542	0.06154	3.18	0.0027
Assault	Assault	1	0.01567	0.00570	2.75	0.0085
larceny	larceny	1	-0.00211	0.00067048	-3.15	0.0029



3. On décide faire un second modèle en se basant sur les corrélations les plus élevées entre Meurtre et les autres variables. Commentez les résultats.

Analyse de variance					
Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	4	359.26154	89.81539	10.82	<.0001
Erreur	45	373.38166	8.29737		
Total sommes corrigées	49	732.64320			

Root MSE	2.88052	R carré	0.4904
Moyenne dépendante	7.44400	R car. ajust.	0.4451
Coeff Var	38.69580		

Résultats estimés des paramètres						
Variable	Libellé	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	Intercept	1	2.60179	1.34630	1.93	0.0596
Rape	Rape	1	0.11988	0.06536	1.83	0.0732
Robbery	Robbery	1	0.00879	0.00633	1.39	0.1717
Assault	Assault	1	0.01746	0.00631	2.77	0.0082
Burglary	Burglary	1	-0.00234	0.00148	-1.58	0.1206