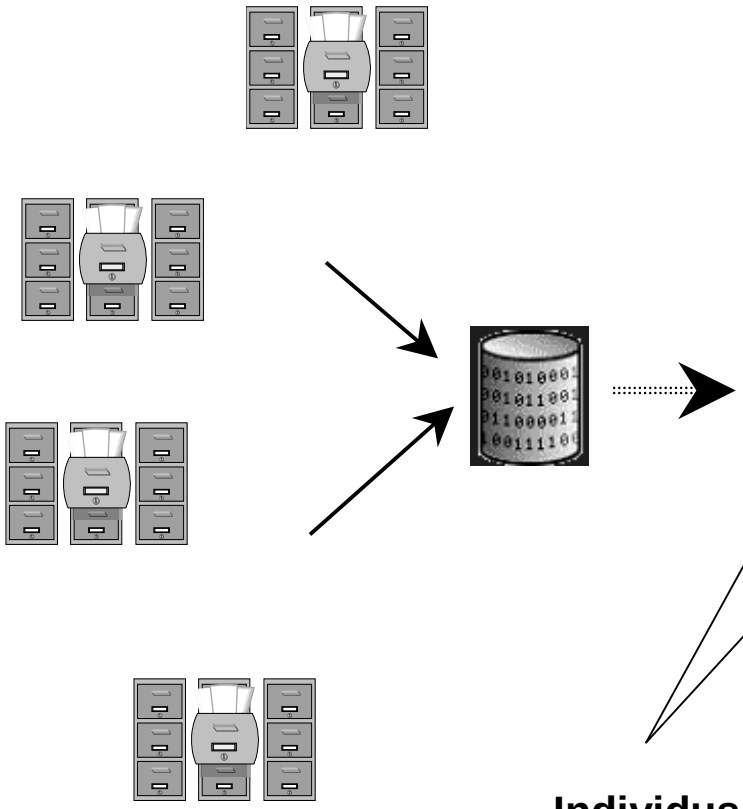
Régression linéaire multiple

Prédire les valeurs d'une variable continue

Ricco Rakotomalala
Ricco.Rakotomalala@univ-lyon2.fr

Tableau de données



Variables, caractères, attributs, Descripteurs, champs, etc.

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

Individus, observations, objets, enregistrements, etc.

Statut des variables

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

Identifiant
(Pas utilisé pour les calculs, mais peut être utilisé pour les commentaires : points atypiques, etc.)

Variables prédictives
Descripteurs
Variables exogènes

Quantitative ou qualitative

Variable à prédire
Attribut classe
Variable endogène

Quantitative

Principes de la régression multiple

Population Ω

$$\begin{cases} Y & \text{variable à prédire (endogène), quantitative} \\ X & \text{variables exogènes (quelconques)} \end{cases}$$

Objet de l 'étude

Une série de variables
$X=(x_1|\ldots|x_p)$

On veut construire une fonction de prédiction (explication) telle que

$$Y = f(X, \alpha)$$

Objectif de l 'apprentissage

Utiliser un échantillon Ω_a (extraite de la population) pour choisir la fonction f et ses paramètres α telle que l 'on minimise la somme des carrés des erreurs

$$S = \sum_{\Omega} [Y - \hat{f}(X, \hat{\alpha})]^2$$

Problèmes :
- ☞ il faut choisir une famille de fonction
- ☞ il faut estimer les paramètres α
- ☞ on utilise un échantillon pour optimiser sur la population

Régression linéaire multiple

- Se restreindre à une famille de fonction de prédiction linéaire
- Et à des exogènes continues (éventuellement des qualitatives recodées)

$$y_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \cdots + a_p x_{i,p} + \varepsilon_i \ ; \ i = 1, \ldots, n$$

Le terme aléatoire ε cristallise toutes les « insuffisances » du modèle :

- le modèle n 'est qu'une caricature de la réalité, la spécification (linéaire notamment) n 'est pas toujours rigoureusement exacte
- les variables qui ne sont pas prises en compte dans le modèle
- les fluctuations liées à l 'échantillonnage (si on change d 'échantillon, on peut obtenir un résultat différent)

ε quantifie les écarts entre les valeurs réellement observées et les valeurs prédites par le modèle

$(a_0, a_1, \ldots, a_p)$ Sont les paramètres du modèle que 'l'on veut estimer à l'aide des données

Régression linéaire multiple

Écriture matricielle

Pour une meilleure concision …

$$\begin{pmatrix} y_1 \\ \\ y_i \\ \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & & x_{1p} \\ 1 & & & \\ 1 & x_{i1} & x_{ij} & x_{ip} \\ 1 & & & \\ 1 & x_{n1} & & x_{np} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \\ \\ a_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \\ \varepsilon_i \\ \\ \varepsilon_n \end{pmatrix}$$

N.B. Noter la colonne représentant la constante

$$Y = Xa + \varepsilon$$

$$(n,1) = (n, p+1) \times (p+1,1) + (n,1)$$

Bien noter les dimensions des matrices

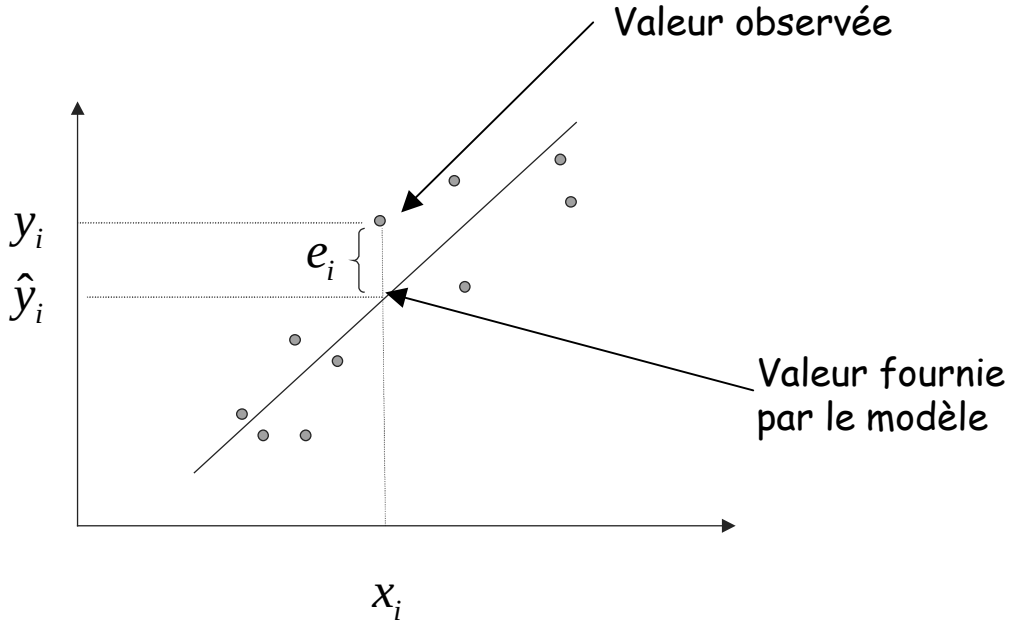
Régression linéaire multiple

Démarche de modélisation

La démarche de modélisation est toujours la même

- estimer les paramètres « a » en exploitant les données
- évaluer la précision de ces estimateurs
- mesurer le pouvoir explicatif du modèle
- évaluer l'influence des variables dans le modèle
    - globalement (toutes les *p* variables)
    - individuellement (chaque variable)
    - un bloc de variables (q variables, q < p)
- sélectionner les variables les plus « pertinentes »
- évaluer la qualité du modèle lors de la prédiction (intervalle de prédiction)
- détecter les observations qui peuvent influencer exagérément les résultats (points atypiques).

La méthode des moindres carrés



La méthode des moindres carrés cherche la meilleure estimation des paramètres « a » en minimisant la quantité

$$SCR = \sum_i e_i^2$$

$$avec\ e_i = Y - X\hat{a}$$

« e », l 'erreur observée est une évaluation du terme résiduel ε

Les hypothèses de la méthode des MCO

« â » deviennent les EMCO (estimateurs des moindres carrés ordinaires)

Hypothèses probabilistes

- *le modèle est linéaire en X*
- *les X sont observés sans erreur*
- *$E(\varepsilon) = 0$, en moyenne le modèle est bien spécifié*
- *$E(\varepsilon^2) = \sigma^2_\varepsilon$ la variance de l 'erreur est constante (hétéroscédasticité)*
- *$E(\varepsilon_i, \varepsilon_j)=0$, les erreurs sont non-corrélés*
- *$Cov(\varepsilon,x)=0$, l 'erreur est indépendante de la variable explicative*
- *$\varepsilon \equiv Normale(0, \sigma^2_\varepsilon)$*

Hypothèses structurelles

- *$Rang(X'X)=p+1$ càd $(X'X)^{-1}$ existe*
- *$(X'X)/n$ tend vers une matrice finie non singulière*
- *$n>p+1$, le nombre d 'observations est supérieur au nombre de variables explicatives*

Idée : rendre les calculs possibles et délimiter les propriétés des estimateurs

EMCO

Pour trouver les paramètres « a » qui minimise S :

$$S = \sum_i \varepsilon_i^{\,2} = \sum_i [y_i - (a_0 + a_{i,1}x_1 + \ldots + a_{i,p}x_p)]^2$$

On doit résoudre $\frac{\partial S}{\partial a} = 0$

Il y a (p+1) équations dites « équations normales » à résoudre

L'estimateur des moindres carrés ordinaires s'écrit : $\hat{a} = (X'X)^{-1}X'Y$

N.B. Compte tenu des hypothèses ci-dessus :
- â est sans biais
- â est convergent
- â est BLUE (c.-à-d. il n'existe pas d'estimateur linéaire sans biais de variance plus petite)

Un premier exemple – Cigarettes

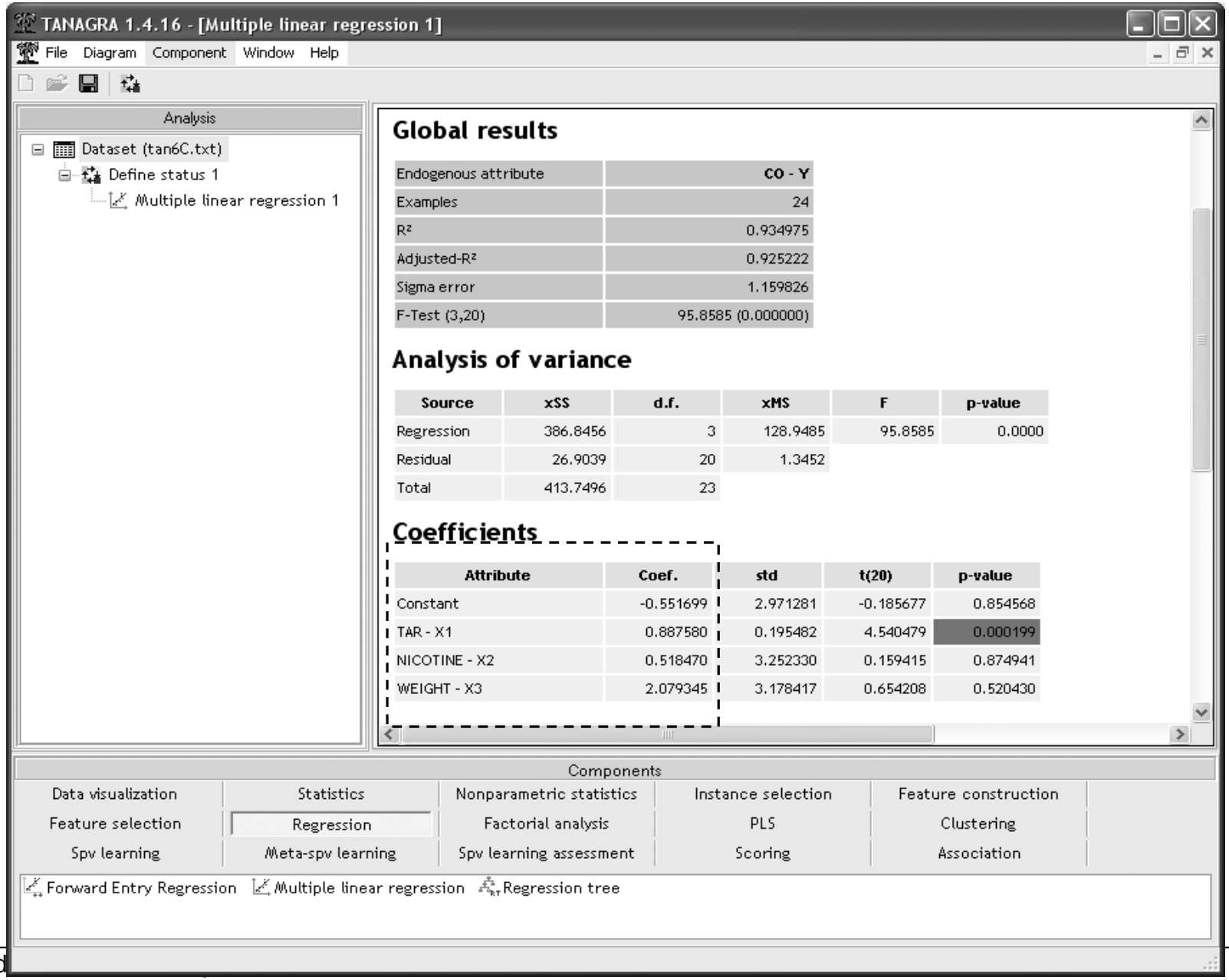
Dans le tableur EXCEL

Cigarette	TAR - X1	NICOTINE - X	WEIGHT - X3	CO - Y	Pred(Y)	e	e^2
Alpine	14.1	0.86	0.9853	13.6	14.46	-0.86	0.7
Benson&Hedge	16	1.06	1.0938	16.6	16.47	0.13	0.0
CamelLights	8	0.67	0.928	10.2	8.83	1.37	1.9
Carlton	4.1	0.4	0.9462	5.4	5.26	0.14	0.0
Chesterfield	15	1.04	0.8885	15	15.15	-0.15	0.0
GoldenLights	8.8	0.76	1.0267	9	9.79	-0.79	0.6
Kent	12.4	0.95	0.9225	12.3	12.87	-0.57	0.3
Kool	16.6	1.12	0.9372	16.3	16.71	-0.41	0.2
L&M	14.9	1.02	0.8858	15.4	15.04	0.36	0.1
LarkLights	13.7	1.01	0.9643	13	14.14	-1.14	1.3
Marlboro	15.1	0.9	0.9316	14.4	15.25	-0.85	0.7
Merit	7.8	0.57	0.9705	10	8.68	1.32	1.7
MultiFilter	11.4	0.78	1.124	10.2	12.31	-2.11	4.4
NewportLights	9	0.74	0.8517	9.5	9.59	-0.09	0.0
Now	1	0.13	0.7851	1.5	2.04	-0.54	0.3
OldGold	17	1.26	0.9186	18.5	17.10	1.40	2.0
PallMallLight	12.8	1.08	1.0395	12.6	13.53	-0.93	0.9
Raleigh	15.8	0.96	0.9573	17.5	15.96	1.54	2.4
SalemUltra	4.5	0.42	0.9106	4.9	5.55	-0.65	0.4
Tareyton	14.5	1.01	1.007	15.9	14.94	0.96	0.9
TrueLight	7.3	0.61	0.9806	8.5	8.28	0.22	0.0
ViceroyRichLight	8.6	0.69	0.9693	10.6	9.45	1.15	1.3
VirginiaSlims	15.2	1.02	0.9496	13.9	15.44	-1.54	2.4
WinstonLights	12	0.82	1.1184	14.9	12.85	2.05	4.2
						SCR	26.9

a3	a2	a1	a0
2.0793	0.5185	0.8876	-0.5517

Un premier exemple – Cigarettes

Dans le logiciel TANAGRA



Évaluation globale de la régression

Tableau d'analyse de variance et Coefficient de détermination

Équation d'analyse de variance – Décomposition de la variance

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

SCT
Variabilité totale

SCE
Variabilité expliquée par le modèle

SCR
Variabilité non-expliquée (Variabilité résiduelle)

Source de variation	**Somme des carrés**	**Degrés de liberté**	**Carrés moyens**
Modèle	SCE	p	SCE/p
Résiduel	SCR	n-p-1	SCR/(n-p-1)
Total	SCT	n-1	

Tableau d'analyse de variance

Un indicateur de qualité du modèle : le coefficient de détermination, il exprime la proportion de variabilité de Y qui est traduite par le modèle

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

R^2#1, le modèle est intéressant
R^2#0, le modèle est mauvais

Test associé à l'évaluation globale du modèle

Test de Fisher

Quelques formulations du test de « signification » globale :
- le modèle est-il pertinent pour expliquer les valeurs de Y ?
- la liaison linéaire Y / X1,…,Xp est-elle licite ?
- test d'hypothèse

$$\begin{cases} H_0 : a_1 = a_2 = \cdots = a_p = 0 \\ H_1 : \text{il en existe au moins } \neq 0 \end{cases}$$

Attention, on n'inclut pas la constante dans le test
c.-à-d. la question est donc, y a-t-il au moins une variable qui emmène de l'information ? Ou il n'est pas possible d'effectuer une prédiction/explication meilleure que la simple constante ?

Statistique du test

$$F = \frac{R^2 / p}{(1-R^2) / (n-p-1)} \equiv Fisher(p, n-p-1)$$

A un niveau de signification donné (ex. 10%, 5%, 1%…)
>> Comparer le F-calculé avec le F-théorique fourni par la table
>> Comparer la p-value avec le niveau de signification

Tableau d'ANOVA, R^2 et Test de signification globale

Données Cigarettes

TANAGRA

Global results

Endogenous attribute	**CO - Y**
Examples	24
R^2	0.934975
Adjusted-R^2	0.925222
Sigma error	1.159826
F-Test (3,20)	95.8585 (0.000000)

$\leftarrow \sqrt{1.3452}$

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	386.8456	3	128.9485	95.8585	0.0000
Residual	26.9039	20	1.3452		
Total	413.7496	23			

EXCEL

a3	a2	a1	a0
2.07934422	0.51846956	0.88758035	-0.55169763
3.17841712	3.25233113	0.19548169	2.97128094
0.93497531	1.15982622	#N/A	#N/A
95.8584963	20	#N/A	#N/A
386.845646	26.9039373	#N/A	#N/A

Évaluation individuelle des coefficients

Une variable contribue-t-elle de manière significative dans la régression ?

Test d'hypothèse associé

$$\begin{cases} H_0 : a_j = 0 \\ H_1 : a_j \neq 0 \end{cases}$$

H0 vérifiée signifie que la variable peut être supprimée du modèle sans en détériorer le pouvoir explicatif

Statistique du test : il s'appuie sur â et l'estimation de son écart-type

$$t = \frac{\hat{a}_j}{\hat{\sigma}_{\hat{a}_j}} \equiv Student(n - p - 1)$$

A un niveau de signification donné (ex. 10%, 5%, 1%…)
>> Comparer le t-calculé avec le t-théorique fourni par la table de Student
>> Comparer la p-value avec le niveau de signification

Test de signification individuelle des variables

Données « Cigarettes »

Tanagra

Coefficients

Attribute	Coef.	std	t(20)	p-value
Constant	-0.551699	2.971281	-0.185677	0.854568
TAR - X1	0.887580	0.195482	4.540479	0.000199
NICOTINE - X2	0.518470	3.252330	0.159415	0.874941
WEIGHT - X3	2.079345	3.178417	0.654208	0.520430

Excel

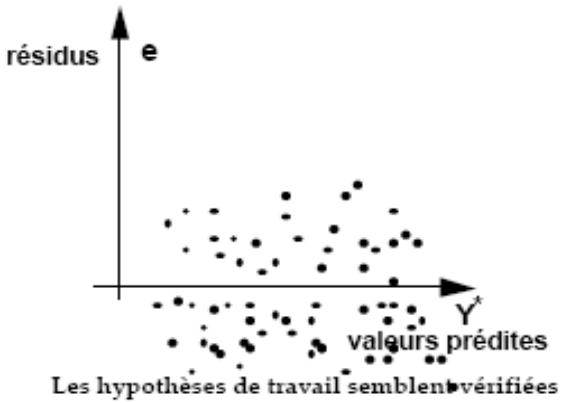
	a3	a2	a1	a0
a	2.07934422	0.51846956	0.88758035	-0.55169763
sigma(a)	3.17841712	3.25233113	0.19548169	2.97128094
	0.93497531	1.15982622	#N/A	#N/A
	95.8584963	20	#N/A	#N/A
	386.845646	26.9039373	#N/A	#N/A
t-calculé	0.65420747	0.15941475	4.54047821	
t-théorique (à 5%)	2.08596248	2.08596248	2.08596248	
p-value	0.52043066	0.87494102	0.00019909	

Diagnostiquer une régression avec les graphiques des résidus

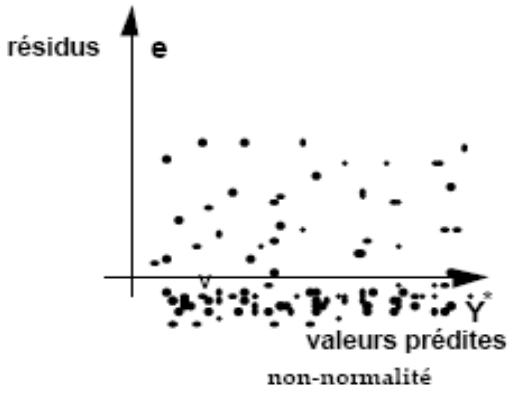
Les graphiques des résidus est un excellent outil de diagnostic de la régression, elles permettent notamment de vérifier l'adéquation aux hypothèses initiales et détecter l'existence éventuelle de points atypiques
» résidus vs. Y
» résidus vs. chaque variable X

→

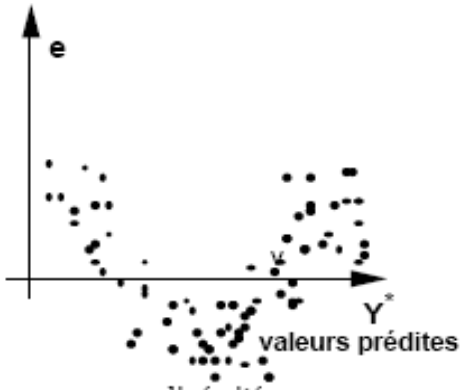
On veut vérifier :
» problème de spécification (non-linéarité du modèle)
» normalité des résidus (distribués aléatoirement et symétriquement autour de 0)
» variance constante (homoscédasticité)
» absence de « structure » dans l'évolution de l'erreur (non auto-corrélation – indépendance)



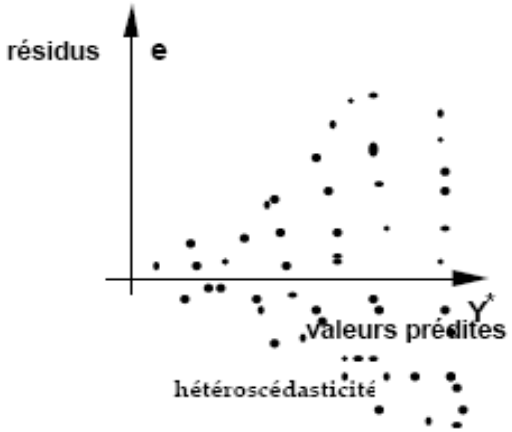
Les hypothèses de travail semblent vérifiées



non-normalité



non-linéarité



hétéroscédasticité

Bibliographique

- *http://fr.wikipedia.org/wiki/Régression_linéaire_multiple*
- *Y.Dodge, V.Rousson, « Analyse de régression appliquée », Dunod, 2004.*
- *R. Bourbonnais, « Économétrie », Dunod, 1998.*
- *M. Tenenhaus, « Méthodes Statistiques en Gestion », Dunod, 1996.*