

# Régression linéaire multiple

Prédire les valeurs d'une variable continue

Ricco Rakotomalala  
Ricco.Rakotomalala@univ-lyon2.fr

# Tableau de données et Statut des variables

Cigarette	TAR (m g)	NICOTINE (m g)	WEIGHT (g)	CO (m g)
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
Salem Ultra	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*

# 1. Le modèle linéaire général

# Régression linéaire multiple

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

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

Le terme aléatoire  $\varepsilon$  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)

$\varepsilon$  quantifie les écarts entre les valeurs réellement observées et les valeurs prédites par le modèle

$(a_0, a_1, \dots, a_p)$

Sont les paramètres/coefficients du modèle que l'on veut estimer à l'aide des données

# Lecture des coefficients

$$\frac{\partial y}{\partial x_j} = a_j$$

→ Le coefficient se lit comme une « propension marginale »

→ Toutes choses égales par ailleurs c.-à-d. l'impact de  $x_j$  sur  $y$  ne tient pas compte de l'influence des autres

→ L'effet des variables est additif c.-à-d. les autres variables étant constantes, si

$$\Delta x_j = 1 \text{ et } \Delta x_{j'} = 1 \Rightarrow \Delta y = (a_j + a_{j'})$$

→ Si on veut analyser les interactions, il faut donc construire des variables synthétiques ex.

$$y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + a_3 (x_{i1} * x_{i2}) + \varepsilon_i$$

Ex. Impact de « fumer » ET  
« boire » sur l'hypertension

# 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 (biais, variance, convergence)
- mesurer le pouvoir explicatif global du modèle
- évaluer l'influence des variables dans le modèle
  - globalement (toutes les  $p$  variables)
  - individuellement (chaque variable)
  - un bloc de variables ( $q$  variables,  $q \leq p$ ) [c'est une généralisation]
- sélectionner les variables les plus « pertinentes »
- évaluer la qualité du modèle lors de la prédiction (intervalle de prédiction)
- détecter les observations qui peuvent fausser ou influencer exagérément les résultats (points atypiques).

# Régression linéaire multiple

## Écriture matricielle

Pour une meilleure concision ...

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

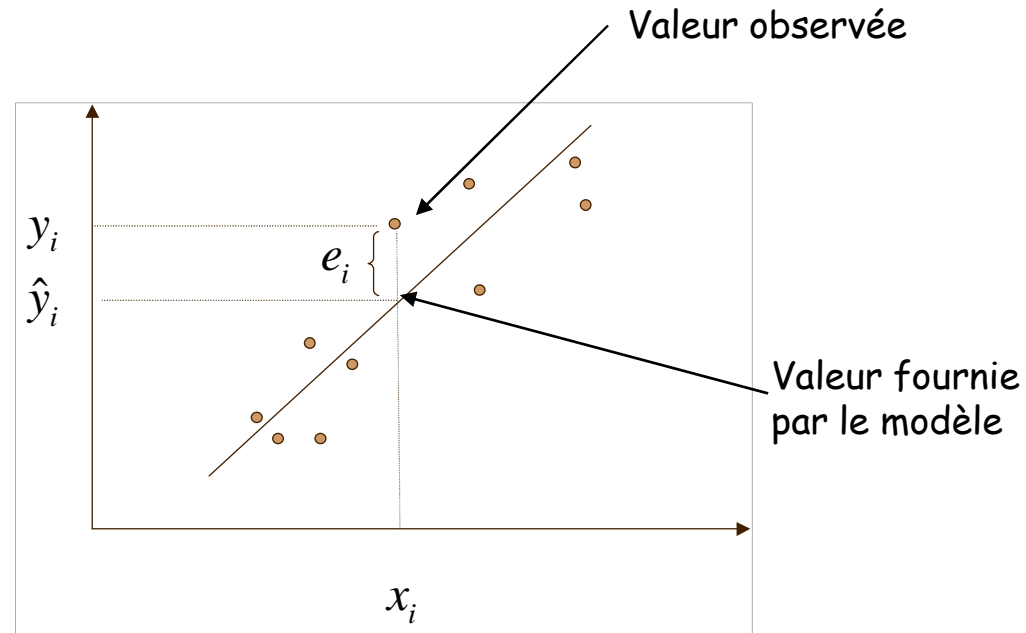
N.B. Noter la colonne  
représentant la constante

$$Y = Xa + \varepsilon$$

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

Bien noter les dimensions des matrices

# La méthode des moindres carrés



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

$$S = \sum_i e_i^2$$

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

« e », l'erreur observée (le résidu) est une évaluation du terme d'erreur  $\mathcal{E}$

# Les hypothèses de la méthode des MCO

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

## Hypothèses probabilistes (hypothèses stochastiques)

- les  $X$  sont observés sans erreur (non aléatoires)
- $E(\varepsilon) = 0$ , en moyenne le modèle est bien spécifié
- $E(\varepsilon^2) = \sigma_\varepsilon^2$  la variance de l'erreur est constante (homoscédasticité)
- $E(\varepsilon_i, \varepsilon_j) = 0$ , les erreurs sont non-corrélées (non-autocorrélation des erreurs)
- $Cov(\varepsilon, x) = 0$ , l'erreur est indépendante des variables explicatives
- $\varepsilon \equiv \text{Normale}(0, \sigma_\varepsilon)$

## Hypothèses structurelles

- $\text{Rang}(X'X) = p+1$  c.-à-d.  $(X'X)^{-1}$  existe, ou encore  $\det(X'X) \neq 0$
- $(X'X)/n$  tend vers une matrice finie non singulière quand  $n \rightarrow +\infty$
- $n > p+1$ , le nombre d'observations est supérieur au nombre de paramètres du modèle (variables explicatives + constante)

Ces hypothèses pèsent sur les propriétés des estimateurs et sur les lois de distribution

# EMCO (Estimateur des moindres carrés ordinaires)

## Principe de calcul - Estimateur

Pour trouver les paramètres « a » qui minimise S :

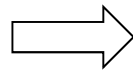
$$S = \varepsilon' \varepsilon$$
$$= \sum_i \varepsilon_i^2 = \sum_i [y_i - (a_0 + a_{i,1}x_1 + \dots + a_{i,p}x_p)]^2$$

On doit résoudre

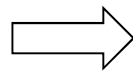
$$\frac{\partial S}{\partial a} = 0$$

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

Les bonnes pistes...



$$S = \varepsilon' \varepsilon = (Y - Xa)'(Y - Xa)$$
$$= Y'Y - 2a'X'Y + a'X'Xa$$



$$\frac{\partial S}{\partial a} = -2(X'Y) + 2(X'X)a = 0$$



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

# EMCO (Estimateur des moindres carrés ordinaires)

## Commentaires

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

$$(X'X) = \begin{pmatrix} n & \sum_i x_{i,1} & \cdots & \sum_i x_{i,p} \\ \sum_i x_{i,1} & \sum_i x_{i,1}^2 & & \sum_i x_{i,1} \times x_{i,p} \\ & & & \\ & & & \sum_i x_{i,p}^2 \end{pmatrix}$$

(p+1,p+1)

Matrice des sommes des produits croisés entre les variables exogènes – Symétrique (son inverse aussi est symétrique)

Si les variables sont centrées

→  $1/n (X'X)$  = matrice de variance covariance

Si les variables sont centrées et réduites

→  $1/n (X'X)$  = matrice de corrélation

$$(X'Y) = \begin{pmatrix} \sum_i y_i \\ \sum_i y_i x_{i,1} \\ \vdots \\ \sum_i y_i x_{i,p} \end{pmatrix}$$

(p+1, 1)

Vecteur des sommes des produits croisés entre l'endogène et les variables exogènes

Si les variables sont centrées

→  $1/n (X'Y)$  = vecteur des covariances entre Y et X

Si les variables sont centrées et réduites

→  $1/n (X'Y)$  = vecteur des corrélations entre Y et X

# Un premier exemple – Cigarettes

Dans le tableur EXCEL

constante	TAR (mg)	NICOTINE (mg)	WEIGHT (g)	CO (mg)
1	14.1	0.86	0.9853	13.6
1	16	1.06	1.0938	16.6
1	8	0.67	0.928	10.2
1	4.1	0.4	0.9462	5.4
1	15	1.04	0.8885	15
1	8.8	0.76	1.0267	9
1	12.4	0.95	0.9225	12.3
1	16.6	1.12	0.9372	16.3
1	14.9	1.02	0.8858	15.4
1	13.7	1.01	0.9643	13
1	15.1	0.9	0.9316	14.4
1	7.8	0.57	0.9705	10
1	11.4	0.78	1.124	10.2
1	9	0.74	0.8517	9.5
1	1	0.13	0.7851	1.5
1	17	1.26	0.9186	18.5
1	12.8	1.08	1.0395	12.6
1	15.8	0.96	0.9573	17.5
1	4.5	0.42	0.9106	4.9
1	14.5	1.01	1.007	15.9
1	7.3	0.61	0.9806	8.5
1	8.6	0.69	0.9693	10.6
1	15.2	1.02	0.9496	13.9
1	12	0.82	1.1184	14.9

(XX)

24	275.6	19.88	23.0921
275.6	3613.16	254.177	267.46174
19.88	254.177	18.0896	19.266811
23.0921	267.46174	19.266811	22.3637325

(XX)<sup>-1</sup>

6.56299	0.06290	-0.93908	-6.71991
0.06290	0.02841	-0.45200	-0.01528
-0.93908	-0.45200	7.86328	-0.39900
-6.71991	-0.01528	-0.39900	7.50993

XY

289.7
3742.85
264.076
281.14508

a<sup>^</sup>

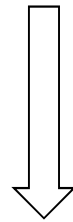
-0.55170	constante
0.88758	tar
0.51847	nicotine
2.07934	weight

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

## 2. Propriétés des estimateurs

## Biais de « $\hat{a}$ »

$$\begin{aligned}\hat{a} &= (X'X)^{-1} X'Y \\ &= (X'X)^{-1} X'[Xa + \varepsilon] \\ \hat{a} &= a + (X'X)^{-1} X'\varepsilon\end{aligned}$$



$$\begin{aligned}E[\hat{a}] &= a + E[(X'X)^{-1} X'\varepsilon] \\ &= a + (X'X)^{-1} X'E[\varepsilon] \\ &= a\end{aligned}$$

Etape 1. Exprimer «  $\hat{a}$  » en fonction de «  $a$  »

Etape 2. Voir sous quelles conditions  $E[\hat{a}] = a$

← Parce que  $X$  non aléatoire

← Parce que  $E[\varepsilon] = 0$  par hypothèse

# Matrice de variance covariance de « $\hat{a}$ »

$$\Omega_{\hat{a}} = E[(\hat{a} - a)(\hat{a} - a)']$$

$$= \begin{pmatrix} V(\hat{a}_0) & COV(\hat{a}_0, \hat{a}_1) & \dots \\ & V(\hat{a}_1) & \\ & & \ddots \\ & & & V(\hat{a}_p) \end{pmatrix}$$

Sur la diagonale, nous disposons de la variance de l'estimation de chaque coefficient. Très utile dans la partie inférentielle.

Puisque

$$\hat{a} - a = (X'X)^{-1} X' \varepsilon \quad \Longrightarrow \quad E[(\hat{a} - a)(\hat{a} - a)'] = (X'X)^{-1} X' E[\varepsilon \varepsilon'] X (X'X)^{-1}$$

Or, par hypothèse :

- $E(\varepsilon^2) = \sigma_\varepsilon^2$  la variance de l'erreur est constante (homoscédasticité)
- $E(\varepsilon_i, \varepsilon_j) = 0$ , les erreurs sont non-corrélées (non-autocorrélation des erreurs)

$$E[\varepsilon \varepsilon'] = \begin{pmatrix} E[\varepsilon_1^2] & E[\varepsilon_1 \varepsilon_2] & \dots & E[\varepsilon_1 \varepsilon_n] \\ & E[\varepsilon_2^2] & & \\ & & \ddots & \\ & & & E[\varepsilon_n^2] \end{pmatrix}$$

(n, n)



$$E[\varepsilon \varepsilon'] = \sigma_\varepsilon^2 I_n$$



$$\Omega_{\hat{a}} = \sigma_\varepsilon^2 (X'X)^{-1}$$

On montre que cette matrice tend vers la matrice nulle (toutes les cellules à 0) lorsque  $n \rightarrow +\infty$  : EMCO est convergent.

On montre de plus que l'EMCO est BLUE (best linear unbiased estimator).

# Variance de l'erreur

$$\Omega_{\hat{a}} = \sigma_{\varepsilon}^2 (X'X)^{-1}$$



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

Pour estimer la variance covariance des coefficients, il faut produire une estimation de la variance de l'erreur.

Développons le résidu

$$\begin{aligned} \hat{\varepsilon} &= Y - \hat{Y} \\ &= (Xa + \varepsilon) - X\hat{a} \\ &= (Xa + \varepsilon) - X \left[ a + (X'X)^{-1} X' \varepsilon \right] \\ \hat{\varepsilon} &= \underbrace{\left[ I - X(X'X)^{-1} X' \right]}_{\Gamma} \varepsilon \end{aligned}$$

Appelée matrice  $\Gamma$ , elle est symétrique ( $\Gamma' = \Gamma$ ) et idempotente ( $\Gamma^2 = \Gamma$ ), de taille  $(n, n)$

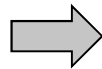
$$\hat{\varepsilon}' \hat{\varepsilon} = \varepsilon' \Gamma \varepsilon$$

On montre alors que :

$$E[\hat{\varepsilon}' \hat{\varepsilon}] = \sigma_{\varepsilon}^2 \times Tr(\Gamma)$$

Variance de l'erreur

Degrés de liberté =  $n - (p+1) = n - p - 1$



Estimateur sans biais de la variance de l'erreur

$$\begin{aligned} \hat{\sigma}_{\varepsilon}^2 &= \frac{\hat{\varepsilon}' \hat{\varepsilon}}{Tr(\Gamma)} \\ &= \frac{\sum_i \hat{\varepsilon}_i^2}{n - p - 1} = \frac{SCR}{n - p - 1} \\ &= \frac{Y' \Gamma Y}{n - p - 1} \end{aligned}$$

Remarque : voir l'analogie avec la régression linéaire simple !!!

# Calculs sous Excel

onstant	TAR (mg)	OTINE (EIGHT)	CO (mg)	Y^	Résidus	Résidus^2	
1	14.1	0.86	0.9853	13.6	14.458	-0.858	0.7359
1	16	1.06	1.0938	16.6	16.474	0.1264	0.016
1	8	0.67	0.928	10.2	8.826	1.374	1.888
1	4.1	0.4	0.9462	5.4	5.2622	0.1378	0.019
1	15	1.04	0.8885	15	15.149	-0.149	0.0221
1	8.8	0.76	1.0267	9	9.7879	-0.788	0.6208
1	12.4	0.95	0.9225	12.3	12.865	-0.565	0.3193
1	16.6	1.12	0.9372	16.3	16.712	-0.412	0.1694
1	14.9	1.02	0.8858	15.4	15.044	0.356	0.1268
1	13.7	1.01	0.9643	13	14.137	-1.137	1.2926
1	15.1	0.9	0.9316	14.4	15.255	-0.855	0.7302
1	7.8	0.57	0.9705	10	8.685	1.315	1.7293
1	11.4	0.78	1.124	10.2	12.308	-2.108	4.445
1	9	0.74	0.8517	9.5	9.5912	-0.091	0.0083
1	1	0.13	0.7851	1.5	2.0358	-0.536	0.2871
1	17	1.26	0.9186	18.5	17.101	1.3995	1.9585
1	12.8	1.08	1.0395	12.6	13.531	-0.931	0.8663
1	15.8	0.96	0.9573	17.5	15.96	1.5396	2.3705
1	4.5	0.42	0.9106	4.9	5.5536	-0.654	0.4272
1	14.5	1.01	1.007	15.9	14.936	0.9642	0.9297
1	7.3	0.61	0.9806	8.5	8.2829	0.2171	0.0471
1	8.6	0.69	0.9693	10.6	9.4547	1.1453	1.3116
1	15.2	1.02	0.9496	13.9	15.443	-1.543	2.3806
1	12	0.82	1.1184	14.9	12.85	2.05	4.2027
					<b>SCR</b>	<b>26.904</b>	

a^			
-0.55169763	constante		
0.887580347	tar		
0.518469559	nicotine		
2.079344216	weight		
n 24			
p 3			
ddl 20			
sigma^2(epsilon) 1.345197			
sigma(epsilon) 1.15982622			
(XX)^-1			
6.56299	0.06290	-0.93908	-6.71991
0.06290	0.02841	-0.45200	-0.01528
-0.93908	-0.45200	7.86328	-0.39900
-6.71991	-0.01528	-0.39900	7.50993
Mat. Var-covar des coefficients			
8.82851	0.08461	-1.26324	-9.03960
0.08461	0.03821	-0.60803	-0.02055
-1.26324	-0.60803	10.57766	-0.53673
-9.03960	-0.02055	-0.53673	10.10234
Ecart-types des coefficients			
constante	tar	nicotine	weight
2.97128	0.19548	3.25233	3.17842

DROITEREG				
coef.	weight	nicotine	tar	constante
	2.07934422	0.51846956	0.88758035	-0.55169763
ecart-type	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

Calcul avec la fonction DROITEREG d'EXCEL

Calcul matriciel sous EXCEL

### **3. Inférence statistique sur les coefficients**

# Distribution de $\hat{a}$

Par hypothèse,

$$\varepsilon \equiv N(0, \sigma_\varepsilon)$$



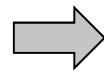
$$\begin{cases} \frac{\hat{a}_j - a_j}{\sigma_{\hat{a}_j}} \equiv N(0,1) \\ (n-p-1) \frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \equiv \chi^2(n-p-1) \end{cases}$$

*Cf. le cours de Régression simple*

*Toujours par analogie avec la régression simple, on peut montrer que*



$$(n-p-1) \frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} = (n-p-1) \frac{\hat{\sigma}_{\hat{a}_j}^2}{\sigma_{\hat{a}_j}^2}$$



$$\frac{\hat{a}_j - a_j}{\hat{\sigma}_{\hat{a}_j}} \equiv \mathcal{S}(n-p-1)$$

Loi de Student à  $(n-p-1)$  degrés de liberté.

On peut la mettre en œuvre dans différents schémas.

Test de conformité à un standard c.-à-d.  $H_0 : a_j = c$  vs.  $H_1 : a_j \neq c$   
*Bilatéral ou unilatéral*

Test de significativité c.-à-d.  $H_0 : a_j = 0$  vs.  $H_1 : a_j \neq 0$   
Permet de déterminer si la variable  $X_j$  a un impact sur  $Y$  !!!

Intervalle de confiance au niveau  $(1 - \alpha)$

# Exemple des cigarettes

$\hat{a}_j$

$\hat{\sigma}_{\hat{a}_j}$

**DROITEREG**

	weight	nicotine	tar	constante
coef.	2.07934422	0.51846956	0.88758035	-0.55169763
ecart-type	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

**Test de significativité à 5%**

t calculé	0.65421	0.15941	4.54048	-0.18568
abs.t.-calculé	0.65421	0.15941	4.54048	0.18568
t théorique (5%)	2.08596	2.08596	2.08596	2.08596

Décision

H0	H0	H1	H0
----	----	----	----

**Intervalles de confiance à 95%**

borne.basse	-4.55072	-6.26577	0.47981	-6.74968
borne.haute	8.70941	7.30271	1.29535	5.64629

$$t = \frac{\hat{a}_j - 0}{\hat{\sigma}_{\hat{a}_j}}$$

$$t_{1-\alpha/2}(20) = 2.08596$$

Rejet de H0 si  $|t| \geq t_{1-\alpha/2}(20)$

$$\hat{a}_j \pm t_{1-\alpha/2} \times \hat{\sigma}_{\hat{a}_j}$$

## 4. Evaluation globale de la régression

# 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	

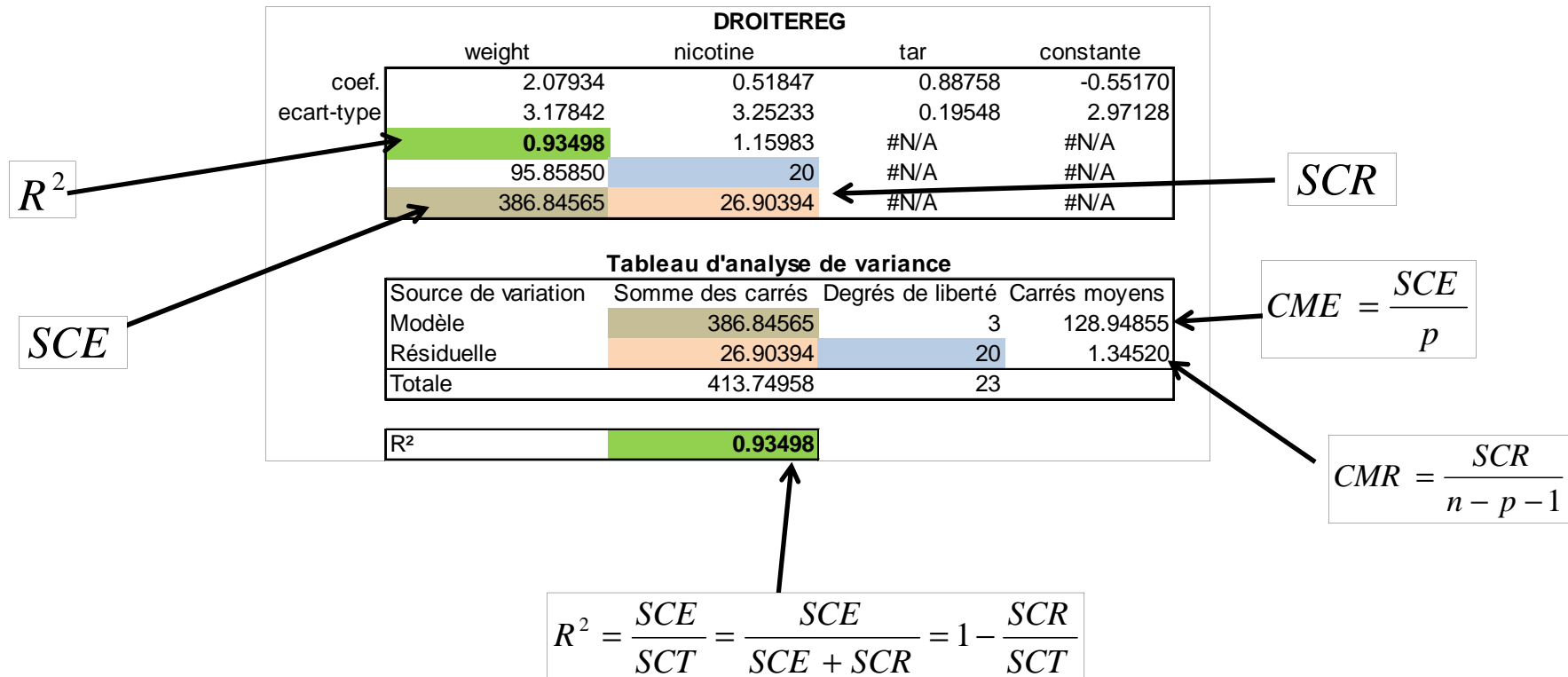
Tableau d'analyse de variance

Un indicateur de qualité du modèle : le coefficient de détermination. Il exprime la proportion de variabilité de Y qui est retranscrite par le modèle

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

*R<sup>2</sup> #1, le modèle est parfait  
R<sup>2</sup> #0, le modèle est mauvais*

# Exemple des cigarettes



# R<sup>2</sup> corrigé pour contrecarrer le sur-ajustement

Problème : Le R<sup>2</sup> augmente mécaniquement avec le nombre de variables. Même si les variables additionnelles ne sont absolument pas pertinentes.

→ On ne peut pas comparer des modèles de complexité différente (avec un nombre d'exogènes différent) sur la base du R<sup>2</sup>

→ Il faut utiliser le R<sup>2</sup> ajusté qui est un R<sup>2</sup> corrigé par les degrés de liberté.

$$\bar{R}^2 = 1 - \frac{SCR/n - p - 1}{SCT/n - 1}$$

TAR (mg)	OTINE	EIGHT	ALEA	CO (mg)
14.1	0.86	0.9853	0.2678	13.6
16	1.06	1.0938	0.3578	16.6
8	0.67	0.928	0.1269	10.2
4.1	0.4	0.9462	0.228	5.4
15	1.04	0.8885	0.109	15
8.8	0.76	1.0267	0.0388	9
12.4	0.95	0.9225	0.3959	12.3
16.6	1.12	0.9372	0.275	16.3
14.9	1.02	0.8858	0.8524	15.4
13.7	1.01	0.9643	0.1624	13
15.1	0.9	0.9316	0.673	14.4
7.8	0.57	0.9705	0.6751	10
11.4	0.78	1.124	0.8474	10.2
9	0.74	0.8517	0.5497	9.5
1	0.13	0.7851	0.4322	1.5
17	1.26	0.9186	0.9799	18.5
12.8	1.08	1.0395	0.3964	12.6
15.8	0.96	0.9573	0.4354	17.5
4.5	0.42	0.9106	0.5534	4.9
14.5	1.01	1.007	0.6546	15.9
7.3	0.61	0.9806	0.5156	8.5
8.6	0.69	0.9693	0.5019	10.6
15.2	1.02	0.9496	0.7209	13.9
12	0.82	1.1184	0.8171	14.9

**DROITEREG (TAR, NICOTINE, WEIGHT) - 1**

	weight	nicotine	tar	constante
	2.07934	0.51847	0.88758	-0.55170
	3.17842	3.25233	0.19548	2.97128
R <sup>2</sup>	<b>0.93498</b>	1.15983	#N/A	#N/A
	95.85850	20	#N/A	#N/A
	386.85	26.90	#N/A	#N/A

**DROITEREG (TAR, NICOTINE, WEIGHT, ALEA) - 2**

	alea	weight	nicotine	tar	constante
	0.81653	1.87048	0.93450	0.85569	-0.72260
	0.96657	3.21095	3.31268	0.20048	2.99961
R <sup>2</sup>	<b>0.93733</b>	1.16822	#N/A	#N/A	#N/A
	71.04289	19	#N/A	#N/A	#N/A
	387.82	25.93	#N/A	#N/A	#N/A

Le modèle (1) est le plus intéressant finalement !!!

R<sup>2</sup> ajusté (1) 0.92522

R<sup>2</sup> ajusté (2) 0.92414

$$\bar{R}^2(1) = 1 - \frac{26.90 / (24 - 3 - 1)}{(386.85 + 26.90) / (24 - 1)} = 1 - \frac{26.90 / 20}{413.75 / 23} = 0.92522$$

$$\bar{R}^2(2) = 1 - \frac{25.93 / (24 - 4 - 1)}{(387.82 + 25.93) / (24 - 1)} = 1 - \frac{25.93 / 19}{413.75 / 23} = 0.92414$$

La réduction du SCR est contrecarrée par la réduction des DDL.

# Test de significativité globale de la régression

Les X emmènent-elles de l'information sur Y ?

Statistiquement,  
le test s'écrit.

$$\begin{cases} H_0 : a_1 = a_2 = \dots = a_p = 0 \\ H_1 : \exists j / a_j \neq 0 \end{cases}$$

← Aucune variable exogène n'est pertinente pour expliquer Y  
← Une des exogènes au moins est porteuse d'information

Statistique de test

$$F = \frac{CME}{CMR} = \frac{SCE/p}{SCR/(n-p-1)} = \frac{R^2/p}{(1-R^2)/(n-p-1)}$$

Distribution sous H0

$$F \equiv Fisher(p, n-p-1)$$

Région critique  
au risque  $\alpha$

$$F \geq F_{1-\alpha}(p, n-p-1)$$

$$F = \frac{386.84565/3}{26.90934/20} = 95.85850$$

DROITEREG				
	weight	nicotine	tar	constante
coef.	2.07934	0.51847	0.88758	-0.55170
-type	3.17842	3.25233	0.19548	2.97128
	0.93498	1.15983	#N/A	#N/A
	<b>95.85850</b>	20	#N/A	#N/A
	386.84565	26.90394	#N/A	#N/A

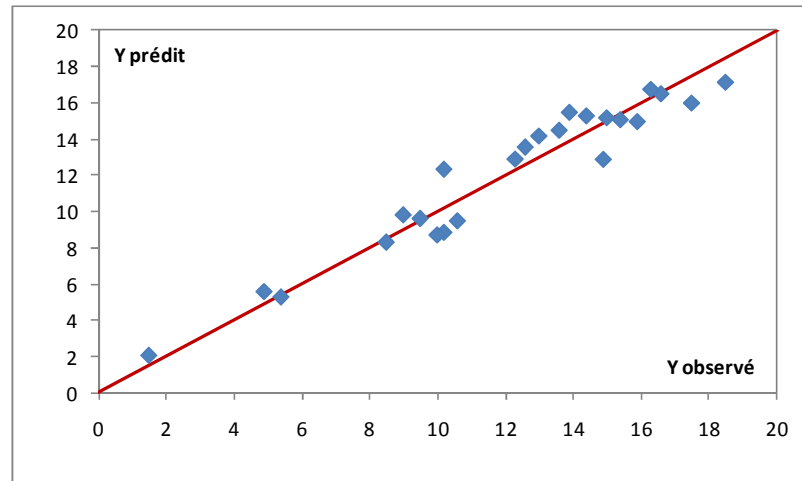
Tableau d'analyse de variance			
Source de variation	Somme des carrés	Degrés de liberté	Carrés moyens
Modèle	386.84565	3	128.94855
Résiduelle	26.90394	20	1.34520
Totale	413.74958	23	

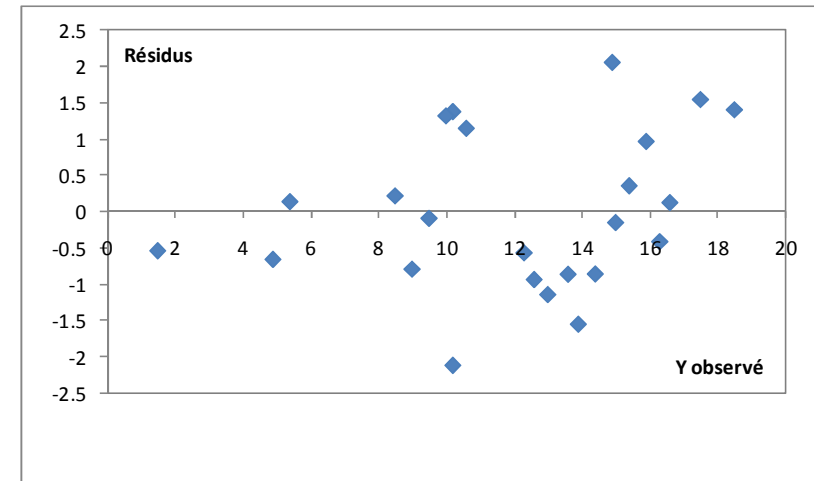
R <sup>2</sup>	0.93498
F	<b>95.85850</b>
ddl1	3
ddl2	20
F-théorique (95%)	3.09839
Conclusion	Rejet de H0

# Diagnostic graphique

Evaluer la qualité de la prédiction – Détecter les cas pathologiques



Y observé vs. Y prédit



Y observé vs. résidu

→ L'analyse des résidus fera l'objet d'un chapitre à part. Elle est très importante pour diagnostiquer la régression.

## 5. Tests généralisés des coefficients

# Test de conformité

Peut être utilisé pour tester la nullité simultanée de plusieurs coefficients

Tester la conformité d'un sous ensemble  $q$  de coefficients à un standard ( $q \leq p$ ).

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

Attention, la notation ne doit pas nous induire en erreur : on teste bien  $q$  paramètres quelconques parmi les  $p$ .

Un des coefficients au moins est conforme au standard.

Statistique de test

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

Distribution sous  $H_0$

$$F \equiv Fisher(q, n - p - 1)$$

Région critique au risque  $\alpha$

$$F \geq F_{1-\alpha}(q, n - p - 1)$$

Est l'inverse de la matrice de variance covariance réduite aux coefficients testés.

Les tests de significativité individuelle des coefficients et le test de significativité globale de la régression sont des cas particuliers.

# Exemple « Cigarettes »

Tester la nullité simultanée des coefficients de WEIGHT et NICOTINE  
(qui individuellement ne sont pas significatives...)

$$\begin{cases} H_0 : \begin{pmatrix} a_{\text{nicotine}} \\ a_{\text{weight}} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ H_1 : \exists j / a_j \neq c_j \end{cases}$$

Coefficients estimés.

a^	
-0.55170	constante
0.88758	tar
0.51847	nicotine
2.07934	weight



Coefficients à tester.

$$\hat{a}_{(q)} = \begin{pmatrix} 0.51847 \\ 2.07934 \end{pmatrix}$$

Mat. Var-covar des coefficients			
constante	tar	nicotine	weight
8.82851	0.08461	-1.26324	-9.03960
0.08461	0.03821	-0.60803	-0.02055
-1.26324	-0.60803	10.57766	-0.53673
-9.03960	-0.02055	-0.53673	10.10234



$$\hat{\Omega}_{\hat{a}_{(q)}}^{-1} = \begin{pmatrix} 10.57766 & -0.53673 \\ -0.53673 & 10.10234 \end{pmatrix}^{-1} = \begin{pmatrix} 0.09479 & 0.00504 \\ 0.00504 & 0.09925 \end{pmatrix}$$

Statistique de test

$$F = \frac{1}{q} [\hat{a}_{(q)} - c_{(q)}]' \hat{\Omega}_{\hat{a}_{(q)}}^{-1} [\hat{a}_{(q)} - c_{(q)}] = \frac{1}{2} \left[ \begin{pmatrix} 0.51847 \\ 2.07934 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right]' \hat{\Omega}_{\hat{a}_{(q)}}^{-1} \left[ \begin{pmatrix} 0.51847 \\ 2.07934 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right] = 0.23274$$

F théorique au risque  $\alpha = 5\%$

$$F_{1-\alpha}(q, n - p - 1) = F_{0.95}(2, 20) = 3.49283$$



L'hypothèse nulle ne peut pas être rejetée au risque  $\alpha = 5\%$

# Test de « q » contraintes linéaires sur les coefficients

Peut être utilisé pour comparer des coefficients

Tester q contraintes linéaires sur les coefficients : R est une matrice de dimension (q , p+1) ; r un vecteur de taille (q, 1)

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

Statistique de test

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

Distribution sous H0

$$F \equiv Fisher(q, n - p - 1)$$

Région critique au risque  $\alpha$

$$F \geq F_{1-\alpha}(q, n - p - 1)$$

Le test de significativité individuelle en est un cas particulier (ex.  $a_1 = 0$ )  $\rightarrow R(1, p+1)$

$$\Rightarrow \begin{cases} R = (0 \ 1 \ \dots \ 0) \\ r = (0) \end{cases}$$

Le test de significativité globale est un cas particulier  $R(p, p+1)$

$$\Rightarrow \begin{cases} R = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}; r = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \end{cases}$$

## Exemple « Cigarettes »

Tester l'égalité des coefficients de TAR et NICOTINE

$$\begin{cases} H_0 : a_{\text{tar}} = a_{\text{nicotine}} \\ H_1 : a_{\text{tar}} \neq a_{\text{nicotine}} \end{cases} \Leftrightarrow \{H_0 : 0 \times a_{\text{constante}} + 1 \times a_{\text{tar}} - 1 \times a_{\text{nicotine}} + 0 \times a_{\text{weight}} = 0\}$$

$$\Rightarrow \begin{cases} R = (0 & 1 & -1 & 0) \\ r = (0) \end{cases}$$

Statistique de test

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

$$= \frac{\frac{1}{1} (0.3691 - 0)' [8.7957]^{-1} (0.3691 - 0)}{26.904 / 20}$$

$$= \frac{0.0155}{1.3452} = 0.0115$$

a^	
-0.55170	constante
0.88758	tar
0.51847	nicotine
2.07934	weight

(X'X)^-1			
6.56299	0.06290	-0.93908	-6.71991
0.06290	0.02841	-0.45200	-0.01528
-0.93908	-0.45200	7.86328	-0.39900
-6.71991	-0.01528	-0.39900	7.50993

F théorique au risque  $\alpha = 5\%$

$$F_{1-\alpha}(q, n - p - 1) = F_{0.95}(1, 20) = 4.3512$$



L'hypothèse nulle ne peut pas être rejetée au risque  $\alpha = 5\%$

## 6. Prédiction et intervalle de prédiction

# Prédiction

## Prédiction ponctuelle et intervalle de prédiction

Ne pas oublier la constante  
en notation matricielle

$$X_{i^*} = \begin{pmatrix} 1 & x_{i^*,1} & \cdots & x_{i^*,p} \end{pmatrix}$$

Prédiction  
ponctuelle

$$\begin{aligned} \hat{y}_{i^*} &= \hat{y}(x_{i^*}) = \hat{a}_0 + \hat{a}_1 x_{i^*,1} + \cdots + \hat{a}_p x_{i^*,p} \\ &= X_{i^*} \hat{a} \end{aligned}$$

Prédiction sans biais

$$\begin{aligned} E(\hat{y}_{i^*}) &= y_{i^*} \\ \Leftrightarrow E(\hat{\varepsilon}_{i^*}) &= E(\hat{y}_{i^*} - y_{i^*}) = 0 \end{aligned}$$

Estimation de la  
variance de l'erreur  
de prédiction

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

Dépend de la qualité du modèle (variance de l'erreur) et de l'éloignement du point par rapport au barycentre (cf. lorsque variables centrées)

Distribution

$$\frac{\hat{\varepsilon}_{i^*}}{\hat{\sigma}_{\hat{\varepsilon}_{i^*}}} = \frac{\hat{y}_{i^*} - y_{i^*}}{\hat{\sigma}_{\hat{\varepsilon}_{i^*}}} \equiv \mathcal{J}(n - p - 1)$$

Au niveau de  
confiance  $(1 - \alpha)$



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

# Prédiction – Exemple « cigarettes »

## Prédiction ponctuelle et intervalle de prédiction

constante	TAR (mg)	NICOTINE (mg)	WEIGHT (g)	CO (mg)
1	14.1	0.86	0.9853	13.6
1	16	1.06	1.0938	16.6
1	8	0.67	0.928	10.2
1	4.1	0.4	0.9462	5.4
1	15	1.04	0.8885	15
1	8.8	0.76	1.0267	9
1	12.4	0.95	0.9225	12.3
1	16.6	1.12	0.9372	16.3
1	14.9	1.02	0.8858	15.4
1	13.7	1.01	0.9643	13
1	15.1	0.9	0.9316	14.4
1	7.8	0.57	0.9705	10
1	11.4	0.78	1.124	10.2
1	9	0.74	0.8517	9.5
1	1	0.13	0.7851	1.5
1	17	1.26	0.9186	18.5
1	12.8	1.08	1.0395	12.6
1	15.8	0.96	0.9573	17.5
1	4.5	0.42	0.9106	4.9
1	14.5	1.01	1.007	15.9
1	7.3	0.61	0.9806	8.5
1	8.6	0.69	0.9693	10.6
1	15.2	1.02	0.9496	13.9
1	12	0.82	1.1184	14.9

a^				
-0.55170	constante			
0.88758	tar			
0.51847	nicotine			
2.07934	weight			
n	24			
p	3			
ddl	20			
sigma^2(epsilon)	1.34520			sigma(epsilon) 1.15983
(XX)^-1				
6.56299	0.06290	-0.93908	-6.71991	
0.06290	0.02841	-0.45200	-0.01528	
-0.93908	-0.45200	7.86328	-0.39900	
-6.71991	-0.01528	-0.39900	7.50993	

X (à prédire)

constante	TAR (mg)	NICOTINE (mg)	WEIGHT (g)
1	11.5	0.8	0.95

$X_{i^*} = (1 \quad 11.5 \quad 0.8 \quad 0.95)$

Pred. Ponctuelle:  $\hat{y}_{i^*} = -0.55170 + 0.88758 \times 11.5 + 0.51847 \times 0.8 + 2.07934 \times 0.95$   
12.04563

Var.Erreur:  $\hat{\sigma}_\varepsilon^2$   
1.34520

Var.Erreur.Prediction:  $\hat{\sigma}_{\hat{\varepsilon}_{i^*}}^2 = \hat{\sigma}_\varepsilon^2 [1 + X_{i^*} (X'X)^{-1} X_{i^*}']$   
1.41146

t de Student:  $t_{1-\alpha/2} (n - p - 1) = t_{0.975} (20)$   
2.08596

borne.basse: 9.56740  
borne.haute: 14.52385

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

# Bibliographique

- [http://fr.wikipedia.org/wiki/Régression lineaire multiple](http://fr.wikipedia.org/wiki/Régression_lineaire_multiple)
- Y.Dodge, V.Rousson, « *Analyse de régression appliquée* », Dunod, 2004.
- R. Bourbonnais, « *Économétrie* », Dunod, 1998.
- M. Tenenhaus, « *Statistique : Méthodes pour décrire, expliquer et prévoir* », Dunod, 2007.