Année Universitaire 2020-2021

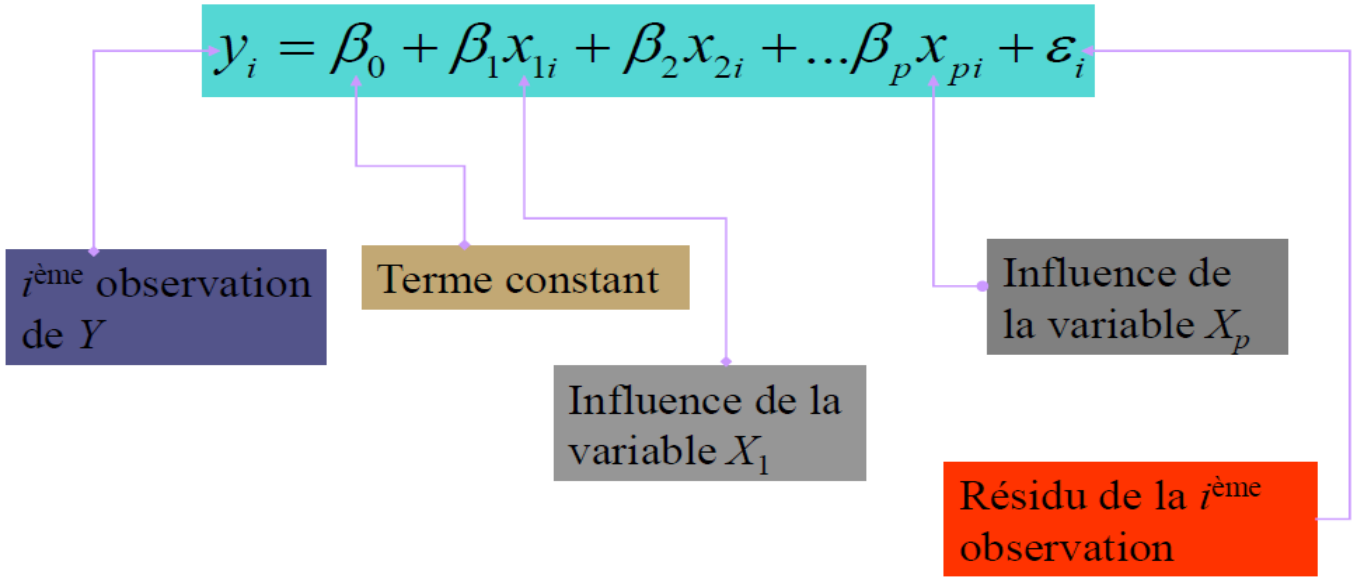
Modèle de Régression Linéaire Multiple (MLRM)

Slim Zouaoui & Walid Barhoumi

Modèle linéaire multiple

Estimer la relation entre une variable dépendante(Y) quantitative et plusieurs variables indépendantes (X1,X2, …)

→ Equation de régression multiple : Cette équation précise la façon dont la variable dépendante est reliée aux variables explicatives :



Démarche de modélisation

La démarche de modélisation est toujours la même

- ✓ estimer les paramètres « β » en exploitant les données
- ✓ évaluer la précision de ces estimateurs
- ✓ mesurer le pouvoir explicatif global du modèle
- ✓ évaluer l'influence des variables dans le modèle

- globalement (toutes les p variable)
- 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).

Ecriture matricielle du modèle

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \cdots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$y = X\beta + \varepsilon$$

La méthode des moindres carrés donne pour résultat :

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

$$E(\hat{\beta}) = \beta \qquad V(\hat{\beta}) = \sigma^2 \left(X^T X\right)^{-1}$$

Le principale du test et l'intervalle de confiance sont les même comme dans le cas du modèle linéaire simple pour chaque paramètre β_i quelconque.

II – Les paramètres du modèle :

Estimateur des moindres carrés ordinaires

Pour trouver les paramètres «β_i» qui minimise S :

$$S = \varepsilon'\varepsilon = \sum_i \varepsilon_i^2 = \sum_i [y_i - (\beta_0 + \beta_{i,1} x_1 + \ldots + \beta_{i,p} x_p)]^2$$

On doit résoudre

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

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

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

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

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

II – Les paramètres du modèle :

Commentaires

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

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

(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

II – Les paramètres du modèle :

Exemple :

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

Variable à prédire
Attribut classe
Variable endogène

Quantitative

II – Les paramètres du modèle :

Exemple des cigarettes :

constante	TAR (mg)	ICOTINE (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

(X'X)

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

(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

X'Y

289.7
3742.85
264.076
281.14508

$\hat{\beta}$

-0.55170	constante
0.88758	tar
0.51847	nicotine
2.07934	weight

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

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

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

$R^2 \# 1$, le modèle est parfait
$R^2 \# 0$, le modèle est mauvais

Évaluation globale de la régression
Tableau d'analyse de variance et Coefficient de détermination

Exemple des cigarettes :

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

R^2

SCR

SCE