MESURES D'ASSOCIATION ENTRE DEUX VARIABLES

- Covariance

  (1) population : $\sigma_{xy} = \frac{1}{n}\sum_i (x_i - \mu_x)(y_i - \mu_y)$

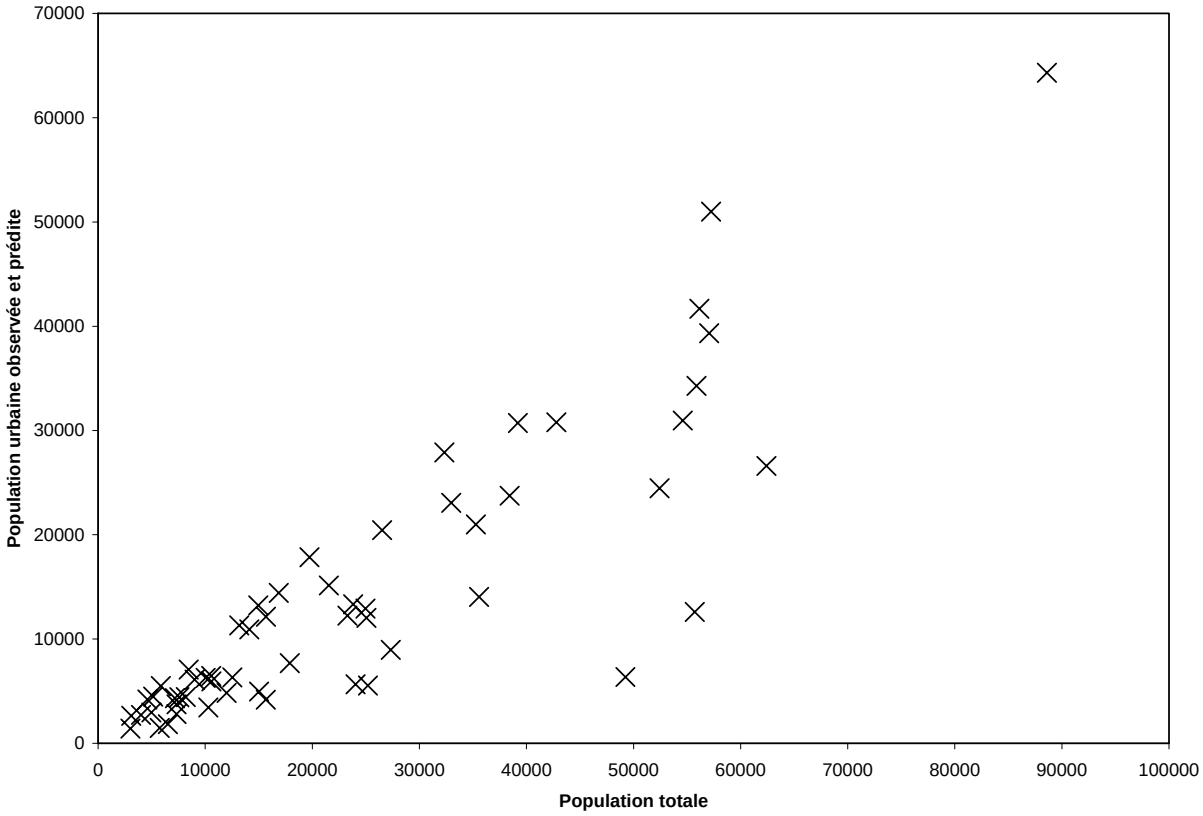
  (2) échantillon : $s_{xy} = \frac{1}{n-1}\sum_i (x_i - m_x)(y_i - m_y)$

- Coefficient de corrélation simple

  (1) population : $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$, avec $-1 < \rho < +1$

  (2) échantillon : $r = \frac{s_{xy}}{s_x s_y}$, avec $-1 < r < +1$

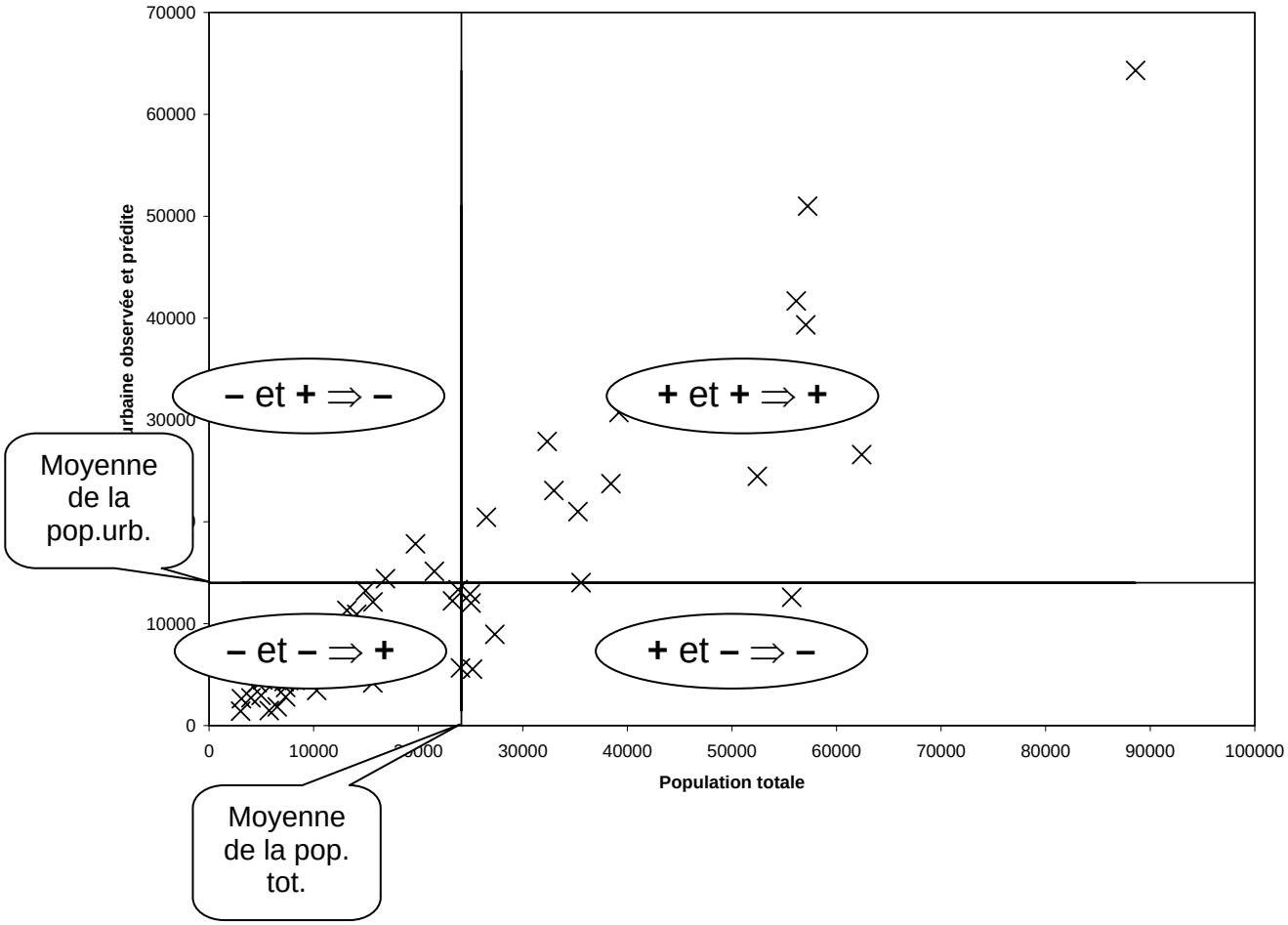
EXEMPLE DE DIAGRAMME DE DISPERSION POPULATION URBAINE ET POPULATION TOTALE (PAYS DE MOINS DE 100 MILLIONS D'HAB. EN 1990)



Source : Lemelin et Polèse (1995) et Primate4.xls, onglet « PTOT<100 (6) »

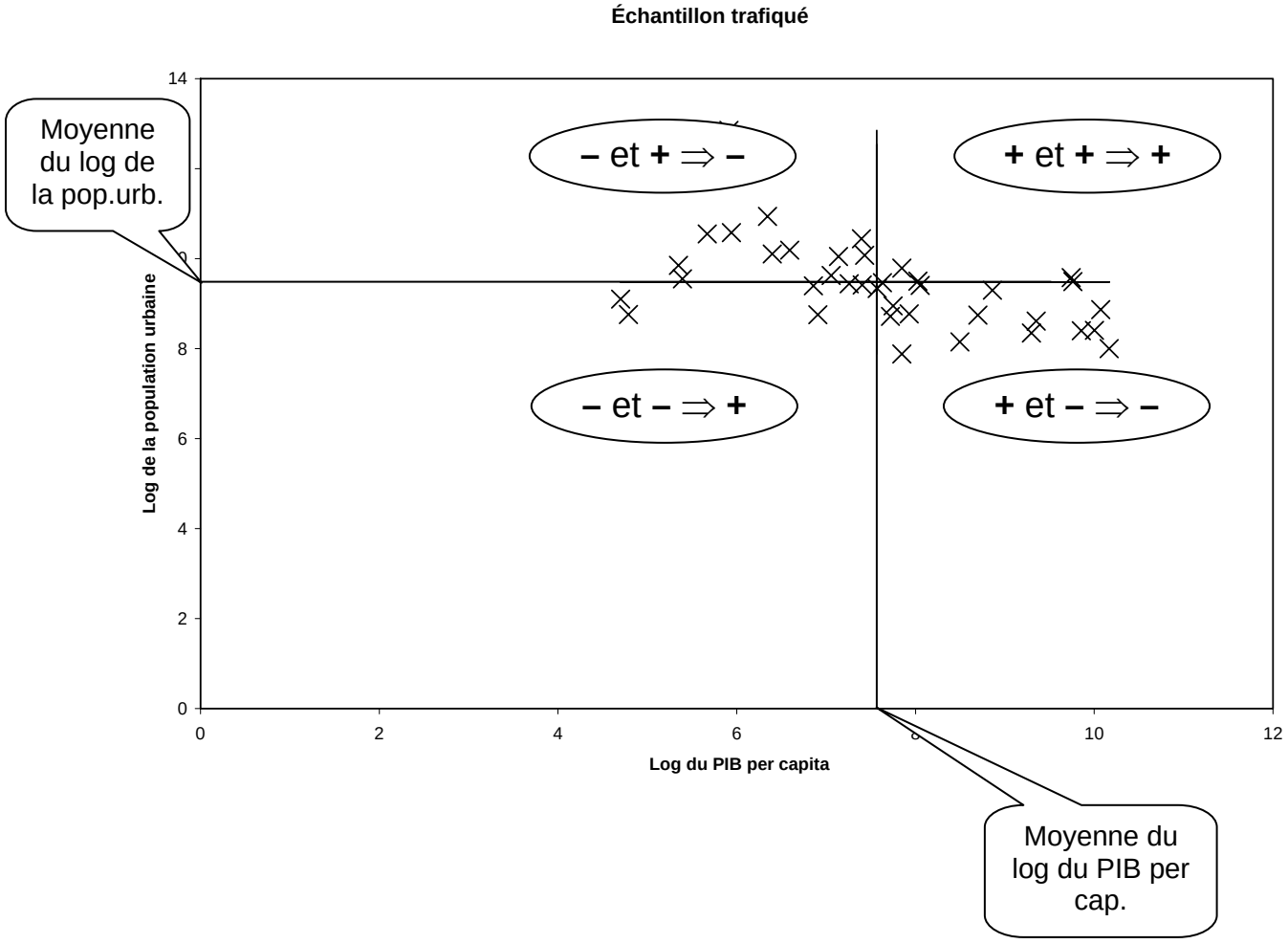
LA COVARIANCE : POURQUOI CETTE FORMULE ?

UN EXEMPLE DE CORRÉLATION POSITIVE



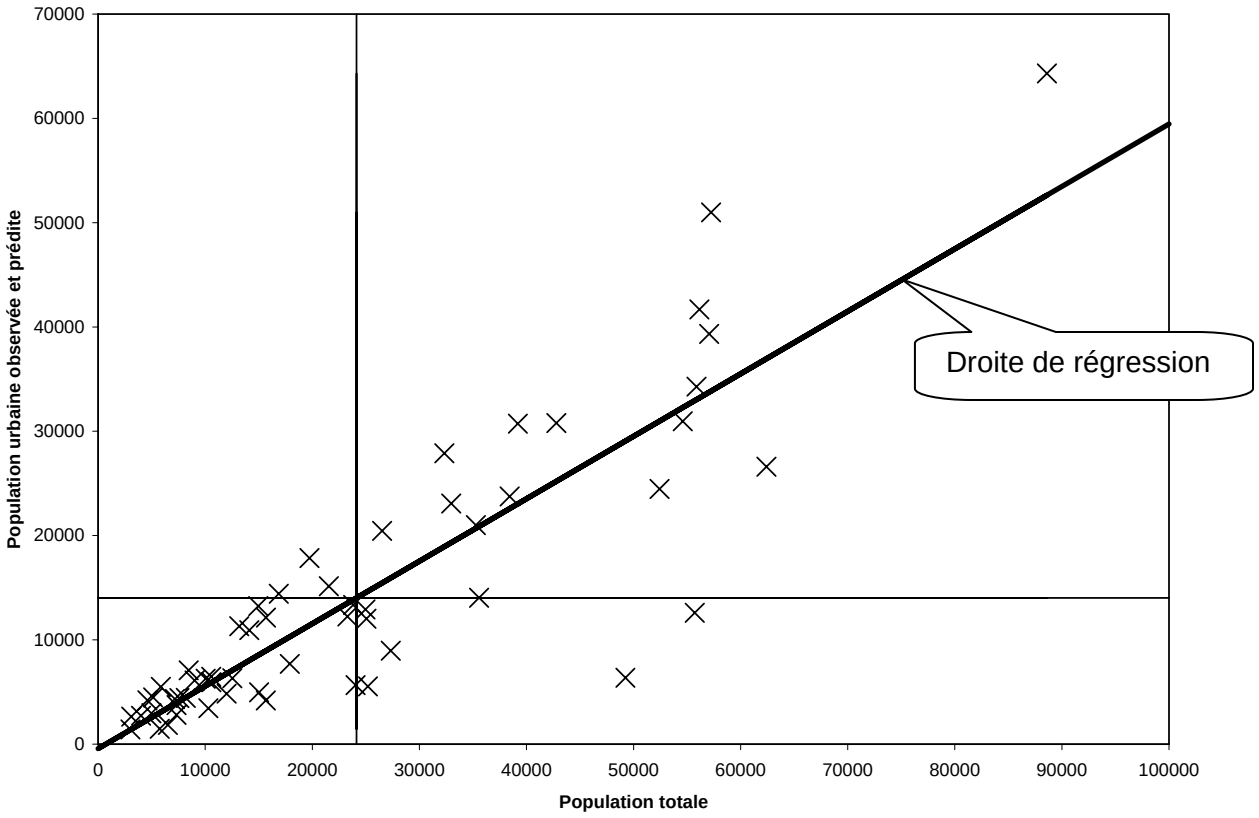
Source : Lemelin et Polèse (1995) et Primate4.xls, onglet « PTOT<100 (5) »

LA COVARIANCE : POURQUOI CETTE FORMULE ?
UN EXEMPLE DE CORRÉLATION NÉGATIVE



Source : Lemelin et Polèse (1995) et Primate4.xls, onglet « Fig.corr.neg. »

RÉGRESSION SIMPLE
POPULATION URBAINE SUR POPULATION TOTALE
(PAYS DE MOINS DE 100 MILLIONS D'HAB. EN 1990)



Droite de régression

PURB = *a* + *b PTOT*

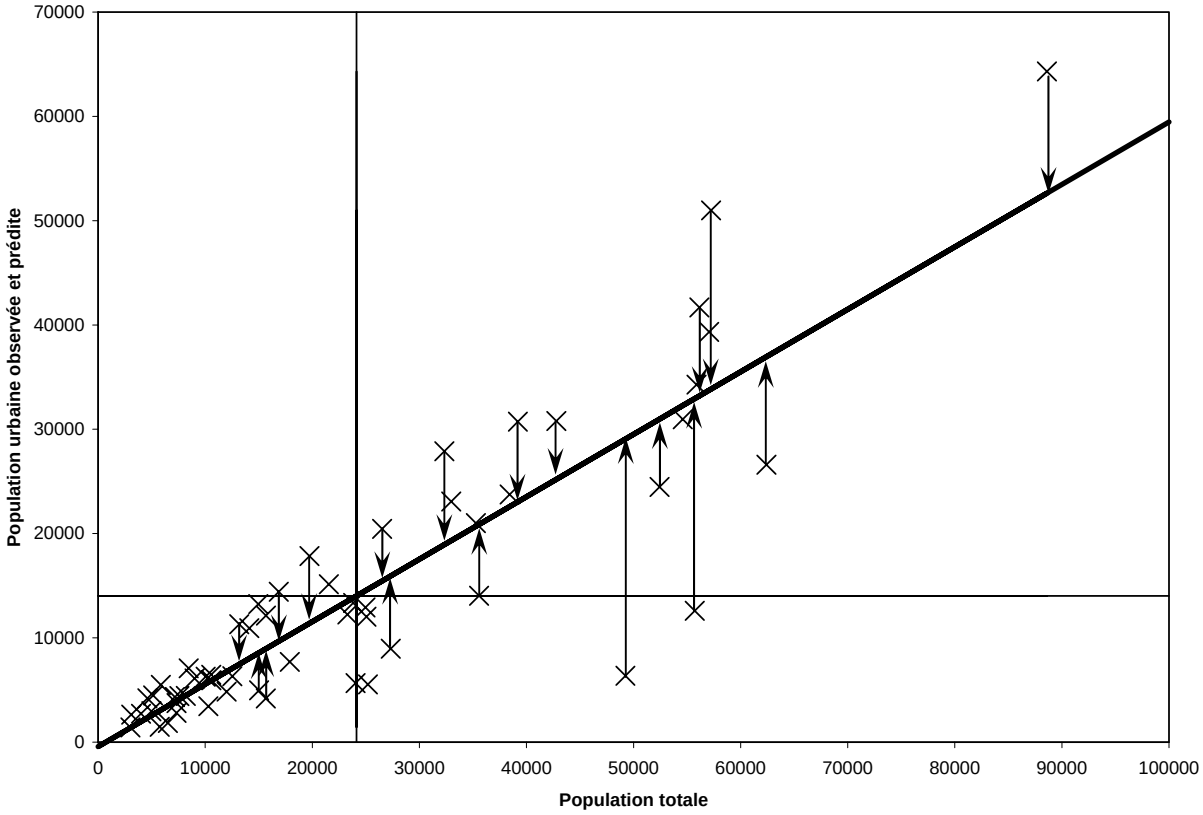
a = –417,6808443

b = 0,598875215

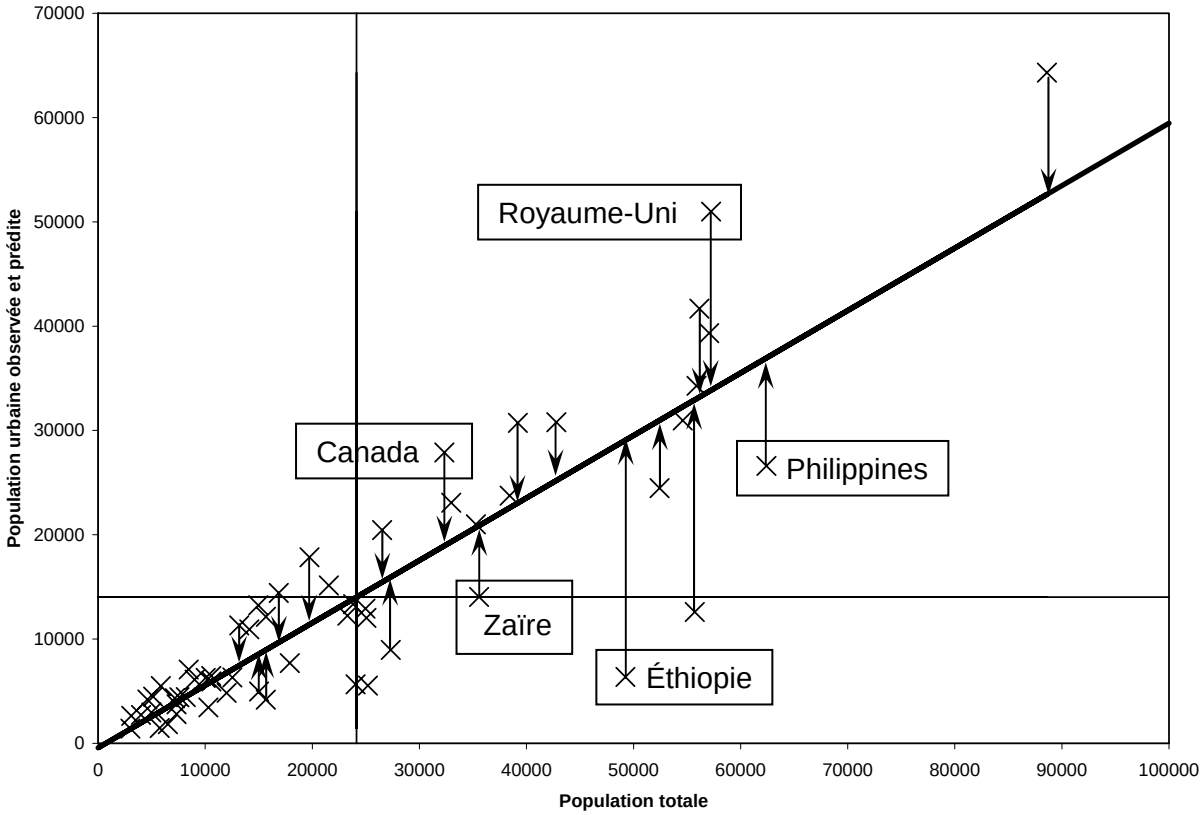
PURB = –417,6808443 + 0,598875215 * *PTOT*

Source : Lemelin et Polèse (1995) et Primate4.xls, onglet « PTOT<100 (5) »

RÉSIDUS DE LA RÉGRESSION SIMPLE
POPULATION URBAINE SUR POPULATION TOTALE
(PAYS DE MOINS DE 100 MILLIONS D'HAB. EN 1990)

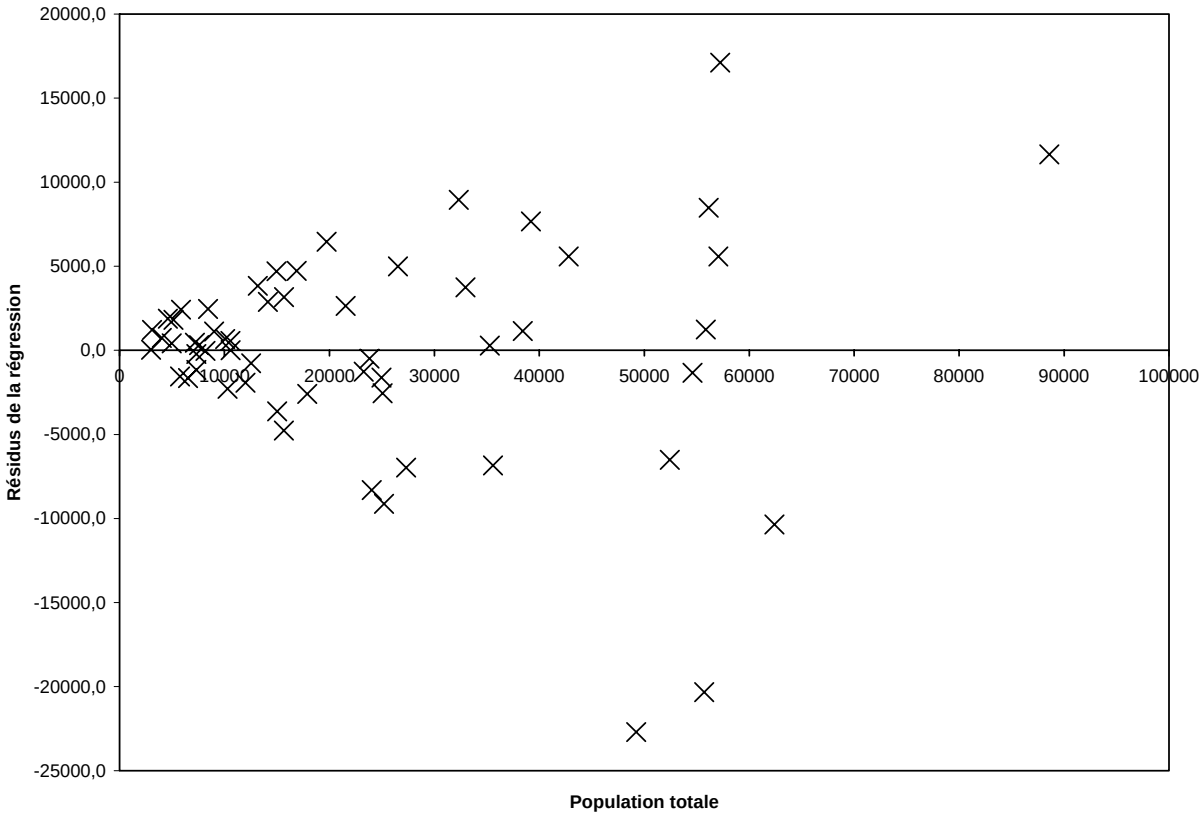


CALCUL DES RÉSIDUS POUR QUELQUES PAYS



PAYS	PURB	PTOT	Régression	Résidu
Canada	20458	26521	15465.1	4992.9
Zaïre	14043	35568	20883.1	−6840.1
Éthiopie	6367	49240	29070.9	−22703.9
Royaume-Uni	50980	57237	33860.1	17119.9
Philippines	26602	62413	36959.9	−10357.9

RÉSIDUS DE LA RÉGRESSION SIMPLE
POPULATION URBAINE SUR POPULATION TOTALE
(PAYS DE MOINS DE 100 MILLIONS D'HAB. EN 1990)



Source : Lemelin et Polèse (1995) et Primate4.xls, onglet « Résidus »

LA RÉGRESSION LINÉAIRE
PARMI LES MÉTHODES D'ANALYSE MULTIVARIÉE

Une classification de quelques méthodes d'analyse multivariée

<table>
<tr><th colspan="2">Variable dépendante</th><th>Variables indépendantes</th><th colspan="2">Méthode</th></tr>
<tr><td colspan="2" rowspan="2">Aucune</td><td>2 variables catégoriques</td><td rowspan="2">Analyse de table de contingence</td><td>... à 2 dimensions</td></tr>
<tr><td>Plus de 2 var. catégo.</td><td>... à plus de 2 dimensions</td></tr>
<tr><td colspan="2" rowspan="2">Continue</td><td>Discrètes (catégoriques)</td><td colspan="2">Analyse de variance
OU
Régression multiple</td></tr>
<tr><td>Continues ou discrètes</td><td colspan="2">Régression multiple</td></tr>
<tr><td rowspan="2">Catégorique</td><td>2 catégories</td><td rowspan="2">Continues ou discrètes</td><td rowspan="2">Logit ou probit</td><td>... binomial</td></tr>
<tr><td>Plus de 2 cat.</td><td>... multinomial</td></tr>
</table>

L'analyse de régression =

une méthode d'analyse des données...

qui s'applique lorsque...

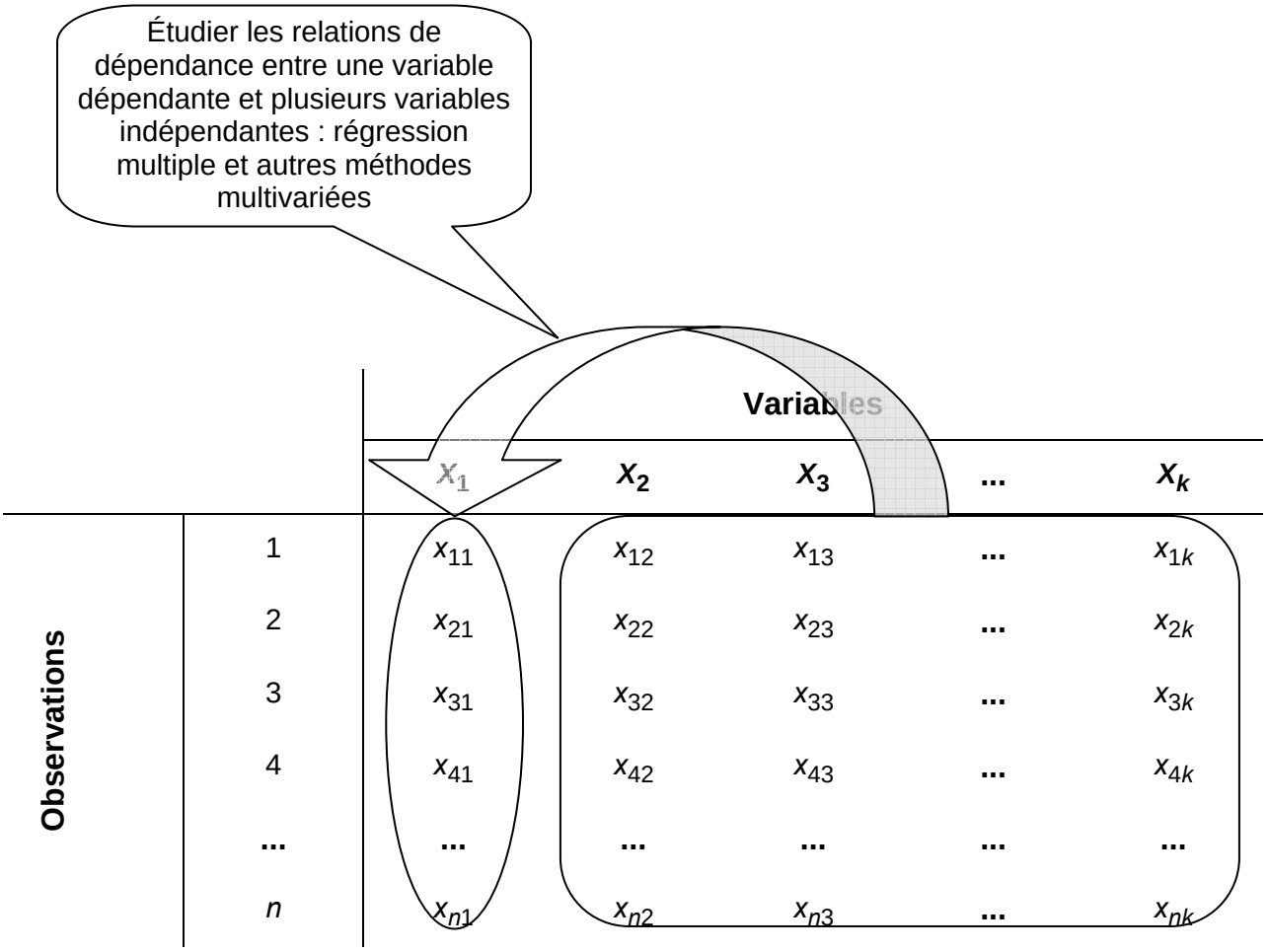
le modèle théorique sur lequel s'appuie l'analyse...

est formalisé par une relation entre...

une variable dépendante continue et...

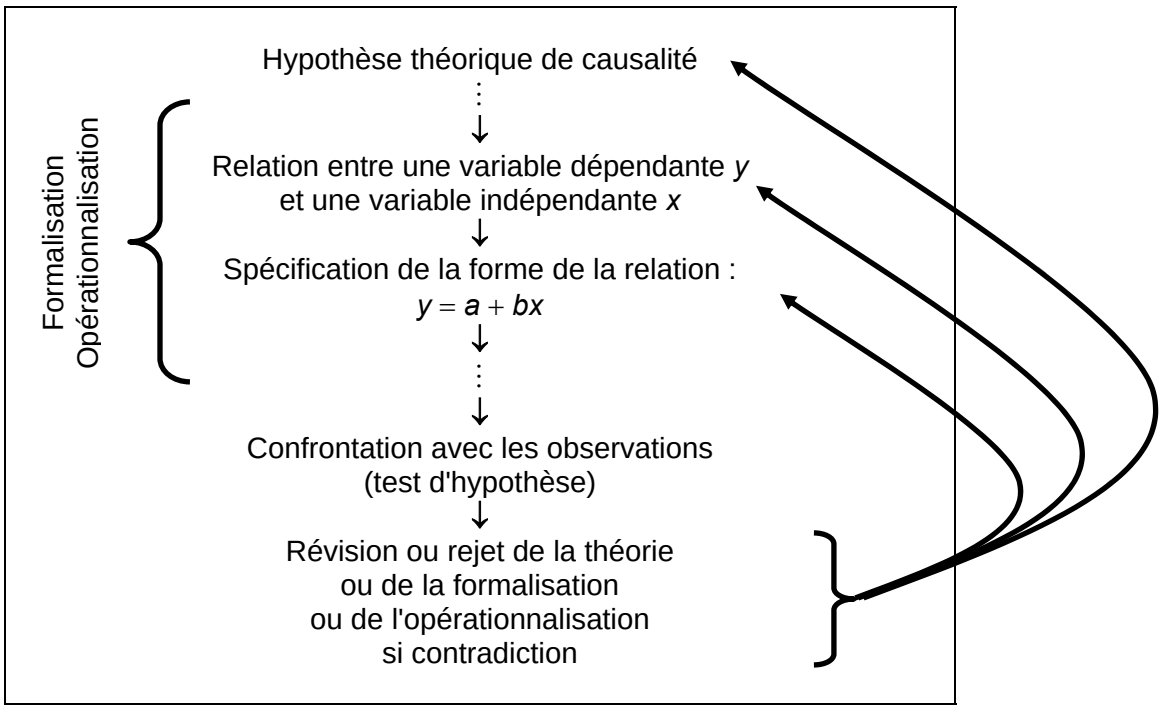
une ou plusieurs variables indépendantes discrètes ou continues.

STRUCTURE DES DONNÉES (5)
POINT DE VUE HORIZONTAL : DÉPENDANCE ENTRE UNE VARIABLE DÉPENDANTE ET PLUSIEURS VARIABLES INDÉPENDANTES



		x_1	x_2	x_3	...	x_k
Observations	1	x_{11}	x_{12}	x_{13}	...	x_{1k}
	2	x_{21}	x_{22}	x_{23}	...	x_{2k}
	3	x_{31}	x_{32}	x_{33}	...	x_{3k}
	4	x_{41}	x_{42}	x_{43}	...	x_{4k}
	...	...	...	...	...	...
	n	x_{n1}	x_{n2}	x_{n3}	...	x_{nk}

CAUSALITÉ ET VARIABLES DÉPENDANTES ET INDÉPENDANTES



UN MODÈLE LINÉAIRE (TROP) SIMPLE

Modèle

$$PLAR_i = \beta_1 + \beta_2\, PTOT_i + \beta_3\, GNPC_i$$

Valeurs des paramètres

$\beta_1 = 3500$

$\beta_2 = 0{,}01$

$\beta_3 = 0{,}1$

Valeurs des variables (1990)

		PLAR ('000)	*PTOT* ('000)	*PURB* ('000)	*GNPC* ($ US)
7 Brazil	Sao Paulo	17395	150368	112643	2680
13 Costa Rica	San Jose CR	1016	3015	1420	1900

Prédictions du modèle

Sao Paulo : $3500 + 0{,}01 \times 150368 + 0{,}1 \times 2680 = 5272$

San José, CR : $3500 + 0{,}01 \times 3015 + 0{,}1 \times 1900 = 3720$

LA CONSTANTE DU MODÈLE

Le **paramètre β_1** est la constante du modèle.

La **variable associée** à ce paramètre est une constante

Valeurs des variables (avec constante explicite)

		Constante	*PLAR* ('000)	*PTOT* ('000)	*PURB* ('000)	*GNPC* ($ US)
7 Brazil	Sao Paulo	1	17395	150368	112643	2680
13 Costa Rica	San Jose CR	1	1016	3015	1420	1900

Prédictions du modèle

Sao Paulo : $3500 \times 1 + 0{,}01 \times 150368 + 0{,}1 \times 2680 = 5272$

San José, CR : $3500 \times 1 + 0{,}01 \times 3015 + 0{,}1 \times 1900 = 3720$

MODÈLE LINÉAIRE GÉNÉRAL
(À UNE SEULE VARIABLE DÉPENDANTE)

Forme conceptuelle d'un modèle déterministe

« La valeur de la variable y est déterminée
par les valeurs des variables x_1, x_2, ... et x_k »

Forme mathématique générale du modèle déterministe

$$y_i = f(x_{i1}, x_{i2}, \ldots, x_{ik})$$

Forme générale du modèle linéaire déterministe

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \sum_{j=1}^{k} \beta_j x_{ij}$$

Variable dépendante = y_i

Variables indépendantes = x_{i1}, x_{i2}, ... et x_{ik}

Paramètres (coefficients) = β_1, β_2, ... et β_k

Écriture du modèle linéaire déterministe avec constante

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \sum_{j=1}^{k} \beta_j x_{ij}$$

... où l'on remarque qu'il n'est pas nécessaire d'écrire explicitement x_{i1}.

LA CONSTANTE DU MODÈLE

Un modèle linéaire avec une constante pourrait s'écrire

$$y_i = \alpha + \sum_{j=1}^{g} \gamma_j z_{ij}$$

Dans la forme générale, il y a une variable qui correspond à chaque coefficient (paramètre). Quelle variable pourrait correspondre à la constante α ?

Cette variable, appelons-la z_0, doit être égale à 1 pour toutes les observations :

$z_{i0} = 1$ pour tout i

Ainsi,

$$y_i = \alpha z_{i0} + \sum_{j=1}^{g} \gamma_j z_{ij} = \alpha + \sum_{j=1}^{g} \gamma_j z_{ij}$$

La variable z_0 qui correspond à la constante est donc une constante, qui a la même valeur dans toutes les observations.

Mais ce modèle est en réalité identique au modèle linéaire général. Posons

$k = g + 1$

$\beta_1 = \alpha$

$\beta_{j+1} = \gamma_j$ pour $j = 1, ..., g$

$x_{i1} = z_{i0}$

$x_{i,j+1} = z_{ij}$ pour $j = 1, ..., g$

Alors,

$$y_i = \alpha z_{i0} + \sum_{j=1}^{g} \gamma_j z_{ij} = \beta_1 x_{i1} + \sum_{j=2}^{k} \beta_j x_{ij} = \sum_{j=1}^{k} \beta_j x_{ij}$$

LA REPRÉSENTATION DES RELATIONS NON LINÉAIRES DANS LE MODÈLE LINÉAIRE

Exemple 1.0 : la transformation logarithmique

$PLAR_i = K\ PURB_i^h$

$\ln PLAR_i = \ln K + h \ln PURB_i$

Variable dépendante : ln *PLAR* Variables indépendantes : constante et ln *PURB* Paramètres : ln *K* et *h*

ln *K* = 2,067

h = 0,636

		ln *PLAR* ('000)	ln *PURB* ('000)
7 Brazil	Sao Paulo	9,76	11,63
13 Costa Rica	San Jose CR	6,92	7,26

Sao Paulo : $\text{EXP}(2{,}067 + 0{,}636 \times 11{,}63) = \text{EXP}(9{,}46) = 12883$

San José, CR : $\text{EXP}(2{,}067 + 0{,}636 \times 7{,}26) = \text{EXP}(6{,}68) = 800$

Exemple 1.1 : une courbe de tendance temporelle

$y_t = y_0\ (1+r)^t$

$\log y_t = \log y_0 + t \log(1+r)$

Variable dépendante : log y_t Variables indépendantes : constante et *t* Paramètres : log y_0 et log(1+*r*)

Exemple 1.2 : une fonction de production Cobb-Douglas

$Y_i = A\ K_i^B\ T_i^C$

Y : quantité produite *K* : quantité de capital utilisée *T* : quantité de main-d'oeuvre utilisée *A*, *B* et *C* sont des paramètres.

$\log Y_i = \log A + B \log K_i + C \log T_i$

Variable dépendante : log Y_i Variables indépendantes : constante, log K_i et log T_i Paramètres : log *A*, *B* et *C*

LA REPRÉSENTATION DES RELATIONS NON LINÉAIRES DANS LE MODÈLE LINÉAIRE

Exemple 2.0 : l'ajout de variables indépendantes

$$\ln PURB_i = a + b \ln PTOT_i + c \ln GNPC_i + d (\ln GNPC_i)^2$$

Variables indépendantes :

- constante
- $\ln PTOT_i$
- $\ln GNPC_i$
- $(\ln GNPC_i)^2$

Exemple 2.1 : relation cubique

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \beta_4 x_i^3$$

Variables indépendantes :

$z_{i1} = 1$ (constante)

$z_{i2} = x_i$

$z_{i3} = x_i^2$

$z_{i4} = x_i^3$

Relation linéaire :

$$y_i = \beta_1 + \beta_2 z_{i2} + \beta_3 z_{i3} + \beta_4 z_{i4} = \sum_{j=1}^{4} \beta_j z_{ij}$$

Exemple 2.2 : surface de tendance

$$Z_i = \beta_0 + X_i \beta_1 + Y_i \beta_2 + X_i^2 \beta_3 + Y_i^2 \beta_4 + X_i Y_i \beta_5$$

Variables indépendantes :

- constante
- X_i
- Y_i
- X_i^2
- Y_i^2
- $X_i Y_i$

CE SERAIT TROP BEAU !

Un modèle exact à deux paramètres :

$$\ln PLAR_i = \ln K + h \ln PURB_i$$

Deux équations linéaires à deux inconnues :

9,76 = ln K + 11,63 h (Brésil)

6,92 = ln K + 7,26 h (Costa Rica)

Solution :

ln K = 2,20

h = 0,65

Toronto, déviante ?

		ln *PLAR* ('000)	ln *PURB* ('000)
7 Brazil	Sao Paulo	9,76	11,63
9 Canada	Toronto	8,15	9,93
13 Costa Rica	San Jose CR	6,92	7,26

$$\ln PLAR = 2{,}20 + 0{,}65 \ln PURB = 2{,}20 + 0{,}65 \times 9{,}93 = 8{,}65 \neq 8{,}15$$

Trois équations, deux inconnues, pas de solution :

9,76 = ln K + 11,63 h (Brésil)

8,15 = ln K + 9,93 h (Can.)

6,92 = ln K + 7,26 h (Costa Rica)

Le modèle est une approximation

Le modèle avec un terme d'erreur :

$$\ln PLAR_i = \ln K + h \ln PURB_i + u_i$$

Formulation générale :

$$y_i = \sum_{j=1}^{k} \beta_j x_{ij} + u_i$$

L'ESTIMATEUR DES MOINDRES CARRÉS ORDINAIRES

Notation

b_i : valeur estimée du paramètre β_i.

$\hat{y}_i \equiv \sum_j b_j x_{ij}$ (valeur de y_i «prédite»)

$e_i \equiv y_i - \hat{y}_i \equiv y_i - \sum_j b_j x_{ij}$ (résidu)

Principe

Minimiser la somme des carrés des résidus :

$$\sum_i (y_i - \hat{y}_i)^2 = \sum_i \left(y_i - \sum_j b_j \, x_{ij} \right)^2 = \sum_i e_i^2$$

C'est-à-dire :

Minimiser (le carré de) la distance euclidienne généralisée entre les y_i et les $\hat{y}_i$.

Propriétés :

1. Estimateur linéaire

$$b_j = \sum_i w_{ji} \, y_i$$

2. Somme des résidus nulle

Lorsque la régression comporte une constante,

$$\sum_i e_i = 0$$

3. Relation entre les moyennes

Lorsque la régression comporte une constante,

$$m_y = m_{\hat{y}} = \sum_j b_j \, m_{x_j}$$

LE COEFFICIENT DE DÉTERMINATION MULTIPLE ET L'ANALYSE DE LA VARIANCE (1)

Première étape : décomposition de la variablilité

$$\sum_i (y_i - m_y)^2 = (n-1)\, s_y^2$$

Si le modèle comporte une constante,

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

Esquisse de démonstration :

$$\sum_i (y_i - m_y)^2 = (n-1)\, s_y^2$$

Si l'on développe le membre de gauche de cette expression, on obtient

$$\sum_i (y_i - m_y)^2 = \sum_i \left[(y_i - \hat{y}_i) + (\hat{y}_i - m_y)\right]^2$$

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

On peut montrer que, **si le modèle comporte une constante**, le dernier terme est nul [1], de sorte que

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

[1] Cette démonstration fait appel à l'écriture matricielle.

LE COEFFICIENT DE DÉTERMINATION MULTIPLE ET L'ANALYSE DE LA VARIANCE (2)

Deuxième étape : interprétation des éléments de la décomposition

Variabilité totale		Variabilité résiduelle		Variabilité «expliquée»
SST		*SSR*		*SSM*
Sum of Squares, Total	=	*Sum of Squares, Residuals*	+	*Sum of Squares, Model*
$\sum_i (y_i - m_y)^2 = (n-1)s_y^2$		$\sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2$		$\sum_i (\hat{y}_i - m_{\hat{y}})^2 = \sum_i (\hat{y}_i - m_y)^2 = (n-1)s_{\hat{y}}^2$

Troisième étape : construction d'une mesure d'ajustement («goodness of fit»)

$$R^2 = \frac{\text{Variabilité «expliquée»}}{\text{Variabilité totale}} = \frac{SSM}{SST}$$

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_i e_i^2}{(n-1)s_y^2}$$

DOMAINE DE VARIATION DU COEFFICIENT DE DÉTERMINATION MULTIPLE (VALEURS EXTRÊMES)

Mathématiquement

1. $R^2 \geq 0$

Démonstration

$SST \geq 0$ (somme de carrés)

$SSM \geq 0$ (somme de carrés)

$R^2 = SSM / SST$ = (somme de carrés) / (somme de carrés) ≥ 0

2. $R^2 \leq 1$

Démonstration

$SST \geq 0$ (somme de carrés)

$SSM \geq 0$ (somme de carrés)

$SSR \geq 0$ (somme de carrés)

$\boxed{SSR + SSM = SST} \Rightarrow \boxed{SSM \leq SST} \Rightarrow \boxed{R^2 = SSM / SST \leq 1}$

Dans quelles circonstances ?

3. Le modèle reproduit parfaitement les observations

   $\boxed{SSR = 0} \Rightarrow \boxed{R^2 = 1}$

4. Il n'y a pas de relation détectable entre la variable dépendante et les variables indépendantes. Alors les coefficients estimés par la méthode des moindres carrés sont tous nuls, à l'exception de la constante :

   $\boxed{\hat{y}_i = b_1 = m_y \text{ pour tout } i} \Rightarrow \boxed{SSM = \sum_i \left(\hat{y}_i - m_y\right)^2 = 0} \Rightarrow \boxed{R^2 = 0}$

LE COEFFICIENT DE DÉTERMINATION MULTIPLE (COMPLÉMENTS)

R^2 : une mesure de similarité

$$SSR = \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2$$

= carré de la distance euclidienne généralisée

= mesure de dissimilarité

SSR / SST : mesure de dissimilarité *normée* ($0 < SSR / SST < 1$)

$R^2 = 1 - SSR / SST$: mesure de similarité normée ($0 < R^2 < 1$)

Relation entre R^2 et le coefficient de corrélation simple

$$r_{\hat{y}y}^2 = \left(\frac{s_{\hat{y}y}}{s_{\hat{y}} s_y} \right)^2 = R^2$$

Coefficient de détermination ajusté

$$\overline{R}^2 = 1 - \frac{n-1}{n-k}\left(1 - R^2\right) = 1 - \frac{SSR \big/ (n-k)}{SST \big/ (n-1)}$$

Quand on ajoute des variables indépendantes, qu'arrive-t-il à $SSR/(n-k)$?

Est-ce que le numérateur SSR diminue davantage, en proportion, que le dénominateur $(n-k)$?