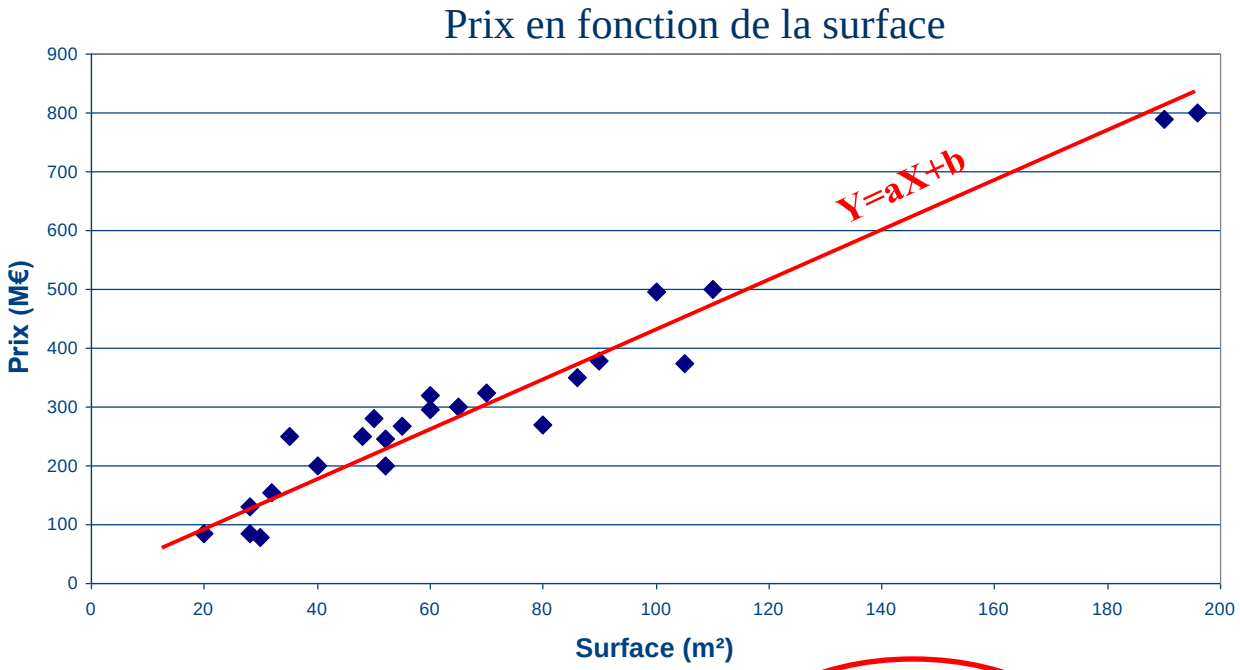
EXEMPLE INTRODUCTIF



Y = Prix

X = Surface

⇒ **$Y=aX+b+\varepsilon$**

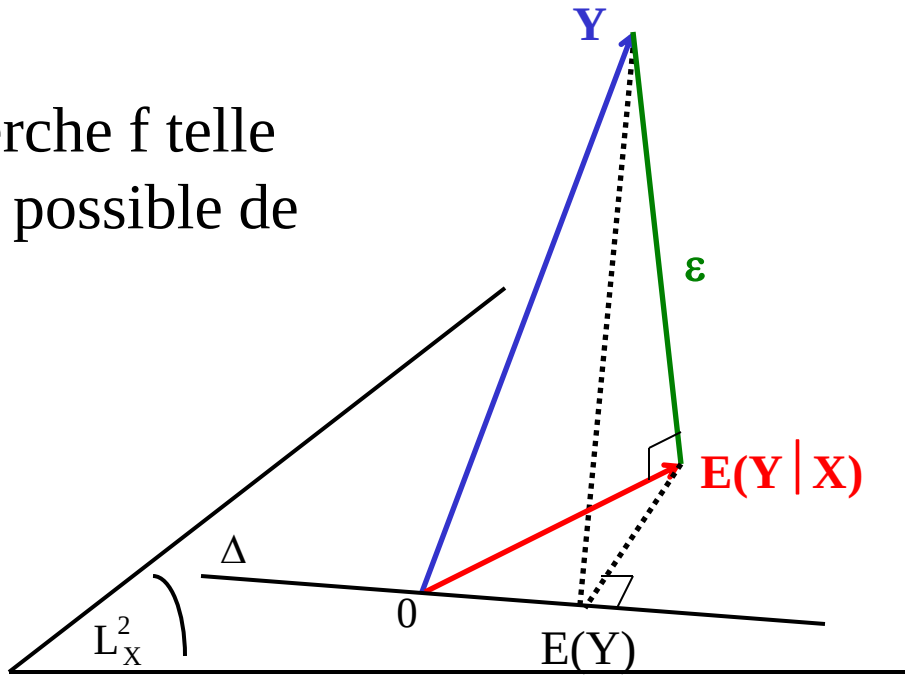
$Y, X_1, X_2, \ldots, X_n$ ⇒ **$Y=a_1X_1+a_2X_2+\ldots+a_nX_n+b+\varepsilon$**

Y: variable expliquée **$X_1,\ldots,X_n$: variables explicatives**

Prix (M €)	Surface (m²)
130	28
280	50
800	196
268	55
790	190
500	110
320	60
250	48
378	90
250	35
350	86
300	65
155	32
245	52
200	40
325	70
85	28
78	30
375	105
200	52
270	80
295	60
85	20
495	100

REGRESSION DE Y EN X

Soient X et Y deux v.a, on cherche f telle que f(X) soit aussi proche que possible de Y en moyenne quadratique



Solution : $f(X)=E(Y \mid X)$

Modèle : $Y=E(Y \mid X)+\varepsilon$

ε est un résidu aléatoire tel que :

- $E(\varepsilon)=0$ car $E(Y)=E[E(Y \mid X)]$
- ε et X non corrélés linéairement car $\varepsilon \perp L^2_X$
- $Var(\varepsilon)=\sigma^2$ (constante)

CAS OU LA REGRESSION EST LINEAIRE

On suppose que : $f(X)=E(Y \mid X)$ s'écrit sous la forme $aX+b$

Le modèle devient alors

$$Y=aX+b+\varepsilon$$

avec $E(\varepsilon)=0$ et $var(\varepsilon)=\sigma^2$

Comment estimer les paramètres inconnus a, b, σ^2 à partir de (x_i,y_i) n observations indépendantes du couple (X,Y) ?

Estimation des coefficients : $\hat{y} = \hat{a}x + \hat{b}$

$$\hat{a} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{cov(x,y)}{s_x^2} = r\frac{s_y}{s_x} \quad \text{et} \quad \hat{b} = \bar{y} - \hat{a}\bar{x}$$

r est le coefficient de corrélation linéaire empirique entre X et Y :

- Si r>0 alors X et Y varient dans le même sens
- Si r<0 alors X et Y varient en sens contraire

HYPOTHESES SUR LE MODELE

Exogénéité : La variable X est indépendante de l'erreur ε. En probabilité, cela s'écrit $E(\varepsilon / X) = 0$

Homoédasticité : L'erreur a une variance indépendante des observations. En probabilité, cela s'écrit $E(\varepsilon / X^i) = \sigma^2$ où X^i représente la variable X pour la $i^{ème}$ observation.

Non corrélation des erreurs : D'une observation à une autre les erreurs sont non corrélées.

Normalité des erreurs : Les erreurs suivent des lois normales centrées et de variances σ^2

ETUDE DES RESIDUS

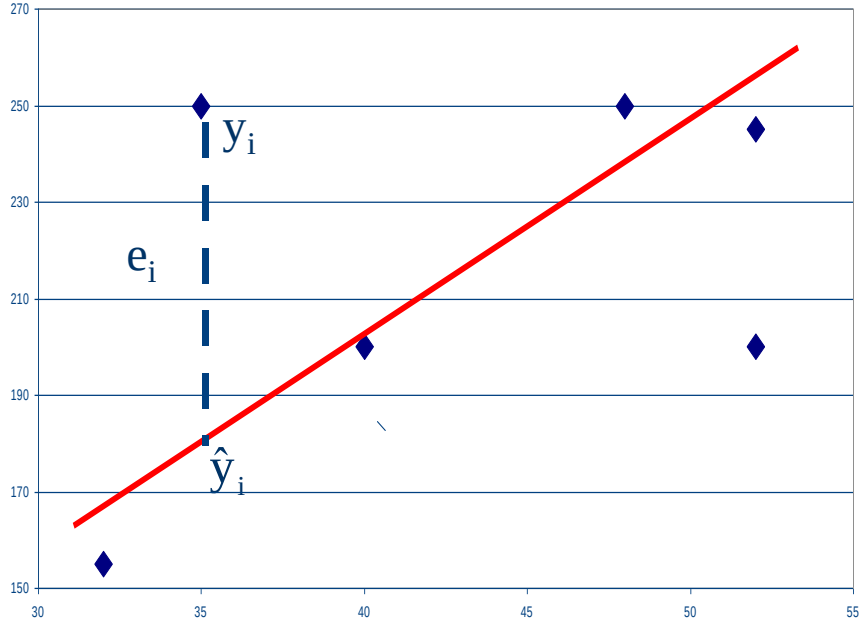
A chaque observation, on a

$$y_i = ax_i + b + \varepsilon_i \text{ et } \hat{y}_i = \hat{a}x_i + \hat{b}$$

Le résidu est donc représenté par les écarts résiduels :

$$e_i = y_i - \hat{y}_i$$

Prix en fonction de la surface



➢ Les écarts résiduels sont de moyenne nulle :

$$\sum_{i=1}^{n} e_i = 0$$

➢ La variance des résidus σ^2 est alors estimée sans biais par :

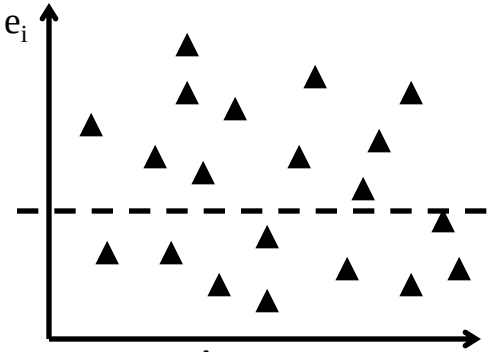
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$$

➢ Les résidus sont non corrélés :

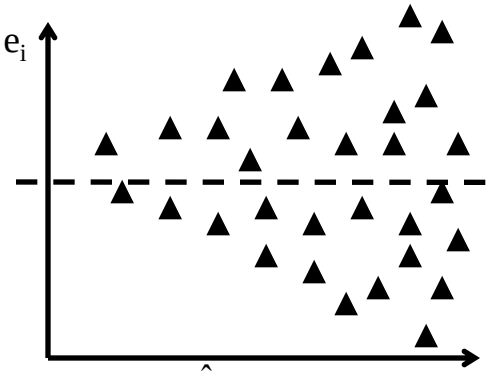
- Représentation graphique des écarts résiduels

- Test statistique d'indépendance

DETECTION DES NON LINEARITES

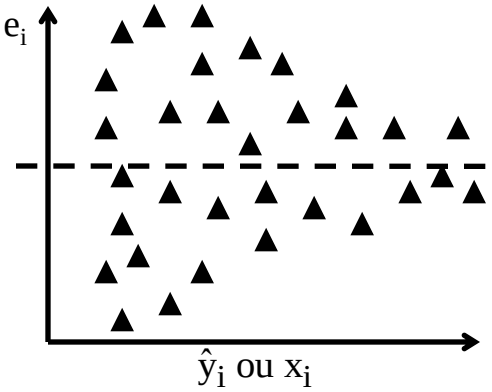


Régression normale



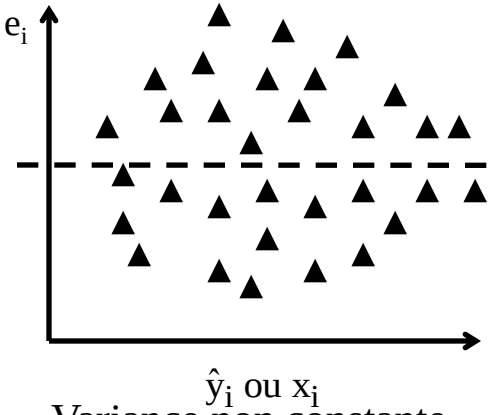
Variance non constante
(grandes valeurs de y)

$y' = \sqrt{y}$, $y' = \log y$, $y' = \log(y+1)$



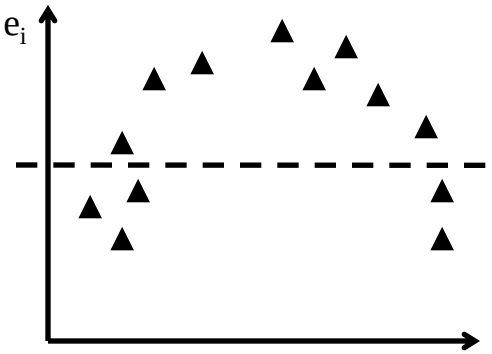
Variance non constante
(petites valeurs de y)

$y' = \frac{1}{y+1}$, $y' = \frac{1}{y}$

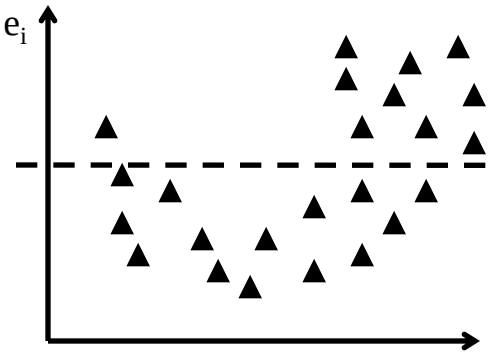


Variance non constante
(y pourcentage)

$y' = \log y$, $y' = \log(y+1)$



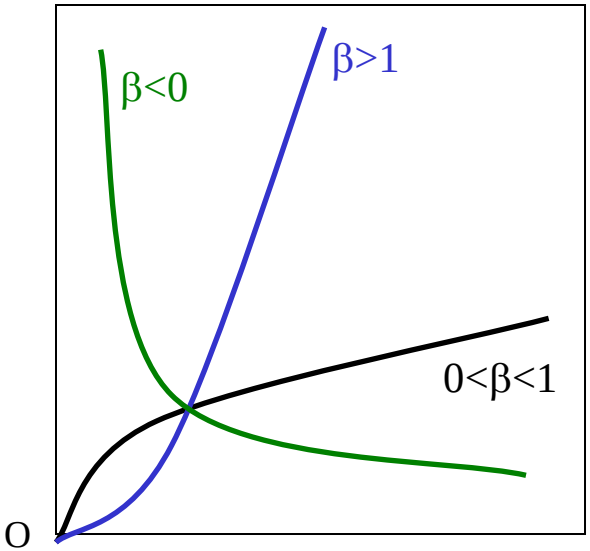
Résidus corrélés



Résidus corrélés

LINEARISATION DES DONNEES (1/2)

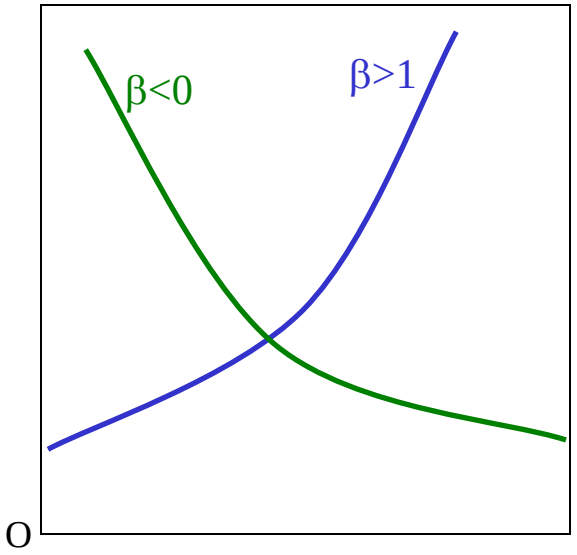
Dans le cas où les données ne présentent pas de relation linéaire, la régression linéaire n'a plus lieu d'être sauf dans certains cas particuliers où les données peuvent être linéarisées



Fonction : $y=\alpha x^{\beta}$

Transformations :
$y'=\log(y)$, $x'=\log(x)$

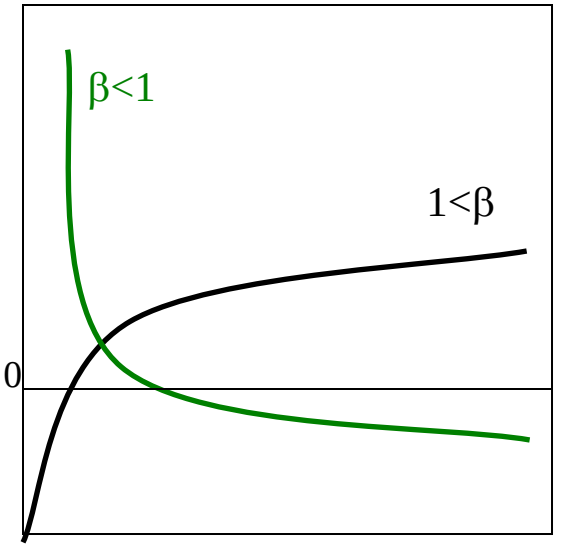
Forme linéaire :
$y'=\log(\alpha)+\beta x'$



Fonction : $y=\alpha e^{\beta x}$

Transformations :
$y'=\log(y)$

Forme linéaire :
$y'=\log(\alpha)+\beta x$

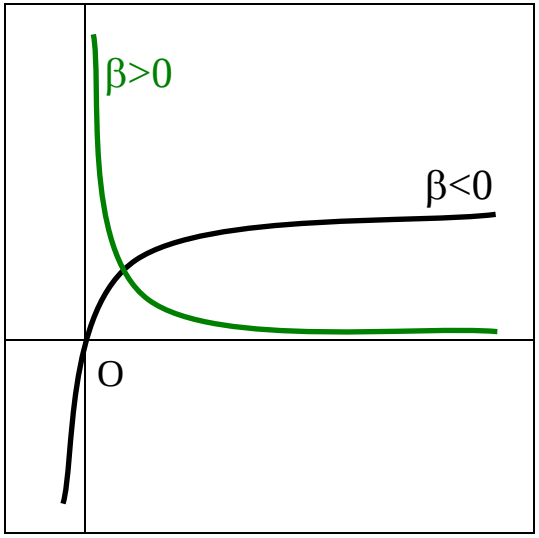


Fonction : $y=\alpha+\beta\log x$

Transformations :
$x'=\log(x)$

Forme linéaire :
$y'=\alpha+\beta x'$

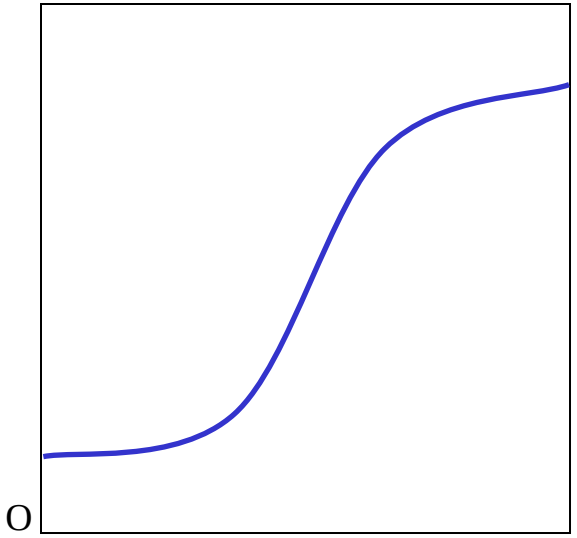
LINEARISATION DES DONNEES (2/2)



Fonction : $y = \frac{x}{\alpha x - \beta}$

Transformations :
y'=1/y, x'=1/x

Forme linéaire :
y'=α-βx'



Fonction : $y = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$

Transformations : $y' = \log(\frac{y}{1-y})$

Forme linéaire :
y'=α+βx

COEFFICIENT DE CORRELATION LINEAIRE

Plus le coefficient de corrélation linéaire empirique,

$$r = \frac{cov(x,y)}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

est proche de 1 ou –1, plus le modèle linéaire est bon. En général, on calcule $\mathbf{r^2 \approx 1}$. De plus, si on considère que r est la réalisation d'une v.a. R et que les hypothèses du modèle sont vérifiées,

alors $F = \frac{R^2}{1-R^2}(n-2)$ suit une loi de Fisher $F_\alpha(1;n-2)$

On peut alors tester l'hypothèse de linéarité **H_0 : « r=0 ».** Si cette hypothèse est rejetée ($F<F_\alpha(1;n-2)$) alors on admet qu'il n'y a pas de relation linéaire entre X et Y.

Remarque : $\hat{a} = r\frac{s_y}{s_x}$ donc tester H_0 : « r=0 » revient à tester H_0 : « a=0 »

PREDICTION D'UNE VALEUR

Supposons que l'on souhaite prévoir à l'aide du modèle la valeur y_0 pour une valeur de x_0 non observée,

$$\hat{y}_0 = \hat{a}x_0 + \hat{b}$$

Soit Y_0 et $\hat{Y}_0$ les v.a. considérées, alors

$$\frac{Y_0 - \hat{Y}_0}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}}$$

suit une loi de Student t_{n-2}

Toutes les valeurs de l'expression ci-dessus sont connues sauf y_0, on peut donc en déduire un intervalle de confiance pour $\hat{Y}_0$

L'intervalle sera d'autant plus grand que x_0 sera éloigné de $\bar{x}$

EXEMPLE (suite)

A partir des données du tableau, on calcule : $\bar{x} = 70.08\ m^2$ et $\bar{y} = 309.33 \times 10^3\ €$

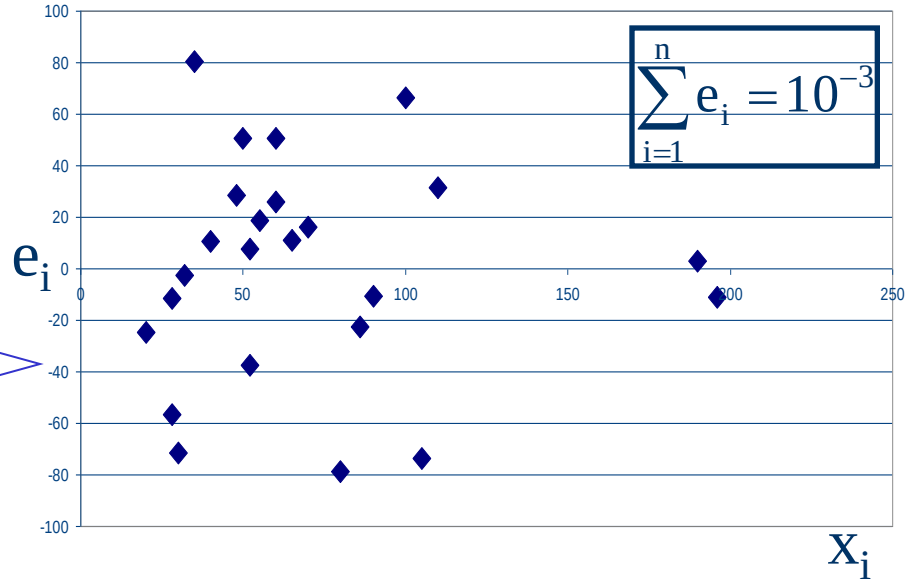
$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 44.69\ m^2$ et $s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} = 182.95 \times 10^3\ € \Rightarrow$ $r = 0.97$

On en déduit : $\hat{y} = 3.98x + 30.09$

Validité de la régression

- r^2=0.95 proche de 1
- $22r^2(1-r^2)=396 >> F_{1,22}=61.7$
- Etude des résidus

Ecarts résiduels en fonction de x_i



Prévision

Dans la table on a $P(|t_{22}| < 2.074) = 0.95$

Si on prend x_0=100m², on a $\hat{y}_{100} = 428.53\ €$. De plus, on a $\hat{\sigma} = 43.84$, d'où l'intervalle de confiance à 95%

$$\left|\frac{y_{100} - 428.53}{45.15}\right| < 2.074 \Rightarrow 334.89€ < y_{100} < 522.17€$$

REGRESSION MULTILINEAIRE

V.A A EXPLIQUER		V.A. EXPLICATIVES			
Y	Cste	X_1	X_2	...	x_p
y_1	1	x_{11}	x_{12}	...	x_{1p}
y_2	1	x_{21}	x_{22}	...	x_{2p}
.....	1				
y_n	1	x_{n1}	x_{n2}	...	x_{np}

$= \mathbf{y}$ *Vecteur des observations*

$= X$ *Matrice du modèle*

Modèle

$$Y = a_0 + a_1X_1 + a_2X_2 + ... + a_pX_p + \varepsilon$$

avec $E(\varepsilon)=0$ et $var(\varepsilon)=\sigma^2$ et X_i indptes

Ecriture matricielle

$$\mathbf{y} = X\mathbf{a} + \varepsilon$$

où $\mathbf{a} = {}^t(a_0\ a_1\ ...\ a_p)$ *Vecteur des coefficients du modèle*

De même que précédemment, on cherche $\hat{y}$ aussi proche possible de **y** , où

$$\hat{y} = \hat{a}_0 + \hat{a}_1x_1 + \hat{a}_2x_2 + ... + \hat{a}_px_p$$

Estimation des coefficients

$$\hat{\mathbf{a}} = ({}^tXX)^{-1}\,{}^tX\mathbf{y}$$

$$\Rightarrow \quad \hat{\mathbf{y}} = X\hat{\mathbf{a}} = X({}^tXX)^{-1}\,{}^tX\mathbf{y}$$

Estimation de la variance de ε

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|}{n\text{-}p\text{-}1}$$

HYPOTHESES SUR LE MODELE

Exogénéité : La variables X_j sont indépendantes de l'erreur ε

Homoédasticité : L'erreur a une variance indépendante des observations. En probabilité, cela s'écrit $E(\varepsilon / X^i) = \sigma^2$ où X^i représente le vecteur de variable X pour la $i^{ème}$ observation.

Non corrélation des erreurs : D'une observation à une autre les erreurs sont non corrélées.

Normalité des erreurs : Les erreurs suivent des lois normales centrées et de variances σ^2

ATTENTION : La validité des tests qui vont suivre suppose que ces hypothèses sont vérifiées. Il est de même pour les IDC.

TESTS DANS LE MODELE (1/2)

- Test de l'hypothèse de non-regression

$H_0 : a_1=a_2=\ldots=a_p=0$ (a_0 quelconque)

Coefficient de détermination : $R^2 = \dfrac{\Sigma(y_i-\bar{y}_i)^2 - \Sigma(y_i-\hat{y}_i)^2}{\Sigma(y_i-\bar{y}_i)^2}$

Variance expliquée par la régression

Variance totale

Si H_0 est vraie alors $F = \dfrac{R^2}{1-R^2}\dfrac{n-p-1}{p}$ suit une loi F(p,n-p-1)

Connaissant la valeur de R^2, on peut donc

- refuser H_0 si $\dfrac{R^2}{1-R^2}\dfrac{n-p-1}{p} \geq F(p;n-p-1)$

avec comme erreur de se tromper $P\left(\dfrac{R^2}{1-R^2}\dfrac{n-p-1}{p} \geq F(p;n-p-1)\right)$

- accepter H_0 sinon

Remarque : si p=1, on retrouve la loi du coefficient de corrélation usuel ρ.

TESTS DANS LE MODELE (2/2)

- Test du caractère significatif d'un coefficient de la régression

$$H_0 : a_j = 0$$

Variance de $\hat{a}_j$: $s_j^2 = \sigma^2 \left[({}^t XX)^{-1} \right]_{jj}$

Si l'hypothèse H_0 est vraie alors $\dfrac{|\hat{a}_j|}{s_j}$ suit une loi t_{n-p-1}

Connaissant la valeur de $\hat{a}_j$ et s_j on peut donc

- refuser H_0 si $\dfrac{|\hat{a}_j|}{s_j} \geq t_{n-p-1}$

avec comme erreur de se tromper $P\left(\dfrac{|\hat{a}_j|}{s_j} \geq t_{n-p-1}\right)$

- accepter H_0 sinon

PREVISION D'UNE VALEUR

Soit $\mathbf{x}_0$ le vecteur des valeurs des variables explicatives pour lesquelles on souhaite connaître y_0, alors

$$\frac{Y_0 - \hat{Y}_0}{\hat{\sigma}\sqrt{1 + {}^t x_0 ({}^t XX)^{-1} x_0}} = t_{n-p-1}$$

Ce qui permet de donner un intervalle de confiance pour y_0.

METHODOLOGIE

Modèle : $Y = a_0 + a_1X_1 + a_2X_2 + ... + a_pX_p + \varepsilon$

1. Si p=1 étude du nuage de points
2. Calcul des coefficients $\hat{a}_j$ et de leur variance $\text{var}(\hat{a}_j)$
3. Qualité de la régression
   - $R^2 \approx 1$
   - Test sur R^2 : au moins un coefficient significatif, *i.e*

$$\frac{R^2}{1-R^2}\frac{n-p-1}{p} \geq F(p; n-p-1)$$

Risque d'erreur α

   - Test sur chaque coefficient : aj significativement non nul ssi

$$\frac{|\hat{a}_j|}{\sqrt{\text{var}(\hat{a}_j)}} \geq t_{n-p-1}$$

4. Vérification hypothèses du modèle :
   - Etude des résidus : $e_i = y_i - \hat{y}_i$ (moyenne nulle, non corrélation, variance constante)
   - X_i indépendantes (matrice de corrélation)

SELECTION DES VARIABLES EXPLICATIVES

Test de tous les modèles impossible

Méthodes pas à pas :

élimination successive ou ajout successif de variables explicatives

• La méthode descendante consiste à éliminer la variable la moins significative parmi les p : celle qui a le *t* de Student le moins significatif. On recalcule alors la régression puis on recommence jusqu'à être satisfait.

• La méthode ascendante procède en sens inverse : On part de la meilleure régression à une seule variable puis on ajoute la variable la plus significative.

Analyse multivariée – Régression linéaire

Exemple

PAYS	DENSITE	ESPVIEF	LITTERATURE	FERTILITE
Arab.Saoud.	0.80	70.00	62.00	6.67
Argentine	1.20	75.00	95.00	2.80
Belgique	32.90	79.00	99.00	1.70
Bolivie	0.70	64.00	78.00	4.21
Canada	0.30	81.00	97.00	1.80
CoréeS.	44.70	74.00	96.00	1.65
France	10.50	82.00	99.00	1.80
GB	23.70	80.00	99.00	1.83
Indonésie	10.20	65.00	77.00	2.80
Iran	3.90	67.00	54.00	6.33
Liban	34.30	71.00	80.00	3.39
Roumanie	9.60	75.00	96.00	1.82
Russie	0.90	74.00	99.00	1.83
Sénégal	4.30	58.00	38.00	6.10
Somalie	1.00	55.00	24.00	7.25
Suède	1.90	81.00	99.00	2.10
Suisse	17.00	82.00	99.00	1.60
Syrie	7.40	68.00	64.00	6.65
Turquie	7.90	73.00	81.00	3.21
Ukraine	8.70	75.00	97.00	1.82
Vénézuela	2.20	76.00	88.00	3.05

Peut-on expliquer la fertilité par les variables densité, littérature et l'espérance de vie?

Résultat du modèle complet sous R

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.01405 -0.36801 -0.07267  0.28287  1.72874

Coefficients:
                Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)      7.81343   2.42653    3.220    0.00503 **
DENSITE         -0.01498   0.01361   -1.100    0.28648
ESPVIEF          0.05114   0.04995    1.024    0.32028
LITTERATURE -0.09779   0.01768   -5.533    3.65e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7356 on 17 degrees of freedom
Multiple R-Squared: 0.8841,     Adjusted R-squared: 0.8636
F-statistic: 43.22 on 3 and 17 DF,  p-value: 3.605e-08
```

Résultat du modèle complet sous Scilab

REGRESSION MULTIPLE
Nombre d'observations utilisées 21
Nombre de variables utilisées 4
Variable à expliquer FERTILITE
1e variable explicative DENSITE
2e variable explicative ESPVIEF
3e variable explicative LITTERATURE

MESURES GLOBALES DE LA QUALITÉ DE LA RÉGRESSION
Coefficient de corrélation multiple R 0.9403
Carré du coef de corrélation multiple R^2 .. 0.8841
Valeur du test F de signification de R^2 ... 43.22 à 3 et 17 ddl

---

COEFFICIENTS DE LA RÉGRESSION

	Coeff.	Err. Stand.	t_value
Constante	7.8134	2.4265	3.22
DENSITE	-0.0150	0.0136	-1.1
ESPVIEF	0.05114	0.0499	1.024
LITTERATURE	-0.0978	0.0177	-5.533

Sélection pas à pas

Première étape : Test des modèles à une variable explicative

Modèle	a_1 $V(a_1)$ t	a_2 $V(a_2)$ t	a_3 $V(a_3)$ t	b	R^2	d.d.l p,n-p-1	t-table	F-table	F-calculé
X_1 (Densité)	-0.06 0.03 -1.83			3.99	0.15	1;19		**4.38**	**3.34**
X_2 (Espvief)		-0.21 0.03 -5.82		18.54	0.64	1;19		**4.38**	**33.87** → **OK**
X_3 (Litt.)			-0.08 0.01 -11.16	10.28	0.87	1;19		**4.38**	**124.5** → **OK**

Meilleur score de X_3 ⇒ X_3 sélectionné

Sélection pas à pas

Deuxième étape : Test des modèles à deux variables explicatives

Modèle	a_1 $V(a_1)$ t	a_2 $V(a_2)$ t	a_3 $V(a_3)$ t	b	R^2	d.d.l p,n-p-1	t-table	F-table	F-calculé
X_1, X_3	-0.16 0.01 -1.17		-0.08 0.01 -10.31	10.21	0.74	2;18	1.73	3.55	64.14
X_2, X_3		0.05 0.05 1.09	-0.1 0.02 -5.84	7.7	0.87	2;18	1.73	3.55	63.48

NON

NON

OK
Au moins une var. explicative

Conclusion : Le modèle retenu est le modèle à une seule variable explicative : littérature