Régression Linéaire

Besnik PUMO

Agrocampus Ouest - INHP - ANGERS

7 février 2011

Introduction : Les objectifs

Les objectifs

1. Reconnaître les applications où le modèle de la Régression Linéaire peut être utilisée
2. Savoir ajuster un modèle à des données numériques : savoir choisir le modèle et estimer ces paramètres inconnus
3. Savoir évaluer la qualité du modèle ajusté sur des critères statistiques
4. Savoir réaliser le processus d'estimation numérique en utilisant le logiciel R et ces outils

Introduction : Modalités d'organisation

Contenu et Organisation

0. 2h TD : Rappels et logiciel R
1. La Régression Linéaire Simple : 2h C + 2h TD + 1h C + 2h TD (CC)
2. La Régression Linéaire Multiple (on se limitera dans le cas de deux variables) : 2h TD + 2h C + 2h TD (CC)

Evaluation

- RL = 1 ECTS (European Credits Transfert System)
- Examen écrit 1h avec documents + Contrôle Continu

Pédagogie

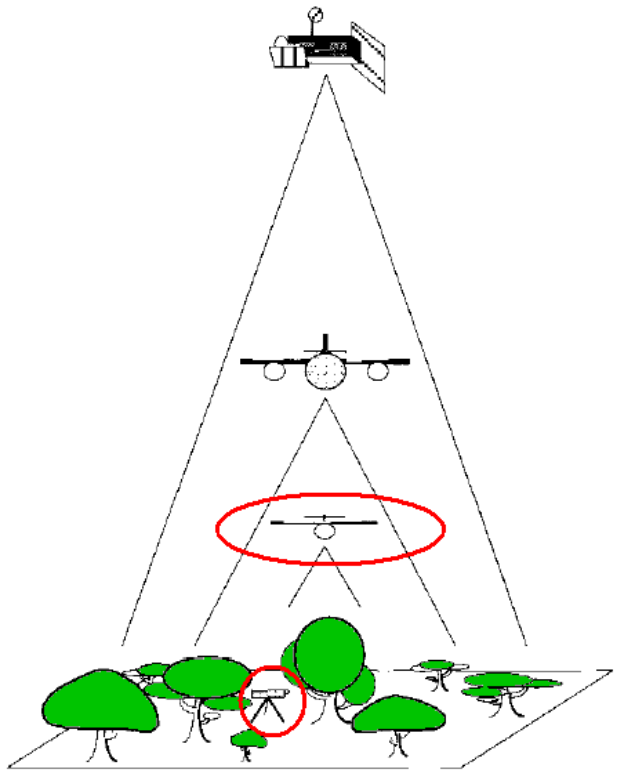
- Une plateforme e-learning : `http://tice.agrocampus-ouest.fr/`
- Des documents sur papier (polycopié du cours sur demande)
- Des TD/TP en salle informatique

La régression linéaire simple

1. Estimation de la biomasse ligneuse

Objectif.[1] Estimer les ressources ligneuses en région tropicale sèche.

Acquisition de données sur plusieurs plat-formes : investigation au sol, vol à basse altitude, couverture aérienne



Hypothèse On peut estimer la biomasse d'un arbre à partir de la surface terrière (1m du sol)

Etape 1 Mesure de la surface terriére (**ST**)

Mesure de la surface de la couronne (**SC**)

Etablissement du lien entre ST et SC

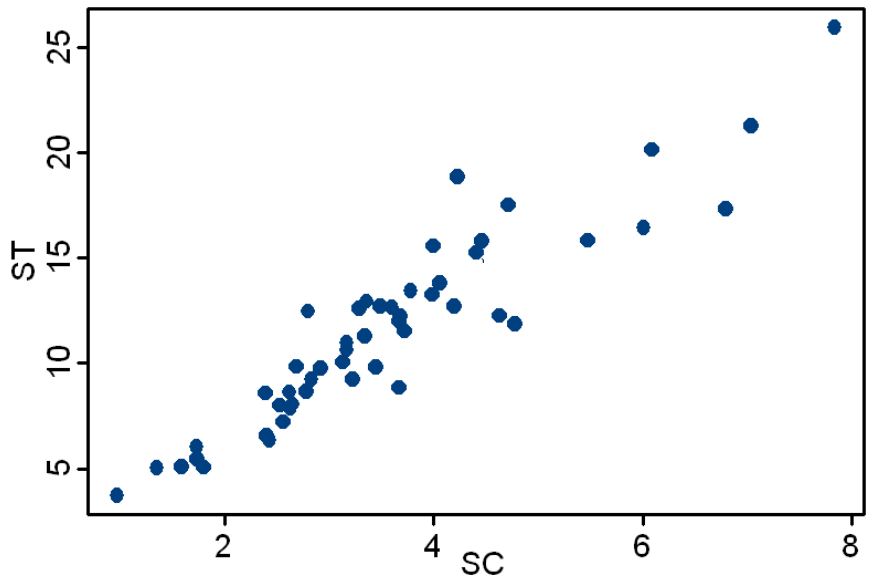
$$ST \approx a + b \cdot SC$$

Etape 2 Couverture aérienne (et traitement d'images) pour estimer la quantité de la végétation ligneuse à partir de la surface globale recouverte

1. Source : P. Defourny (1990) Méthode d'évaluation quantitative de la végétation ligneuse en région soudano-saharieenne, dans *T*élédétection et sécheresse, p. 63-74

Les données

Nr	ST	SC	Nr	ST	SC	Nr	ST	SC
1	6.34	2.43	18	12.46	2.80	35	21.27	7.04
2	6.03	1.73	19	20.15	6.09	36	9.24	3.23
3	12.61	3.29	20	13.43	3.78	37	8.66	2.78
4	17.34	6.80	21	3.73	0.97	38	9.22	2.83
5	15.25	4.41	22	9.80	3.45	39	7.99	2.53
6	15.57	4.00	23	5.46	1.74	40	6.55	2.40
7	15.80	4.46	24	12.24	3.68	41	12.71	3.49
8	12.93	3.36	25	16.44	6.01	42	10.97	3.17
9	13.82	4.06	26	18.86	4.23	43	9.83	2.69
10	11.28	3.34	27	12.25	4.63	44	5.02	1.35
11	12.00	3.67	28	17.52	4.72	45	7.88	2.63
12	11.52	3.72	29	5.07	1.59	46	8.57	2.39
13	8.05	2.65	30	11.88	4.78	47	9.75	2.92
14	5.05	1.80	31	12.65	3.60	48	13.27	3.99
15	10.63	3.17	32	8.63	2.62	49	10.05	3.13
16	15.83	5.48	33	25.94	7.84	50	8.85	3.67
17	12.71	4.20	34	7.22	2.56			



Notations

n - nombre d'individus (arbres) mesurés ($n = 50$ dans notre application)

x - la variable explicative - c'est une variable quantitative - notée SC dans notre application - On notera $x_1, \ldots, x_n$ les n mesures disponibles de la variable x

y - la variable à expliquer ou dépendante - c'est une variable quantitative - notée ST dans notre application - On notera $y_1, \ldots, y_n$ les n mesures disponible de la variable y

2. Le modèle et les postulats

Le modèle sous-jacent à la régression linéaire simple (RLS) est :

$$y = b_0 + b_1 x + e$$

- e désigne l'erreur ou l'écart entre la partie expliquée par le modèle et l'observation
- b_0, b_1 sont les paramètres inconnus du modèle
- ce modèle est
  - linéaire par rapport aux paramètres b_0, b_1 et
  - simple ou à une variable explicative

La Régression Linéaire Simple

Le modèle RLS permet en "choisissant convenablement" les paramètres b_0, b_1 d'ajuster ou de modéliser un nuage de points (x_i, y_i)

- Les valeurs des paramètres du modèle ajusté s'appellent des estimations de ces paramètres.
- Des conditions ou des postulats sont nécessaires pour pouvoir estimer les paramètres b_0, b_1

Les postulats

Soient donc $(x_1, y_1), \cdots, (x_1, y_1)$ n mesures.

Le postulat d'indépendance : On suppose que les y_i sont indépendantes

P0 $E(e_i) = 0$ pour tout i

l'erreur moyenne est nulle ; $E(y_i) = b_0 + b_1 x_i$ et $e_i = y_i - E(y_i)$

P1 $V(e_i) = \sigma^2$ pour tout i : Le postulat d'homoscédasticité)

ceci nous permet d'estimer b_0, b_1 sans connaître σ^2

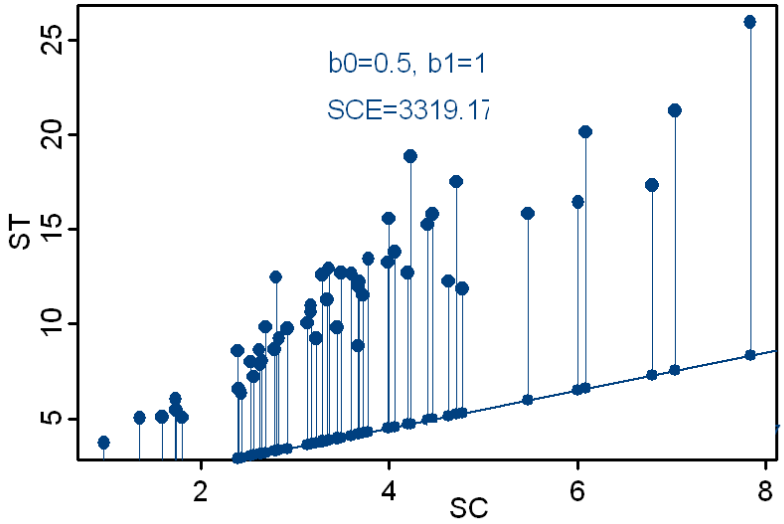
P2 e_i suit une loi Gaussienne $G(0, \sigma^2)$ pour tout i : Le postulat de normalité

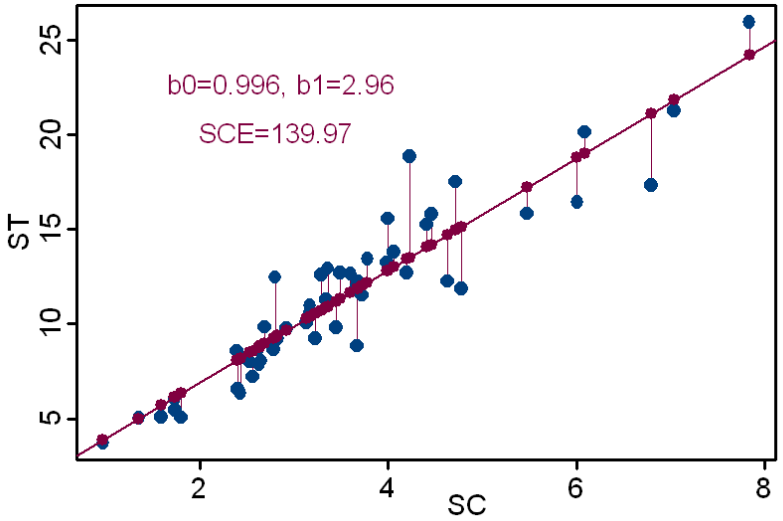
Pour tout x fixé la valeur y_x associée à x suit une loi Gaussienne $G(b_0 + b_1 x, \sigma^2)$; on peut fournir un intervalle de confiance pour y_x (dépendant de σ^2)

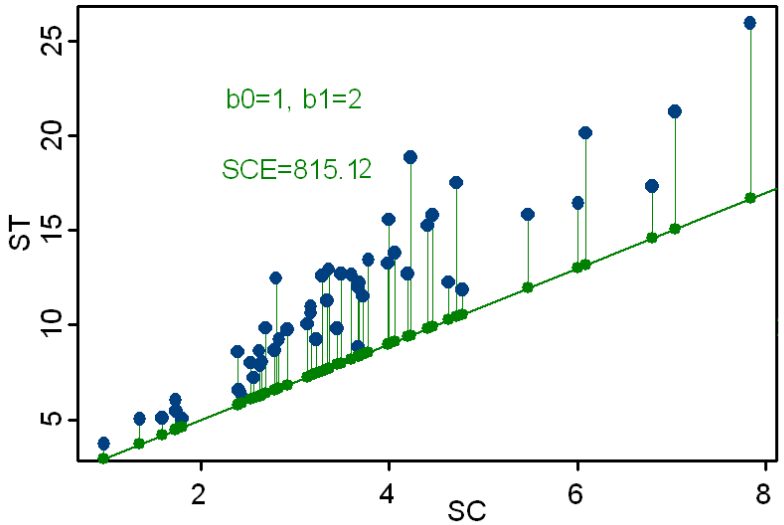
3. Estimation des paramètres par les moindres carrés

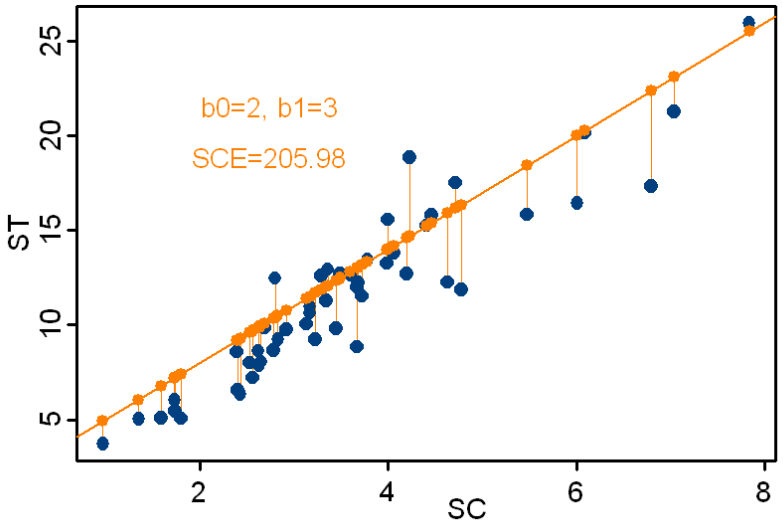
1. Approche graphique (basée sur l'exemple Télédétection)
   Le principe de la méthode des moindres carrés
2. Approche formelle
   Les estimateurs des moindres carrés

Approche graphique



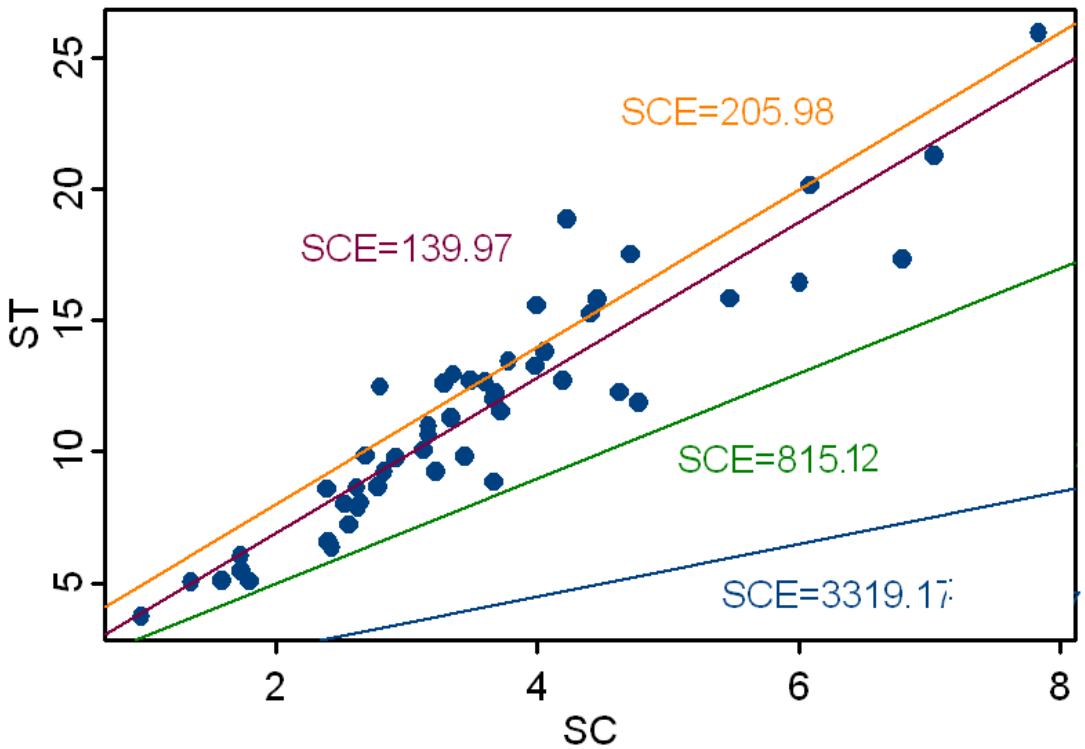






Comparaison des 4 choix

sur le critère $SCE(b_0, b_1) = \sum_{i=1,50} \{ST_i - (b_0 + b_1 SC_i)\}^2$



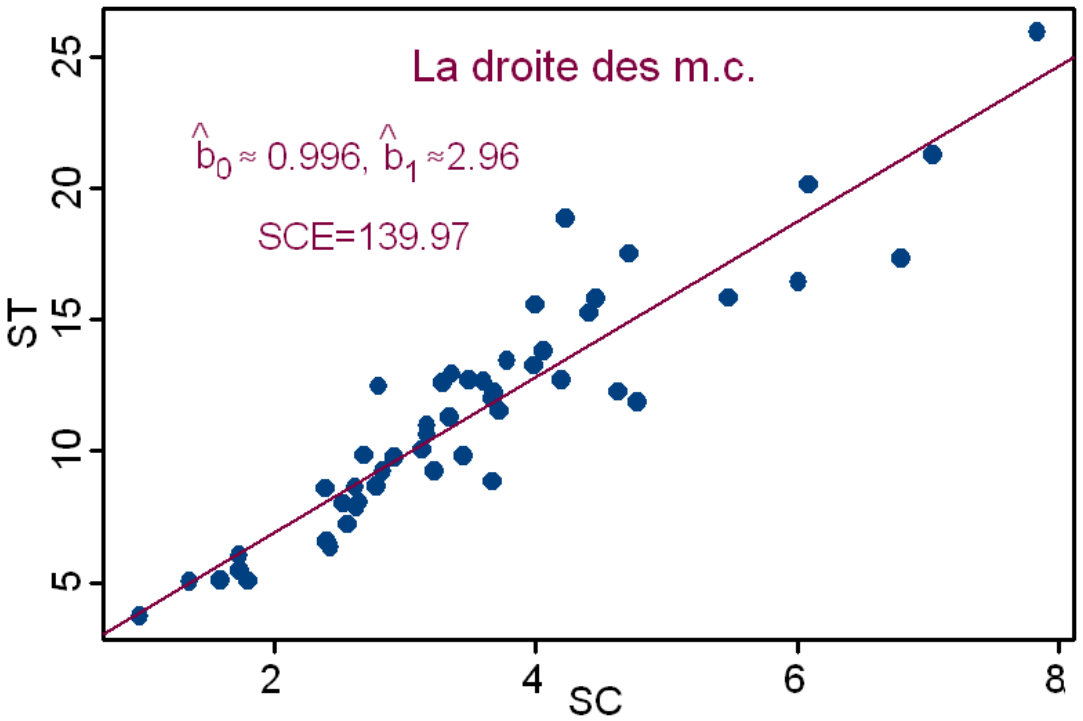
Les estimations des moindres carrés

SCE minimum pour

$\hat{b}_0 = 0.996, \hat{b}_1 = 2.9596$

La droite des moindres carrés :

$$ST = 0.996 + 2.9596 \cdot SC$$



Approche formelle

- On cherche $\hat{b}_0, \hat{b}_1$ qui minimisent

$$SCE(b_0, b_1) = \sum_{i=1,n} \{y_i - (b_0 + b_1 x_i)\}^2$$

- Un problème d'optimisation de $SCE(b_0, b_1)$ - une fonction à deux variables b_0, b_1
- $\hat{b}_0, \hat{b}_1$ sont la solution des équations (*)

$$\left\{ \begin{array}{rcl} \frac{\partial S(b_0,b_1)}{\partial b_0} & = & 0 \\ \frac{\partial S(b_0,b_1)}{\partial b_1} & = & 0 \end{array} \right. \qquad \text{équations (*)}$$

car $SCE(b_0, b_1) \to +\infty$ quand $b_0 \to +\infty$ (idem pour b_1)

Les estimateurs des moindres carrés

- $a) \sum_i x_i = n\bar{x}, \sum_i y_i = n\bar{y}$ $\quad b) \sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n[\bar{x}]^2$
  $c) \sum_i (x_i - \bar{x})\bar{y} = 0$ $\quad d) \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - n\bar{x}\bar{y}$

- Les équations (*) donnent :

$$\begin{cases} \sum_i 2(y_i - \hat{b}_0 - \hat{b}_1 x_i) \cdot -1 &= 0 \\ \sum_i 2(y_i - \hat{b}_0 - \hat{b}_1 x_i) \cdot (-x_i) &= 0 \end{cases}$$

- En utilisant a. On en obtient

$$\begin{cases} \sum_i y_i - \sum_i \hat{b}_0 - \hat{b}_1 \sum_i x_i &= 0 \\ \sum_i y_i x_i - \hat{b}_0 \sum_i x_i - \hat{b}_1 \sum x_i^2 &= 0 \end{cases} \Rightarrow \begin{cases} n\bar{y} - n\hat{b}_0 - n\hat{b}_1\bar{x} &= 0 \\ \sum_i y_i x_i - n\hat{b}_0\bar{x} - \hat{b}_1 \sum x_i^2 &= 0 \end{cases}$$

- Finalement en utilisant d. et en remplaçant $\hat{b}_0$ dans la 2 équation :

$$\begin{cases} \hat{b}_0 &= \bar{y} - \hat{b}_1\bar{x} \\ \hat{b}_1 &= \frac{\sum_i y_i x_i - n\bar{x}\bar{y}}{\sum x_i^2 - n[\bar{x}]^2} \end{cases} \Rightarrow \begin{cases} \hat{b}_0 &= \bar{y} - \hat{b}_1\bar{x} \\ \hat{b}_1 &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \end{cases}$$

Les estimateurs

- Estimateur de b_0 (resp. b_1)

$$\hat{b}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

- Estimateur de σ^2

$$s^2 = \frac{1}{n-2} \sum_i (y_i - \hat{y}_i)^2$$

Nota. $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$

(*) Ecriture matricielle du modèle

$$\boxed{\mathbf{Y} = \mathbf{X}\beta + \mathbf{E}}$$

où

- β est le vecteur colonne $(b_0, b_1)^t$
- X est une matrice $nx2$ et Y et E sont des vecteurs de taille n

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \; Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \; E = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

- On montre que $SCE(b_0, b_1) = (Y - X\beta)^t (Y - X\beta)$ et

$$\boxed{(Y - X\beta)^t (Y - X\beta) \geq (Y - X\hat{\beta})^t (Y - X\hat{\beta}) := SCE(\hat{b}_0, \hat{b}_1)}$$

où :

$$X^t X = \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}, \; X^t Y = \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix}, \; (X^t X)^{-1} X^t Y = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \end{pmatrix}$$

- (*) Justification. Soit $Y - X\beta = Y - X\hat{\beta} + X\hat{\beta} - X\beta$. Puisque $(\beta - \hat{\beta})^t X^t (Y - X\hat{\beta}) = 0$ alors $(Y - X\beta)^t (Y - X\beta) = (Y - X\hat{\beta})^t (Y - X\hat{\beta}) + (X\beta - X\hat{\beta})^t (X\beta - X\hat{\beta}) \geq (Y - X\hat{\beta})^t (Y - X\hat{\beta})$.

4. Les problèmes statistiques

a) Estimation des paramètres : Propriétés et tests

b) Vérification des postulats du modèle et diagnostique sur les résidus

c) Qualité du modèle

d) Les intervalles de confiance pour $E(y_x)$ et de prédiction pour y_x

a-1) Estimation des paramètres : propriétés

- Les estimateurs $\hat{b}_0$) et $\hat{b}_1$ sont sans biais pour les paramètres respectif et de variance minimale (TH. de Gauss-Markov)

  $E(\hat{b}_0) = b_0)$, $E(\hat{b}_1) = b_1$; Ils sont de variance minimale dans la classe des estimateurs sans biais.

  Les variances de ces deux estimateurs sont $Var(\hat{b}_0)$, $Var(\hat{b}_1)$; on note respectivement $\sigma_{\hat{b}_0}$ et $\sigma_{\hat{b}_1}$ leurs écart-types :

$$\sigma_{\hat{b}_0} = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}},\ \sigma_{\hat{b}_1} = \sigma \sqrt{\frac{1}{\sum_i (x_i - \bar{x})^2}}.$$

- $E(s^2) = \sigma^2$; on estime alors $\sigma_{\hat{b}_0}$ et $\sigma_{\hat{b}_1}$ par :

$$s_{\hat{b}_0} = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}},\ s_{\hat{b}_1} = s \sqrt{\frac{1}{\sum_i (x_i - \bar{x})^2}}.$$

a-2) Estimation des paramètres : tests

- Les tests suivants sont valables si P2 est vérifié

- Est-ce que la valeur de y dépend-t-elle de x ?

  S'écrit $H0_{b_1} : b_1 = 0$. A un risque fixé α l'hypothèse $H0_{b_1}$ est rejetée si $t_{b_1} > qt_{n-2}(1 - \frac{\alpha}{2})$.

  Définition. $t_{b_1} = \frac{\hat{b}_1}{s_{\hat{b}_1}}$ où $s_{\hat{b}_1}$ est fourni à la page 17

  $qt_{n-2}(1 - \frac{\alpha}{2})$ est le quantile $1 - \frac{\alpha}{2}$ d'une v.a. de student à $n - 2$ ddl

- Est-ce que $y = 0$ quand $x = 0$?

  Reformulée comme hypothèse nulle $H0_{b_0} : b_0 = 0$. $H0_{b_0}$ est rejetée au seuil α si $t_{b_0} > qt_{n-2}(1 - \frac{\alpha}{2})$.

- Test de $H0 : b_1 = a$ où a est une valeur quelconque - On rejette $H0$ au seuil α si $t_{b_1;a} > qt_{n-2}(1 - \alpha)$ où $t_{b_1;a} = \frac{\hat{b}_1 - a}{s_{\hat{b}_1}}$.

b) Vérification des postulats du modèle et diagnostique sur les résidus

- On peut vérifier le postulat P2 par un test d'ajustement
  D'autres procédures permettent de vérifier P0, P1 et l'indépendance des résidus dans des cas particuliers

- Deux diagnostiques importantes sont proposées par les logiciels statistiques :
  - Individu suspect : Si son résidus "studentisés" r_i est supérieur à 2
    $$r_i = \frac{\hat{e}_i}{s_{-i}\sqrt{1-h_i}}$$
    où h_i est le terme i de la diagonale de $X(X^tX)^{-1}X^t$.
  - Individu influent : Si la distance de Cook est supérieur à 1.
    $$dCook_i = \frac{r_i^2}{1}\ \frac{h_i}{1-h_i}$$

    voir Christensen, pg. 349 : Pour p variables on a $dCook_i = \frac{r_i^2}{p}\ \frac{h_i}{1-h_i}$

c) Qualité du modèle

- La qualité du modèle est mesurée à l'aide de R^2 : $\boxed{R^2 = \frac{\sum_i (\hat{y}_i - \bar{\hat{y}})^2}{\sum_i (y_i - \bar{y})^2}}$

  Rappels. $\bar{\hat{y}} = \bar{y}, \hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$

  R^2 est la rapport entre la SCE expliquée (c.a.d. $\sum_i \left(\hat{y}_i - \bar{\hat{y}}\right)^2$) et la SCE totale (c.a.d. $SCE_T = \sum_i (y_i - \bar{y})^2$)

- R^2 est égale au carré du coefficient de corrélation linéaire r de x et y ; donc $0 \leq R^2 \leq 1$

$$\hat{y}_i - \bar{y} = \hat{b}_1 (x_i - \bar{x}) \Rightarrow R^2 = \frac{[\hat{b_1}]^2 \sum_i (x_i - \bar{x})^2}{\sum_i (y_i - \bar{y})^2} \text{ et } r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} := \frac{\hat{b}_1 \sum_i (x_i - \bar{x})^2)}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

- Interprétation. Le modèle est de bonne qualité si R^2 est proche de 1.

d) Prédiction et intervalles de confiances

- Pour tout x_o fixé on estime y_{x_o} par $\boxed{\hat{y}_{x_o} = \hat{b}_0 + \hat{b}_1 x_o}$

  $E(\hat{y}_x) = E(y_x)$; c.a.d. $\hat{y}_x$ est un estimateur sans biais pour $E(y_x)$ (ceci découle de $E(\hat{b}_0) = b_0$ et $E(\hat{b}_1) = b_1$ - voir page 17.

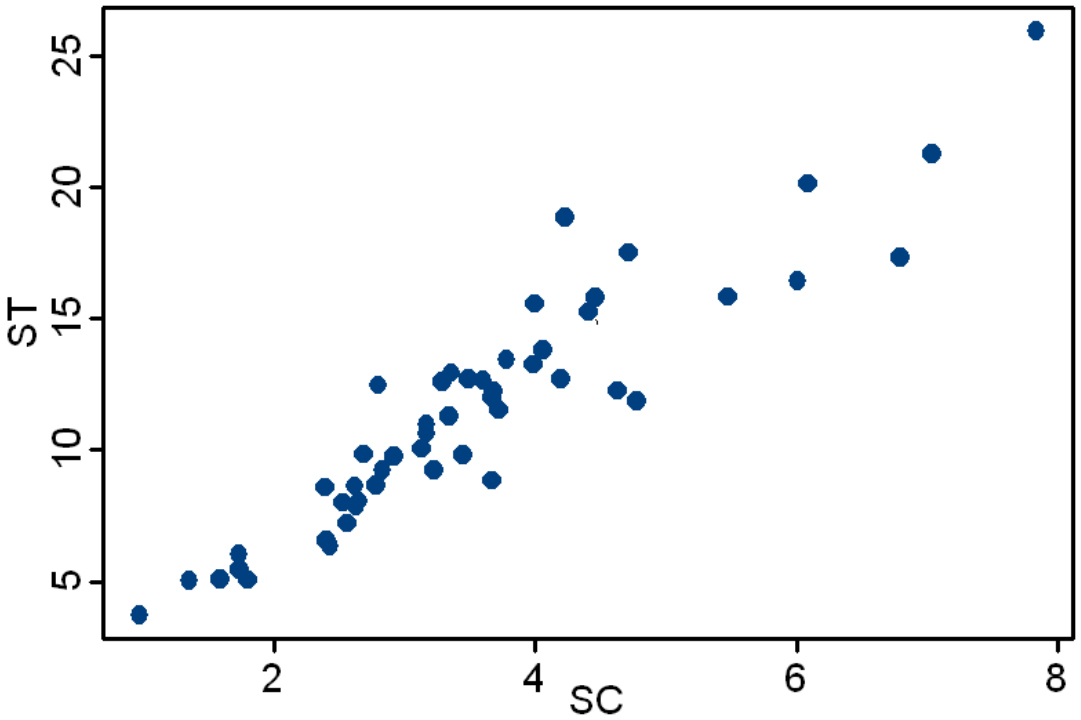
- Si P2 est vérifié on a :
  - L'intervalle de Confiance de seuil α pour $E(y_x)$:
    $$\boxed{IC_\alpha(E(y_x))\hat{y}_{x_o} \pm qt_{n-2}(1-\tfrac{\alpha}{2})\, s_{y_{x_o}}}$$
  - L'intervalle de Prédiction de seuil α pour y_x :
    $$\boxed{IP_\alpha(y_x) = \hat{y}_{x_o} \pm qt_{n-2}(1-\tfrac{\alpha}{2}) \sqrt{s^2_{y_{x_o}} + s^2}}$$

$s_{y_{x_o}} = \sigma \sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$, s^2 est l'estimation de σ^2

5. Application à l'estimation de la biomasse ligneuse

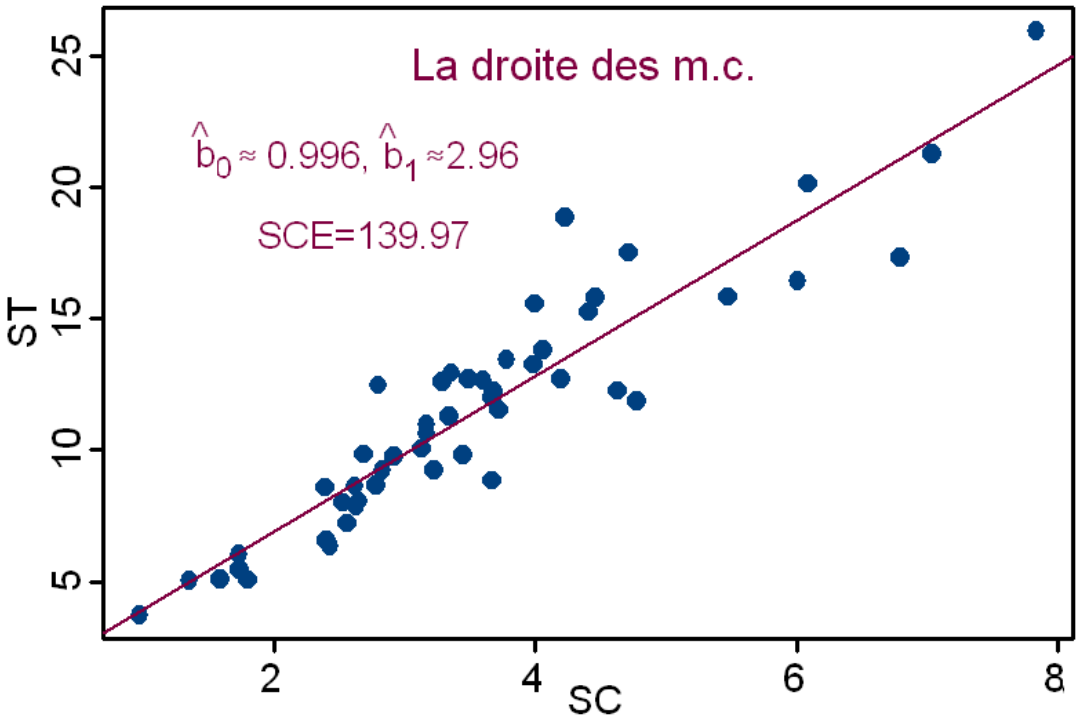
- Présentation graphique des données - a-t'on une liaison linéaire entre ST et SC ?
- La droite de régression, estimation des paramètres et qualité du modèle
- Analyse des résidus
- L'IP de ST et IC de $E(ST)$

Présentation et étude graphique des données



- Le graphique indique une dépendance linéaire entre ST et SC.
- La variabilité autour d'une droite passant "près" du nuage est relativement stable (voir le postulat P1)
- On ne remarque pas de données suspectes

La droite de régression

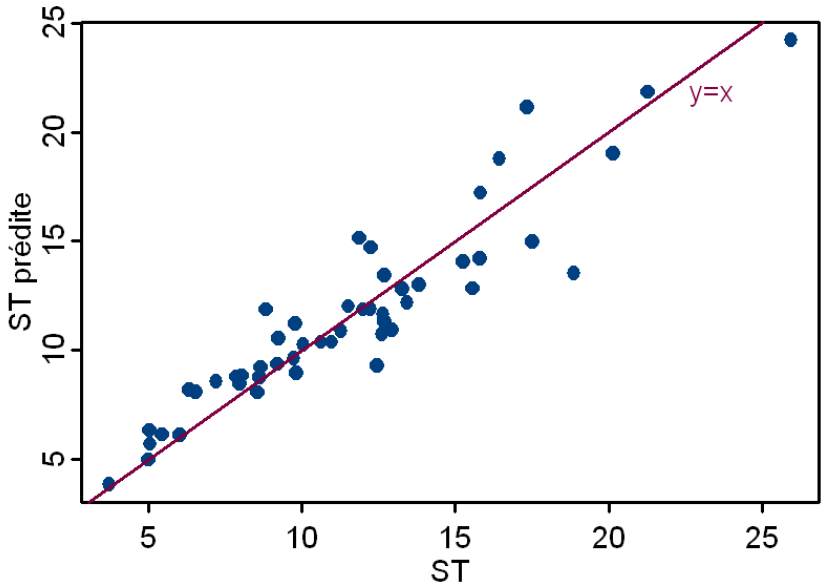


- Le modèle ajusté : $ST = 0.996 + 2.9596\, SC$
- $s^2 = 2.9159$
- $R^2 = 0.8642$

Analyse des résidus

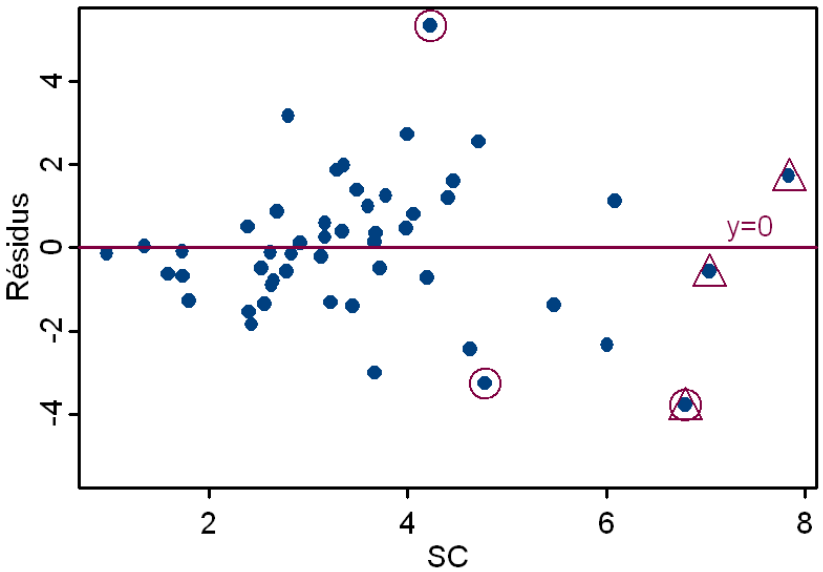
Le résidu est la différence entre ST observé et ST prédite par le modèle

$$\hat{e}_i = ST_i - \widehat{ST}_i$$



Graph $(ST, ST\,\text{prédite})$

Un nuage resserré autour de la diagonale indique un bon modèle de prédiction



Graph $(SC, \text{résidus})$

Identifier des individus suspects (◯) ou influents (△) et détecter les problèmes de hétéroscédasticité (postulat P1)

Unusual Residuals

Row	X	Y	Predicted Y	Residual	Studentized Residual
4	6.8	17.34	21.1215	-3.78149	-2.49
26	4.23	18.86	13.5153	5.34474	3.53
30	4.78	11.88	15.1431	-3.26306	-2.01

Influential Points

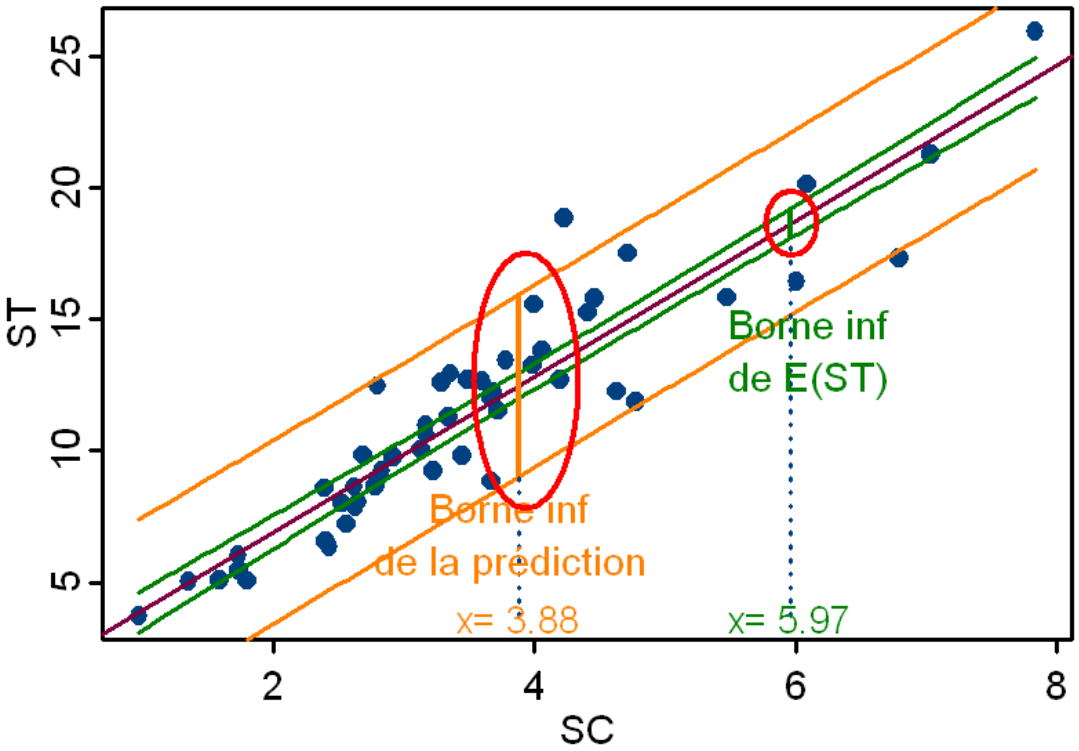
Row	X	Y	Predicted Y	Studentized Residual	Leverage
4	6.8	17.34	21.1215	-2.49	0.1234
33	7.84	25.94	24.1995	1.14	0.2004
35	7.04	21.27	21.8318	-0.35	0.1393

Average leverage of single data point = 0.04

Prédiction et IC de la prédiction

SC	$\widehat{ST}$	$IC_{0.95}(E(ST))$		$IP_{0.95}(ST)$	
3.88	12.493	11.999	12.986	9.024	15.962
5.97	18.654	18.100	19.209	15.176	22.132

Exemple : $SC = 3.88 \Rightarrow ST = 0.996 + 2.9596 \cdot 3.88 := 12.493$



Tests sur les paramètres

```
Linear Model: Y = a + b*X
-------------------------------------------------
Dependent variable: ST
Independent variable: SC
-------------------------------------------------
                     Standard       T
Parameter Estimate     Error      Statistic   P-Value
-------------------------------------------------
Intercept   0.9961     0.6492     1.53429     0.1315

Slope       2.9596     0.1694     17.4742     0.0000
-------------------------------------------------
```

Soit $\alpha = 0.05$: $H0 : b_1 = 0$ est rejetée car le Proba=0.0000 (2e ligne) du tableau. On ne peut peut pas rejeter $H0 : b_0 = 0$ au risque 0.05 car Proba=0.1315 (1e ligne du tableau).

Lecture équivalente. On rejette $H0 : b_1 = 0$ car $t_{b_1} = 17.4742 > 2.0106$, le quantile $1 - \frac{0.05}{2}$ de t_{48}.

6. La régression non-linéaire

Comparison of Alternative Models

Model	Correlation	R-Squared
Double reciprocal	0.9537	90.96%
Multiplicative	0.9455	89.40%
Linear	0.9296	86.42%
Square root-X	0.9288	86.27%
Square root-Y	0.9211	84.84%
Logarithmic-X	0.9054	81.98%
Exponential	0.8975	80.54%
S-curve	-0.8934	79.81%
Reciprocal-X	-0.7872	61.98%

R^2 calculé à partir du modèle "linéarisé"

Multiplicative model: Y = a*X^b

Dependent variable: ST
Independent variable: SC

Parameter	Estimate	Standard Error	T Stat.	P-Value
Intercept	1.24655	0.0588	21.197	0.0000
Slope	0.94079	0.0467	20.124	0.0000

NOTE: intercept = ln(a) => 3.41831

A partir des résultats de gauche on peut dire que le modèle multiplicatif est meilleur que RLS sur le critère R^2 (R^2 calculé à partir du modèle $lnY = lna + b * lnX$).

Cet exemple illustre le cas d'un modèle $f_{a,b}(x) = a \cdot x^b$ non-linéaire mais "linéarisable" par la transformation logarithmique. En général si $f_{a,b}(x)$ est une fonction quelconque dépendant de a et b on estime ces deux paramètres en optimisant $\sum_i (y_i - f_{a,b}(x))^2$.

La régression linéaire à deux variables explicatives

1. Estimation du taux de fécondité de pucerons

Objectif.[2] Estimer le taux de fécondité de pucerons à partir de variables moins coûteuses à mesurer

Parasite "Leptomastix dactylopi" un parasite des cochenilles du manioc au Congo

Variables La Taille (mm) et la Longévité :=Long (jours) sont plus simple à mesurer que la Fécondité :=Fec

Question Peut-on estimer Fécondité connaissant les deux autres variables ?

Le modèle La régression linéaire à deux variables explicatives : $Fec \approx b_0 + b_1 \cdot Taille + b_2 \cdot Long$

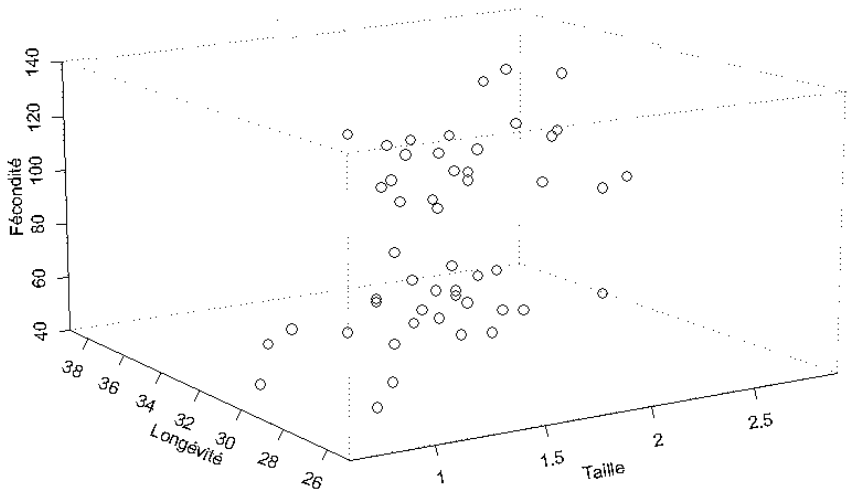
Mesurer la Taille ou la Longévité est plus simple que évaluer la Fécondité :

Mesurer la Taille prend 15 secondes ; Mesurer sa Longévité suppose suivre tous les jours celles qui survivent ; Mesurer la Fécondité suppose élever et isoler des cochenilles saines, les mettre en contact avec chaque animal, mettre ensuite chaque cochenille en élevage, puis les disséquer au bout de quelques jours

2. Source : Jean-Sébastien Pierre (1997) - essai réalisé au Congo sur le manioc

Les données

Nr	Taille	Longevite	Fecondite	Nr	Taille	Longevite	Fecondite
1	0.63	29	52	26	1.76	25	82
2	0.75	25	56	27	1.87	35	86
3	0.85	31	57	28	1.93	34	89
4	0.92	26	59	29	1.85	33	89
5	0.96	31	61	30	1.88	36	89
6	1.12	30	61	31	1.93	36	91
7	1.45	32	62	32	2.14	35	95
8	1.13	28	63	33	2	33	98
9	1.32	29	65	34	2.11	38	98
10	1.44	29	65	35	2.23	38	99
11	1.47	27	66	36	2.17	36	99
12	1.35	31	67	37	2	36	100
13	1.33	27	67	38	2.15	31	101
14	1.62	30	68	39	2	33	101
15	1.57	29	69	40	2.33	37	102
16	1.63	28	69	41	2.26	35	102
17	1.36	29	69	42	2.22	29	103
18	1.62	30	70	43	1.91	38	104
19	1.53	30	71	44	2.34	29	106
20	1.52	31	72	45	2.53	34	107
21	1.62	27	72	46	2.56	34	109
22	1.73	30	74	47	2.46	35	110
23	1.71	31	75	48	2.72	39	115
24	1.64	33	75	49	2.84	39	119
25	1.71	29	79	50	2.92	37	121



Présentation 3D des données - Difficile à affirmer si les points sont proches d'un plan !

Notations

n - nombre d'individus (pucerons) mesurés ($n = 50$ dans notre application)

x_1, x_2 - les deux variables explicatives - ce sont des variables quantitatives - notée *Taille* et *Longevite* dans cette application - On notera $x_{1,1}, \ldots, x_{n,1}$ et $x_{1,2}, \ldots, x_{n,2}$ les n mesures disponibles pour chacune des variables x_1 et x_2

y - la variable à expliquer ou dépendante - c'est une variable quantitative - notée *Fecondite* dans notre application - On notera $y_1, \ldots, y_n$ les n mesures disponibles de la variable y

2. Le modèle et les postulats

Le modèle de la régression linéaire multiple (RLM) à deux variables explicatives s'écrit en général :

$$y = b_0 + b_1 x_1 + b_2 x_2 + \mathrm{e}$$

- e désigne l'erreur ou l'écart entre la partie expliquée par le modèle et l'observation
- b_0, b_1, b_2 sont les paramètres inconnus du modèle
- le modèle est linéaire par rapport aux paramètres b_0, b_1, b_2 et à deux variables explicatives

L'estimation des paramètres b_0, b_1, b_2 du modèle RLM est basée sur un essai donnant n mesures indépendantes $x_{i,1}, x_{i,2}, y_i$ et en écrivant :

$$y_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + e_i, \ \ i = 1, \ldots, n$$

Les postulats

Le postulat d'indépendance : On suppose que les y_i sont indépendantes

P0 $E(e_i) = 0$ pour tout i

l'erreur moyenne est nulle ; $E(y_i) = b_0 + b_1 x_{i,1} + b_2 x_{i,2}$ et $e_i = y_i - E(y_i)$

P1 $V(e_i) = \sigma^2$ pour tout i (Homoscédasticité)

ceci nous permet d'estimer b_0, b_1, b_2 sans connaître σ^2

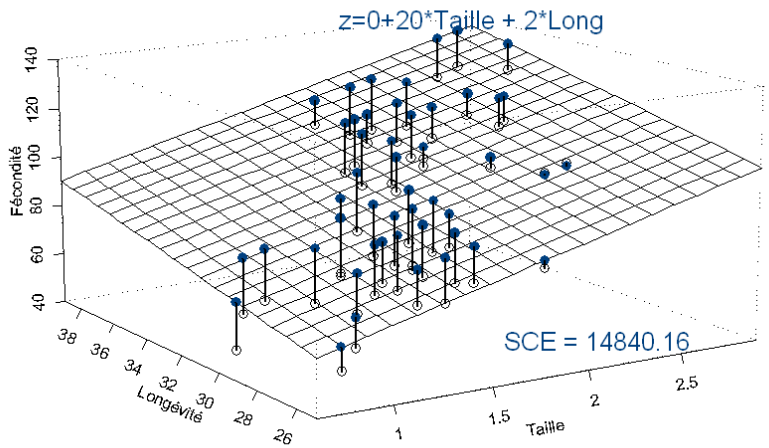
P2 e_i suit une loi Gaussienne $G(0, \sigma^2)$ pour tout i

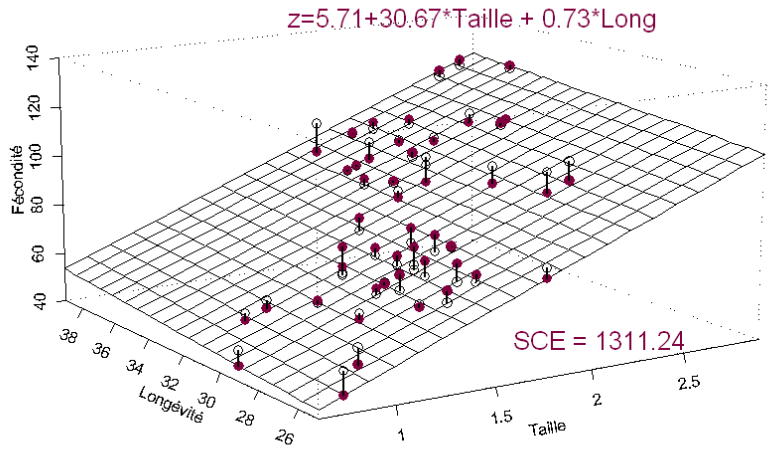
Pour tout $x^o = (x_1^o, x_2^o)'$ fixé la valeur y_{x^o} associée à x^o suit une loi Gaussienne $G(b_0 + b_1 x_1^o + b_2 x_2^o, \sigma^2)$; on peut fournir un intervalle de confiance pour y_{x^o} (dépendant de σ^2)

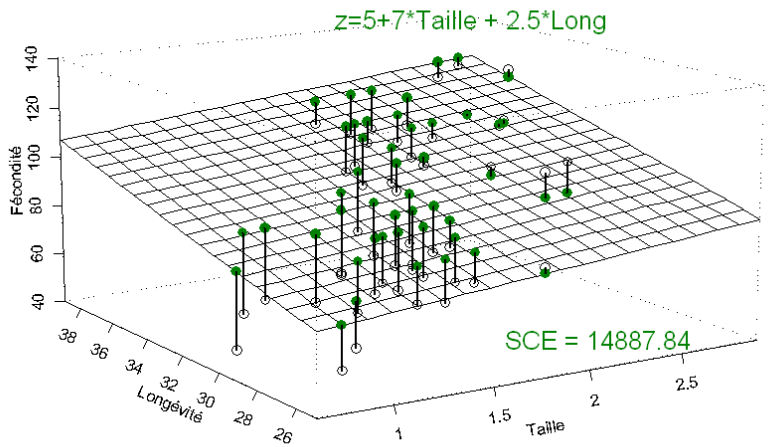
3. L'estimation des paramètres par les moindres carrés

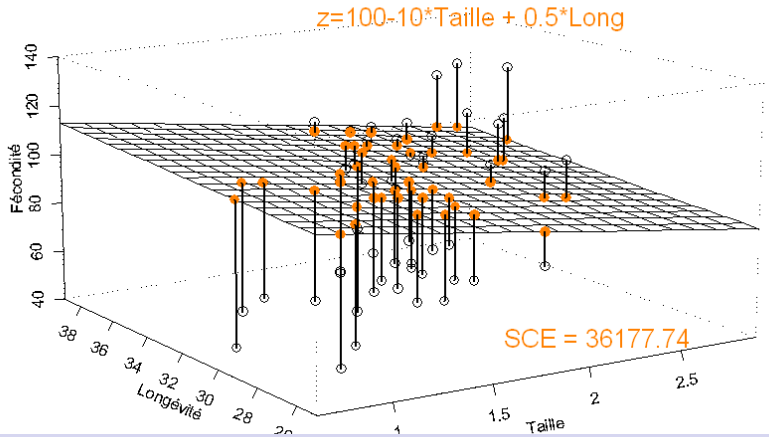
1. Approche graphique - exemple "Fécondité de pucerons" :
   Le principe de la méthode des moindres carrés
2. Approche formelle :
   Les estimateurs des moindres carrés

Approche graphique









Approche graphique : Comparaison des 4 choix

Critère de choix : $SCE(b_0, b_1, b_2) = \sum_{i=1,50} \{Fec_i - (b_0 + b_1 Taille_i + b_2 Long_i)\}^2$

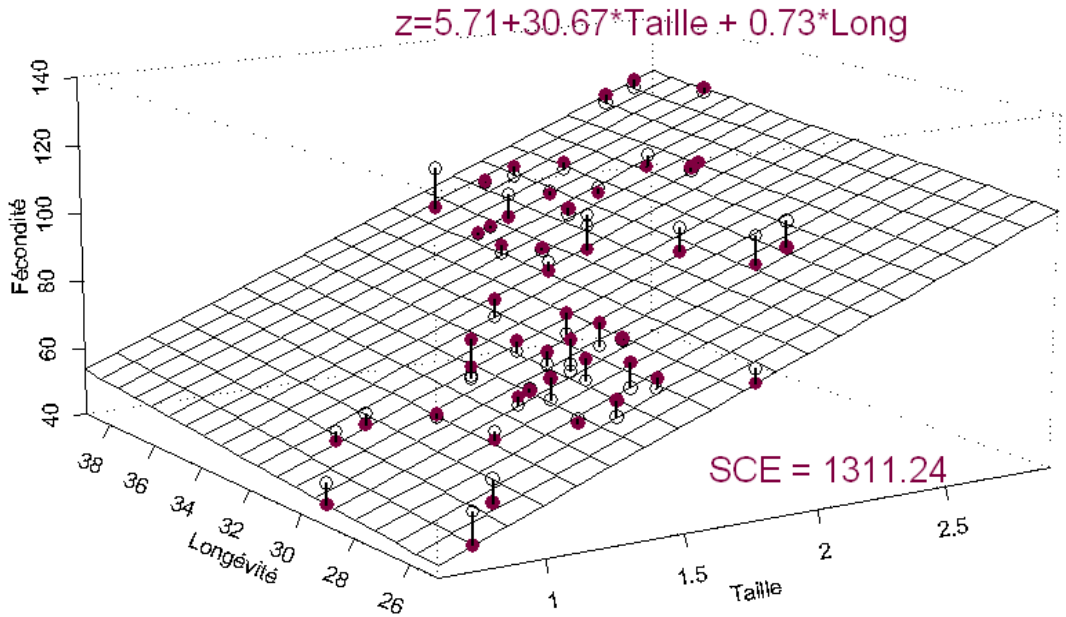
Choix	b_0	b_1	b_2	$SCE(b_0, b_1, b_2)$
Choix 1	0.0	20.0	2.0	14840.160
Choix 2	5.711	30.666	0.727	1311.237
Choix 3	5.0	7.0	2.5	14887.844
Choix 4	100.0	-10.0	0.5	36177.740

Approche graphique : Le plan de régression

SCE minimum pour : $\hat{b}_0 = 5.711, \hat{b}_1 = 30.666, \hat{b}_2 = 0.727$

L'équation des moindres carrés :

$$Fec = 5.711 + 30.666 \cdot Taille + 0.727 \cdot Long$$



Calcule des estimateurs des moindres carrés

- On cherche $\hat{b}_0, \hat{b}_1, \hat{b}_2$ qui minimisent

$$SCE(b_0, b_1, b_2) = \sum_{i=1,n} \{y_i - (b_0 + b_1 x_{i,1} + b_2 x_{i,2})\}^2$$

- Un problème d'optimisation de $SCE(b_0, b_1, b_2)$ - une fonction à trois variables b_0, b_1, b_2
- $\hat{b}_0, \hat{b}_1, \hat{b}_2$ sont la solution des **équations normales**

$\left(\text{car } SCE(b_0, b_1, b_2) \to +\infty \text{ quand } b_0, b_1 \text{ ou } b_2 \to \infty\right)$

$$\begin{cases} \frac{\partial S(b_0,b_1,b_2)}{\partial b_0} = 0 \\ \frac{\partial S(b_0,b_1,b_2)}{\partial b_1} = 0 \\ \frac{\partial S(b_0,b_1,b_2)}{\partial b_2} = 0 \end{cases} \Leftrightarrow \begin{cases} \sum_i 2(y_i - \hat{b}_0 - \hat{b}_1 x_{i,1} - \hat{b}_2 x_{i,2}) \cdot (-1) = 0 \\ \sum_i 2(y_i - \hat{b}_0 - \hat{b}_1 x_{i,1} - \hat{b}_2 x_{i,2}) \cdot -(x_{i,1}) = 0 \\ \sum_i 2(y_i - \hat{b}_0 - \hat{b}_1 x_{i,1} - \hat{b}_2 x_{i,2}) \cdot -(x_{i,2}) = 0 \end{cases}$$

Les estimateurs des paramètres

- Propriété : $\sum_i a_i = n\bar{a}$, $\frac{1}{n}\sum_i (a_i - \bar{a})b_i = \frac{1}{n}\sum_i a_i b_i - \bar{a}\,b_i$, $\frac{1}{n}\sum_i (a_i - \bar{a})\bar{b} = 0$

  Notations 1 : $V_{1,y} = \frac{1}{n}\sum_i (x_{i,1} - \bar{x}_1)(y_i - \bar{y}) := \frac{1}{n}\sum_i x_{i,1}y_i - \bar{x}_1\bar{y}$, $V_{2,y} = \frac{1}{n}\sum_i x_{i,2}y_i - \bar{x}_2\bar{y}$

  Notations 2 : $V_{1,1} = \frac{1}{n}\sum_i [x_{i,1}]^2 - [\bar{x}_1]^2$, $V_{1,2} = \frac{1}{n}\sum_i x_{i,1}x_{i,2} - \bar{x}_1\,\bar{x}_2$, $V_{2,2} = \frac{1}{n}\sum_i [x_{i,2}]^2 - [\bar{x}_2]^2$

- Les estimateurs m.c. de b_0, b_1, b_2 : à partir des équations normales (en divisant par n chaque équation) : On retranche la 1e équation des deux autres pour obtenir les équations de droites

$$\left\{\begin{array}{rcl} \bar{y} - \hat{b}_0 - \hat{b}_1\,\bar{x}_1 - \hat{b}_2\,\bar{x}_2 & = & 0 \\ \hat{b}_0\,\bar{x}_1 - \hat{b}_1\,\frac{1}{n}[x_{i,1}]^2 - \hat{b}_2\,\frac{1}{n}x_{i,2}x_{i,1} & = & 0 \\ \hat{b}_0\,\bar{x}_2 - \hat{b}_1\,\frac{1}{n}x_{i,1}x_{i,2} - \hat{b}_2\,\frac{1}{n}[x_{i,2}]^2 & = & 0 \end{array}\right. \Leftrightarrow \left\{\begin{array}{l} \hat{b}_0 = \bar{y} - \hat{b}_1\,\bar{x}_1 + \hat{b}_2\,\bar{x}_2 \\ \hat{b}_1\,V_{1,1} + \hat{b}_2\,V_{1,2} = V_{1,y} \\ \hat{b}_1\,V_{1,2} + \hat{b}_2\,V_{2,2} = V_{2,y} \end{array}\right.$$

- L'estimateur de σ^2 : $\boxed{s^2 = \frac{1}{n-3}\sum_i (y_i - \hat{y}_i)^2}$ où

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{i,1} + \hat{b}_2 x_{i,2}$$

(*) Ecriture matricielle du modèle

$$\boxed{\mathbf{Y = X\beta + E},\ \mathbf{Var(E) = \sigma^2 I}}$$

- où I est la matrice idéntité $n \times n$ et :

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},\ X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} \\ 1 & x_{2,1} & x_{2,2} \\ \vdots & \vdots & \\ 1 & x_{n,1} & x_{n,2} \end{pmatrix},\ \beta = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix},\ E = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

- On a $SCE(b_0, b_1, b_2) = (Y - X\beta)^t (Y - X\beta)$ et

$$\boxed{(Y - X\beta)^t (Y - X\beta) \geq (Y - X\hat{\beta})^t (Y - X\hat{\beta}) := SCE(\hat{b}_0, \hat{b}_1, \hat{b}_2)}$$

où :

$$X^t X = \begin{pmatrix} n & \sum_i x_{i,1} & \sum_i x_{i,2} \\ \sum_i x_{i,1} & \sum_i x_{i,1}^2 & \sum_i x_{i,1}x_{i,2} \\ \sum_i x_{i,2} & \sum_i x_{i,1}x_{i,2} & \sum_i x_{i,2}^2 \end{pmatrix},\ X^t Y = \begin{pmatrix} \sum_i y_i \\ \sum_i x_{i,1}y_i \\ \sum_i x_{i,2}y_i \end{pmatrix},\ \boxed{\hat{\boldsymbol{\beta}} = \mathbf{(X^t X)^{-1} X^t Y}}$$

- (*) Justification. Soit $Y - X\beta = Y - X\hat{\beta} + X\hat{\beta} - X\beta$. Puisque $(\beta - \hat{\beta})^t X^t (Y - X\hat{\beta}) = 0$ alors $(Y - X\beta)^t (Y - X\beta) = (Y - X\hat{\beta})^t (Y - X\hat{\beta}) + (X\beta - X\hat{\beta})^t (X\beta - X\hat{\beta}) \geq (Y - X\hat{\beta})^t (Y - X\hat{\beta})$.

4. Les problèmes statistiques

a) Estimation des paramètres : propriétés et tests (tableau ANOVA)

b) Qualité du modèle

c) Prédiction et intervalle de prédiction

a-1) Estimation des paramètres : propriétés

- $\hat{b}_0$ (resp. $\hat{b}_1$, $\hat{b}_2$) est BLUE (Best Linear Unbiased Estimation)

  C'est le Théorème de Gauss-Markov : $E(\hat{b}_0 = b_0$, $E(\hat{b}_1) = b_1$, $E(\hat{b}_2) = b_2$ et si $l'Y$ est tel que $E(l'Y) = b_0$ alors $Var(l'Y) \geq Var(\hat{b}_0)$.

- Les variances de ces estimateurs sont $\sigma^2_{\hat{b}_0}$, $\sigma^2_{\hat{b}_1}$, $\sigma^2_{\hat{b}_2}$.

  Ils sont les éléments de la diagonale de la matrice $(X^tX)^{-1}\sigma^2$ car $Var\{(X^tX)^{-1}(X^tY)\} = (X^tX)^{-1}X^t\, Var\{Y\}X\,(X^tX)^{-1} = (X^tX)^{-1}(X^tX)\,(X^tX)^{-1}\sigma^2 = (X^tX)^{-1}\sigma^2$.

- $\boxed{E(s^2) = \sigma^2}$.

  En remplaçant σ^2 par s^2 on obtient les estimateurs $\boxed{s_{\hat{b}_0}}$, $\boxed{s_{\hat{b}_1}}$, $\boxed{s_{\hat{b}_2}}$ de $\sigma_{\hat{b}_0}$, $\sigma_{\hat{b}_1}$ et $\sigma_{\hat{b}_2}$.

a-1) Estimation des paramètres : Test global (tableau ANOVA)

L'équation ANOVA (ANalysis Of VAriance) permet d'évaluer la qualité du modèle et de réaliser le test de Fisher F sur les paramètres.

- Rappels et notations : $V_{1,y} = \frac{1}{n}\sum_i x_{i,1}y_i - \bar{x}_1\bar{y}$, $V_{2,y} = \frac{1}{n}\sum_i x_{i,2}y_i - \bar{x}_2\bar{y}$, $V_{1,1} = \frac{1}{n}\sum_i [x_{i,1}]^2 - [\bar{x}_1]^2$, $V_{1,2} = \frac{1}{n}\sum_i x_{i,1}x_{i,2} - \bar{x}_1\,\bar{x}_2$, $V_{2,2} = \frac{1}{n}\sum_i [x_{i,2}]^2 - [\bar{x}_2]^2$

  Un rsultats immédiat : $\hat{y}_i - \bar{\hat{y}} = \hat{b}_1(x_{i,1} - \bar{x}_1) + \hat{b}_2(x_{i,2} - \bar{x}_2)$

  Les estimateurs des vérifient les équations $\begin{cases} \hat{b}_0 = \bar{y} - \hat{b}_1\,\bar{x}_1 + \hat{b}_2\,\bar{x}_2 \\ \hat{b}_1\,V_{1,1} + \hat{b}_2 V_{1,2} = V_{1,y} \\ \hat{b}_1\,V_{1,2} + \hat{b}_2 V_{2,2} = V_{2,y} \end{cases}$

- Equation ANOVA : $\boxed{SCE_T = SCE_M + SCE_{Res}}$: $\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$

  a. Il suffit de montrer $\sum_i (\hat{y}_i - \bar{\hat{y}})(y_i - \hat{y}_i) = \sum_i (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y} - \hat{y}_i - \bar{\hat{y}}) = 0$ car $\bar{y} = \bar{\hat{y}}$

  c. $\begin{aligned} \frac{1}{n}\sum_i (\hat{y}_i - \bar{y})(y_i - \bar{y}) &= \sum_i (y_i - \bar{y})(\hat{b}_1(x_{i,1} - \bar{x}_1) + \hat{b}_2(x_{i,2} - \bar{x}_2) \\ &= \hat{b}_1 V_{1,y} + \hat{b}_2 V_{2,y} \end{aligned}$

  c. $\begin{aligned} \frac{1}{n}\sum_i (\hat{y}_i - \bar{y})^2 &= \sum_i [\hat{b}_1(x_{i,1} - \bar{x}_1) + \hat{b}_2(x_{i,2} - \bar{x}_2)]^2 \\ &= \hat{b}_1^2 V_{1,1} + 2\hat{b}_1\hat{b}_2 V_{1,2} + \hat{b}_2^2 V_{2,2} \\ &= \hat{b}_1[\hat{b}_1 V_{1,1} + \hat{b}_2 V_{1,2}] + \hat{b}_2[\hat{b}_1 V_{1,2} + \hat{b}_2^2 V_{2,2}] \\ &= \hat{b}_1 V_{1,y} + \hat{b}_2 V_{2,y} \end{aligned}$.

  d. Le résultat découle de b. et c.

Tableau ANOVA

- Les règles de calculs et terminologie

a. $ddl_M = 2$: nombre de paramètres hors terme constant, $ddl_T = n - 1$, $ddl_{Res} = n - 3$

b. $SCE_M = \sum_i (\hat{y}_i - \bar{\hat{y}})^2$, $SCE_T = \sum_i (y_i - \bar{y})^2$, $SCE_R = \sum_i (y_i - \hat{y}_i)^2$

c. $\boxed{SCE_T = SCE_M + SCE_{Res}}$, $\boxed{ddl_T = ddl_M + ddl_{Res}}$, $\boxed{F = \frac{MCE_M}{MCE_{Res}}}$

- Le tableau ANOVA

Source	Sommes des Carrés des Ecarts	Degrés De Libertés	Moyennes des Carrés des Ecarts	Statistique de Fisher
Modèle	SCE_M	2	MCE_M	F
Résiduel	SCE_{Res}	n-3	MCE_{Res}	
Total	SCE_T	n-1	MCE_T	

a-3) Estimation des paramètres : le test global de Fisher

Le test est valable si P2 est validé

- $H0 : b_1 = b_2 = 0$
  ou : "est-ce que y dépend d'au moins une des variables x ?"

- Statistique de test : $\boxed{F = \frac{MCE_M}{MCE_{Res}}}$

- Règle statistique de décision de risque α fixé apriori : On rejette $H0$ si $F > qF_{2,n-3}(1-\alpha)$.

  $qF_{2,n-3}(1-\alpha)$ est le quantile $1-\alpha$ d'une v.a. de Fisher à 2 et $n-3$ ddl

On peut par analogie tester des hypothèses qui concernent des combinaisons linéaires de b_1, b_2

a-3) Estimation des paramètres : les tests pour chaque paramètre

- Le test est valable si P2 est validé
- On présentera le test sur b_1 ; Les tests concernant b_0 et b_2 sont similaires

- $H0 : b_1 = 0$
  Est-ce que y dépend de x_1 (la variable du modèle associée à b_1) ?
- Statistique de test : $\boxed{t = \frac{\hat{b}_1}{s_{\hat{b}_1}}}$
- Règle statistique de décision de risque α fixé apriori : On rejette $H0$ si $t > qt_{n-3}(1-\alpha/2)$.

  $qt_{n-3}(1-\frac{\alpha}{2})$ est le quantile $1-\frac{\alpha}{2}$ d'une v.a. de student à $n-3$ ddl

- Un cas particulier : $H0 : b_0 = 0$. Cette question peut être formulée de façon équivalente : "le terme constant est utile dans le modèle ?". La règle de décision est analogue et basée sur la statistique $\hat{b}_0/s_{\hat{b}_0}$.

b) Qualité du modèle

- La qualité du modèle est mesurée à l'aide de R^2 : $\boxed{R^2 = \frac{\sum_i (\hat{y}_i - \bar{\hat{y}})^2}{\sum_i (y_i - \bar{y})^2}}$

  Rappels. $\bar{\hat{y}} = \bar{y}, \hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{i,1} + \hat{b}_2 x_{i,2}$

  R^2 est la rapport entre SCE_M (variabilité expliquée) et SCE_T (variabilité totale)

- R^2 est égale au carré du coefficient de corrélation linéaire r de $\hat{y}$ et y ; donc $0 \leq R^2 \leq 1$

  Justification(*). Il suffit d'écrire d'une part $R^2 = \frac{\|\hat{Y} - \bar{Y}\|^2}{\|Y - \bar{Y}\|^2}$ et de l'autre $r^2 = \frac{\langle Y - \bar{Y}, \hat{Y} - \bar{Y} \rangle^2}{\|\hat{Y} - \bar{Y}\|^2 \, \|Y - \bar{Y}\|^2} = \frac{\langle \hat{Y} - \bar{Y}, \hat{Y} - \bar{Y} \rangle^2}{\|\hat{Y} - \bar{Y}\|^2 \, \|Y - \bar{Y}\|^2} = R^2$ car $\langle Y - \hat{Y}, \hat{Y} - \bar{Y} \rangle = [Y^t - Y^t X (X^t X)^{-1} X^t][X(X^t X)^{-1} X^t Y - \bar{Y}] = 0.$

- Interprétation. Le modèle est de bonne qualité si R^2 est proche de 1.

c) La prédiction et l'intervalle de la prédiction

- Pour tout $x^o = (x_1^o, x_2^o)$ fixé on estime y_{x^o} par

$$\boxed{\hat{y}_{x^o} = \hat{b}_0 + \hat{b}_1\, x_1^o + \hat{b}_2\, x_2^o}$$

$E(\hat{y}_x) = E(y_x)$: $\hat{y}_x$ est donc sans biais pour $E(y_x)$

- Si P2 est vérifié on a :
  - L'Intervalle de Confiance de seuil α pour $E(y_x)$:

$$\boxed{IC_\alpha(E(y_x)) = \hat{y}_{x^o} \pm qt_{n-3}(1-\tfrac{\alpha}{2})\, s_{y_{x^o}}},$$

Notations matricielle ! : $s^2_{y_{x^o}} = (1, x_1^o, x_2^o)(X^t\, X)^{-1}(1, x_1^o, x_2^o)^t\, s^2$

  - L'Intervalle de Prédiction de seuil α pour y_x :

$$\boxed{IP_\alpha(y_x) = \hat{y}_{x^o} \pm qt_{n-3}(1-\tfrac{\alpha}{2})\,(s^2_{y_{x^o}} + s^2)^{1/2}}$$

5. Comparaison de modèles de régression

- Le problème est de comparer sur un critère objectif les modèles $\boxed{y \sim a_0 + a_1 x_1}$, $\boxed{y \sim c_0 + c_2 x_2}$, $\boxed{y \sim b_0 + b_1 x_1 + b_2 x_2}$.

- Pourquoi ne pas utiliser R^2 ?

  Le modèle à deux variables sera toujours meilleur que les deux autres modèles sur le critère R^2 qui ne peux que "diminuer" quand on enlève une variable quelconque du modèle

- MSC_{Res} est un bon critère

- Deux autres critères sont aussi utilisés : R^2_{aj} et Cp de Mallows

Les critères R^2_{aj} et le Cp de Mallows

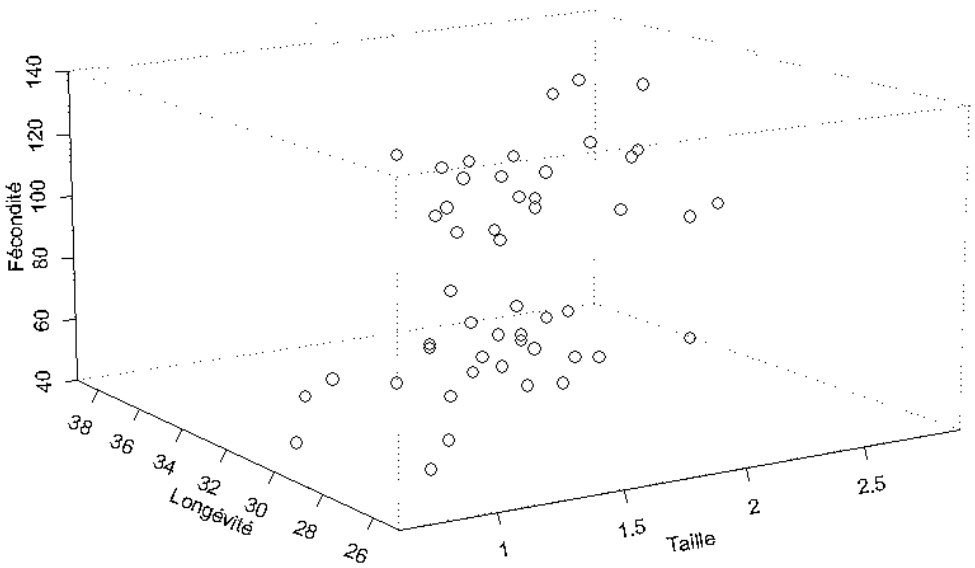
Notations : q est le nombre des variables du modèle, $SCE_{Res}(Mq)$ et $R^2(Mq)$ sont calculés pour le modèle à q variables explicatives, s^2 est l'estimation de σ^2 à partir du modèle complet

- $\boxed{R^2_{aj} = 1 - \frac{n-1}{n-q-1}[1 - R^2(Mq)]}$
  - Critère : $\boxed{R^2_{aj} \text{ proche de } 1}$
  - Justification : $R^2_{aj} = \frac{MCE_T(Mq) - MCE_{Res}(Mq)}{MCE_T(Mq)}$
- $\boxed{Cp = \frac{SCE_{Res}(Mq)}{s^2} - [n - 2(q+1)]}$
  - Critère : $\boxed{Cp \text{ proche de } q+1}$ mais attention pour le modèle complet (à 2 variables) $Cp = 3$ (en général $Cp = p + 1$)
  - Justification : $E(Cp) = q + 1$

6. Application à l'estimation de la fécondité des pucerons

- Présentation graphique des données - a-t'on une liaison linéaire entre *Fec* et *Taille* et/ou *Long* ?
- L'équation de régression, estimation des paramètres et qualité du modèle
- Analyse des résidus
- Les intervalles de confiance et de prédiction de la variable *Fec*

Description graphique des données



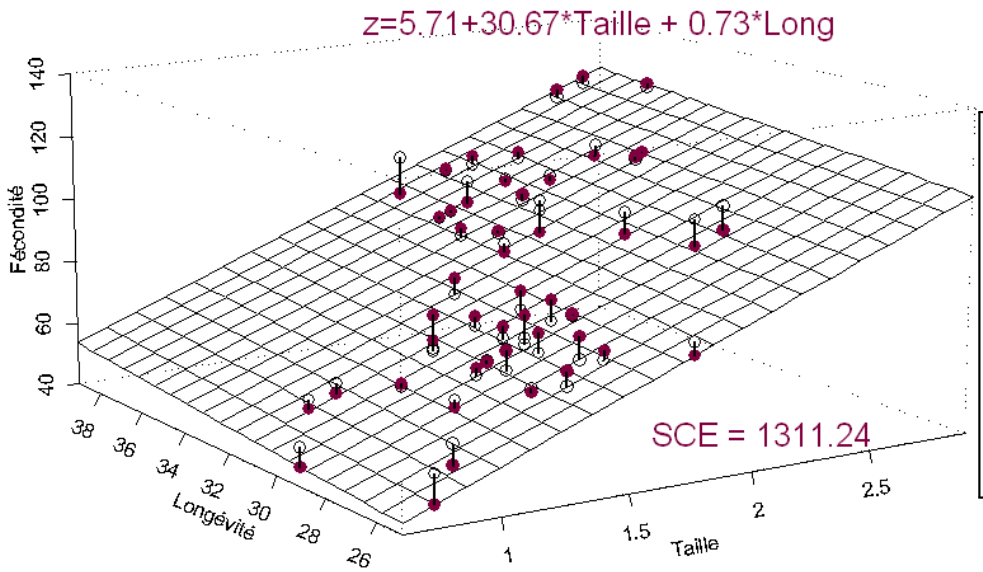
- Le graphique 3D peut être trompeur

  Le modèle adapté :

  $\boxed{Fec \sim Taille, Long}$

- L'étude graphique est possible pour un modèle à deux variables explicative - Mais ceci deviendra impossible à partir de 3 variables explicatives

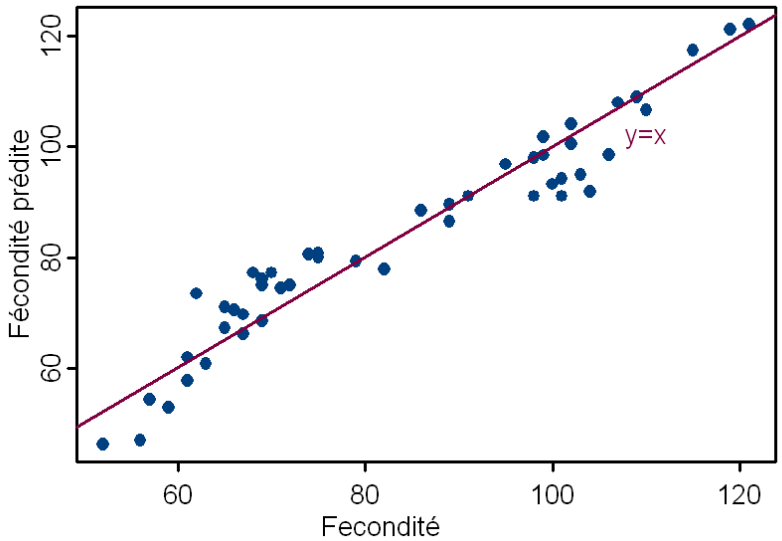
Le plan de régression

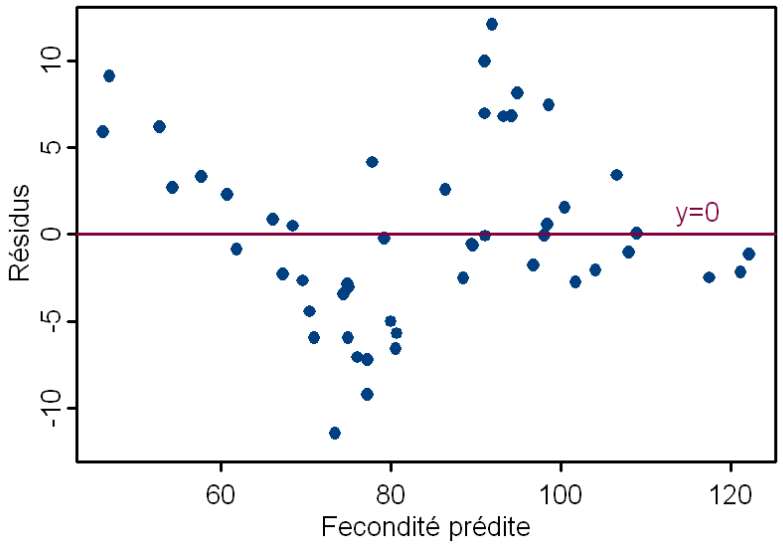


- Le modèle ajusté :
  $Fec = 5.71 + 30.67\,Taille + 0.73\,Long$
- $s^2 = 27.8987$
- $R^2 = 0.926$

Analyse des résidus

Les résidus : $\hat{e}_i = Fec_i - \widehat{Fec}_i$



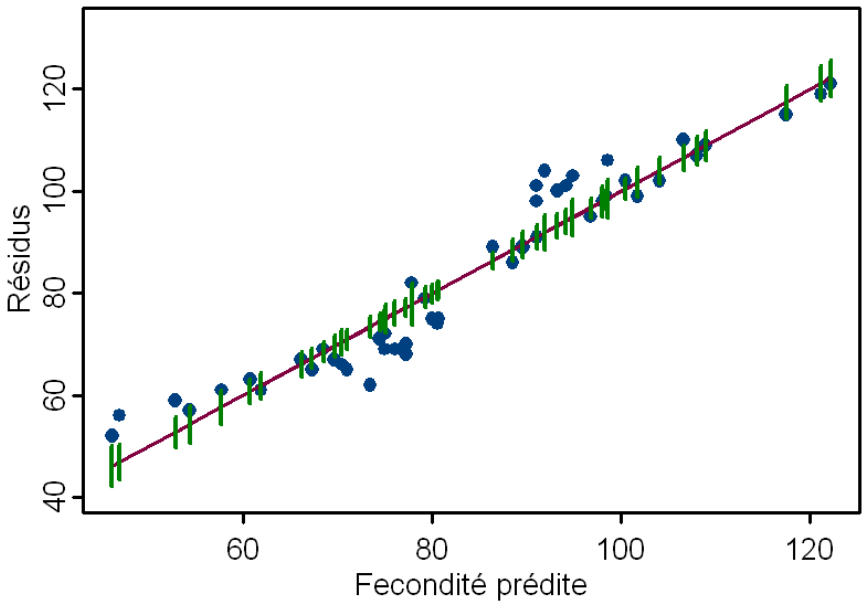


- Un nuage resserré autour de la diagonale indique un bon modèle de prédiction - Les deux graph nous permettent de détecter des problèmes d'hétéroscédasticité (postulat P1)
- Les individus suspects (◯) ou influents (△)

La prédiction et les intervalles de confiances pour la prédiction

Taille	*Long*	$\widehat{Fec}$	$IC_{0.95}(E(Fec))$		$IP_{0.95}(Fec)$	
2.1	30	91.927	89.274	94.5801	80.975	102.879

A noter que $s_{Fec_{xo}} = 5.444$, $s_{E(Fec_{xo})} = 1.319$ où $xo = (1, 2.1, 30)'$.



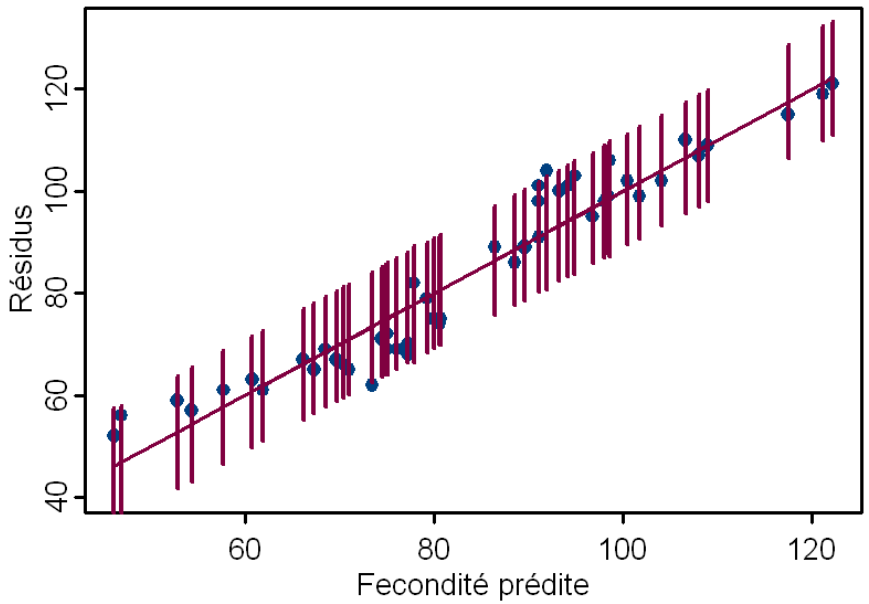


Tableau ANOVA et le test F

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	16526.9	2	8263.47	296.20	0.0000
Residual	1311.24	47	27.8987		
Total (Corr.)	17838.2	49			

R-squared = 92.6493 percent
R-squared (adjusted for d.f.) = 92.3365 percent
Standard Error of Est. = 5.28192

Soit $\alpha = 0.05$: L'hypothése multiple $H0 : b_{Taille} = b_{Long} = 0$ est rejetée car le Proba=0.0000<0.05.

Lecture équivalente. On rejette $H0 : b_{Taille} = b_{Long} = 0$ car $F = 296.20 > 3.195$, le quantile $1 - 0.05$ de $F_{2,47}$.

Tests sur les paramètres

```
Multiple Regression Analysis
-----------------------------------------------------------
Dependent variable: Fecondite
-----------------------------------------------------------
                           Standard       T
Parameter      Estimate       Error    Statistic    P-Value
-----------------------------------------------------------
CONSTANT         5.7111       6.925       0.8247     0.4137
Taille           30.666       1.998       15.351     0.0000
Long            0.72725       0.279       2.6071     0.0122
-----------------------------------------------------------
```

On fixe $\alpha = 0.05$: $H0 : b_{Taille} = 0$ est rejetée car Proba=0.0000<0.05 (2e ligne) du tableau (conclusion identique pour $H0 : b_{Long} = 0$ car Proba=0.0122 < 0.05.

L'hypothése $H0 : b_0 = 0$ est acceptée car Proba=0.4137>0.05.

On peut tester $H0 : b_{Taille} = 0$ en comparant $t_{b_{Taille}} = 15.351$ avec 2.012, le quantile $1 - \frac{0.05}{2}$ de t_{47}.

Comparaison des modèles régressifs

```
Regression Model Selection

Dependent variable: Fecondite
Independent variables: A=Taille B=Longevite

Number of complete cases: 50
      Number of models fit: 4
Model Results
--------------------------------------------------
                      Adjusted           Included
MSE      R-Squared  R-Squared    Cp     Variables
--------------------------------------------------
364.04       0.0        0.0     591.39
31.268       91.586     91.411  7.7972         A
164.29       55.792     54.871  236.67         B
27.899       92.649     92.337  3.0           AB
--------------------------------------------------
```

Le meilleur modèle est $Fec \sim Taille, Long$, car :

- $R^2_{aj} = 92.337$, $s^2 = 27.899$
- On ne peut pas se baser sur le "Cp de Mallows", il est égale 3 pour le modèle complet (à deux variables)