

Module 10 - Régression linéaire simple

Département de mathématiques et de statistique



UNIVERSITÉ
LAVAL

Introduction

Buts possibles d'une analyse de régression

Quels effets les variables explicatives X_1 , X_2 , X_3 , ... ont-elles sur la variable réponse Y ?

- Quels sont les effets de la météo, du jour de la semaine et du prix de l'essence sur le nombre de trajets effectués à vélo?
- Quel est l'impact de la vitesse du vent et de la hauteur des vagues sur les contraintes subies par une éolienne en mer?
- Quel est l'impact de la distance à l'université, du moyen de transport utilisé et de l'horaire de cours sur le taux de présence en classe?

Objectifs plus précis

Étudier les relations entre des variables en se basant sur des données.

- **Prévision** : *Étant donné son âge, son poids, fumeur/non fumeur, combien d'années un individu devrait-il survivre ?*
- **Sélection de variables** : *Parmi la température, l'ensoleillement, la pluie reçue, l'altitude, le trafic, quelles variables ont une influence significative sur l'orniérage des routes ?*
- **Spécification de modèle** : *Comment la durée de vie de transformateurs électriques varie-t-elle en fonction de leur grosseur ?*
- **Estimation de paramètres** : *La luminosité en fonction de la distance des étoiles d'une certaine galaxie est de la forme $L = K_1 + K_2 d + \sigma \epsilon$, où K_1 , K_2 et σ sont des paramètres inconnus à être estimés à partir d'observations.*

Modélisation

Exprimer la valeur de Y **en fonction** des valeurs de $X_1, \dots, X_k$:

$$Y = f(X_1, \dots, X_k) + \text{fluctuation aléatoire}$$



variable réponse
variable dépendante
variable endogène
issue



variables explicatives
variables indépendantes
variables exogènes
facteurs
covariables
“features”

Apprentissage statistique/machine

Estimer la fonction $f(\cdot)$ dans un modèle du type

$$Y = f(X_1, \dots, X_k) + \text{fluctuation aléatoire}$$

à partir de n observations (les informaticiens parlent “d'exemples”)

$(Y_1, X_{11}, \dots, X_{1k}), \dots, (Y_n, X_{n1}, \dots, X_{nk})$ est ce que l'on appelle de l'**apprentissage statistique** (apprentissage machine).

Exemple d'un modèle à une variable

$$Y = f(X) + \text{fluctuation aléatoire}$$



note finale
d'un étu-
diant dans
un certain
cours



nombre d'heures
d'étude de l'étu-
diant

(explique une
partie de la note)



deux étudiants peuvent avoir
étudié le même nombre
d'heures, mais avoir des
notes différentes

(partie de la note non expli-
quée par le temps d'étude)

Modèle linéaire

Un modèle de régression linéaire représente Y comme une **fonction linéaire** des paramètres :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + E,$$

Y : variable réponse

$X_1, \dots, X_k$: variables explicatives

$\beta_0, \dots, \beta_k$: paramètres de valeur inconnue à estimer à l'aide d'observations.

E : erreur aléatoire

La régression linéaire simple

Modèle de régression linéaire simple (RLS)

Une seule variable explicative X qui explique une partie de la valeur de Y
 n paires d'observations indépendantes $(X_1, Y_1), \dots, (X_n, Y_n)$, réparties autour d'une droite imaginaire.

$$Y_i = \beta_0 + \beta_1 X_i + E_i \\ (i = 1, \dots, n)$$

Postulats :

$E_i \sim N(0, \sigma^2)$ indép.

équivalent à

$Y_i|X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$ indép.

Illustration

À noter

Il est important de comprendre que nous obtenons des observations des paires de variables aléatoires $(X_1, Y_1), \dots, (X_n, Y_n)$. Le modèle suppose que ces paires sont générées ainsi :

- $X_1, \dots, X_n$ sont soit des valeurs fixées, soit n réalisations d'une variable aléatoire dont la loi ne nous intéresse habituellement pas.
- Sachant les X_i , les Y_i sont des variables aléatoires normales indépendantes obtenues en ajoutant des erreurs E_i i.i.d. $N(0, \sigma^2)$ à $\beta_0 + \beta_1 X_i$:
 $\Rightarrow Y_i = \beta_0 + \beta_1 X_i + E_i$ et, sachant X_i , on a $Y_i|X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$.

Dans ce cours, nous aurons que $\beta_0, \beta_1, \sigma^2$ sont des paramètres inconnus et X_i est une valeur que l'on considère fixe .

Les postulats du modèle

Conditions d'application à vérifier, en ordre d'importance, avant d'interpréter les intervalles de confiance et les tests d'hypothèses à venir :

P1 **linéarité** : l'espérance de Y_i est linéaire par rapport à X_i ;

$$E(E_i) = 0 \quad \Leftrightarrow \quad E(Y_i) = \beta_0 + \beta_1 X_i$$

P2 **homoscédasticité** : la variance de Y_i autour de la droite est homogène (elle ne dépend pas de X_i);

$$V(E_i) = \sigma^2 \quad \Leftrightarrow \quad V(Y_i) = \sigma^2$$

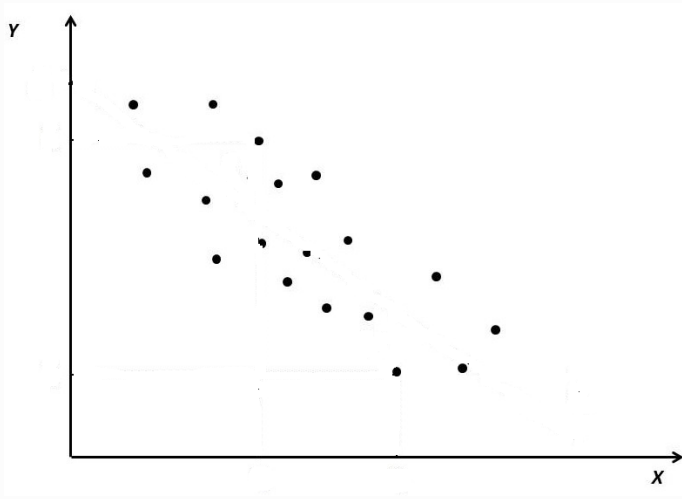
P3 **indépendance** : les observations sont indépendantes;

$$E_i \text{ indépendants} \quad \Leftrightarrow \quad Y_i \text{ indépendants}$$

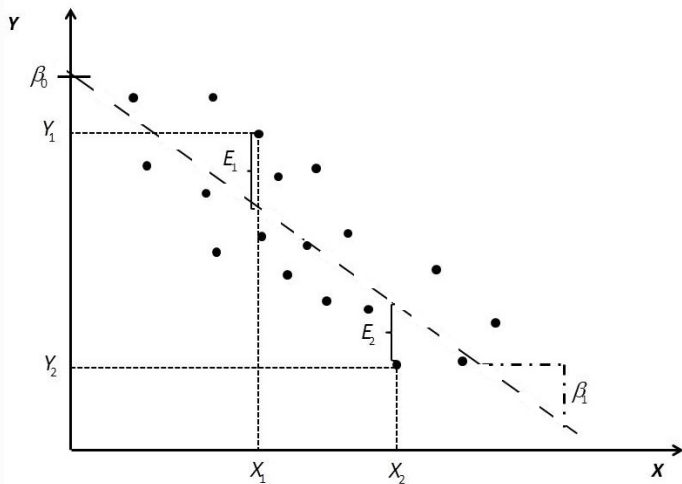
P4 **normalité** : les Y_i sont de loi normale autour de la droite.

$$E_i \sim \text{normale} \quad \Leftrightarrow \quad Y_i \sim \text{normale}$$

Éléments du modèle de RLS



Éléments du modèle de RLS



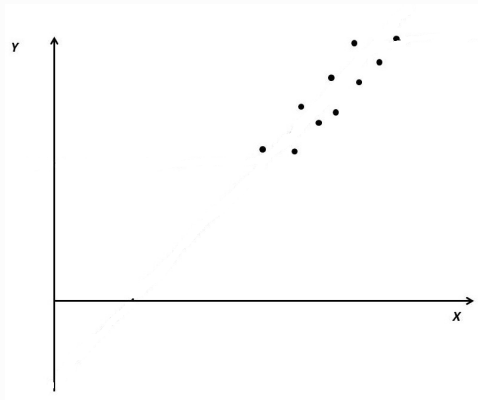
Interprétation des paramètres β_0 et β_1

Si le postulat de linéarité (P1) est vrai :

- β_1 : Si on augmente X d'une unité, alors la valeur moyenne de Y augmente de β_1 unités.
 - $\beta_1 > 0$: grands X associés aux grands Y et vice-versa (corrélation positive entre les variables).
 - $\beta_1 < 0$: petits X associés aux grands Y .
- β_0 : la valeur moyenne de Y quand $X = 0$.
 - Si la valeur $X = 0$ est peu réaliste ou hors du domaine des données observées, β_0 peut représenter une valeur peu significative ou aberrante. Néanmoins, le modèle de régression linéaire simple demeure pertinent pour les interprétations et prédictions lorsque X est proche des valeurs effectivement observées.

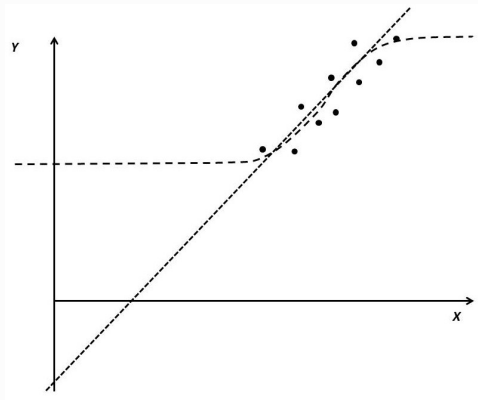
Extrapolation dangereuse!

ATTENTION!!!! On **ignore** la forme de la relation entre X et Y à l'extérieur de l'intervalle observé pour les valeurs de $X_1, \dots, X_n$.



Extrapolation dangereuse!

ATTENTION!!!! On ignore la forme de la relation entre X et Y à l'extérieur de l'intervalle observé pour les valeurs de $X_1, \dots, X_n$.



Estimation des paramètres β_0 et β_1

β_0 et β_1 : paramètres de valeurs inconnues.

$\hat{\beta}_0$ et $\hat{\beta}_1$: estimateurs des paramètres, obtenus en **minimisant la partie inexpliquée de Y** (la fluctuation aléatoire) :

$$SS_E = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2,$$

où

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ est la i^e **valeur ajustée**, ou **valeur prédite**
(la partie de Y_i expliquée par le modèle; hauteur de la ligne de régression estimée en $x = X_i$)

$e_i = Y_i - \hat{Y}_i$ est le i^e **résidu**
(la partie de Y_i inexpliquée par le modèle; différence entre Y_i et hauteur de la ligne estimée de régression en $x = X_i$).

Problème de minimisation

En annulant les dérivées de SS_E par rapport à $\hat{\beta}_0$ et $\hat{\beta}_1$, on obtient

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$$

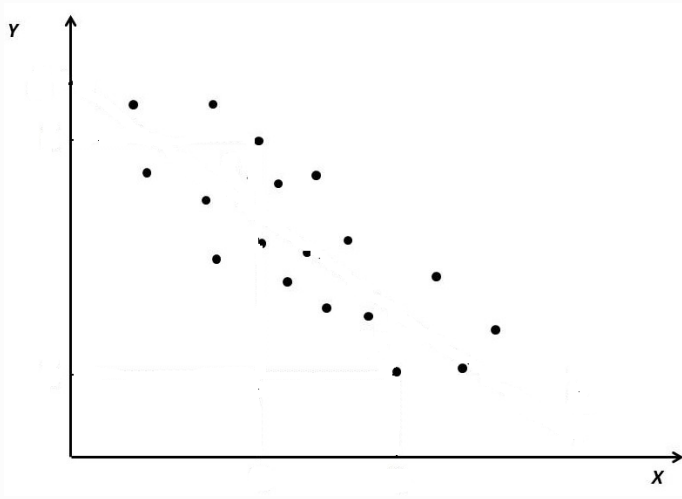
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

où

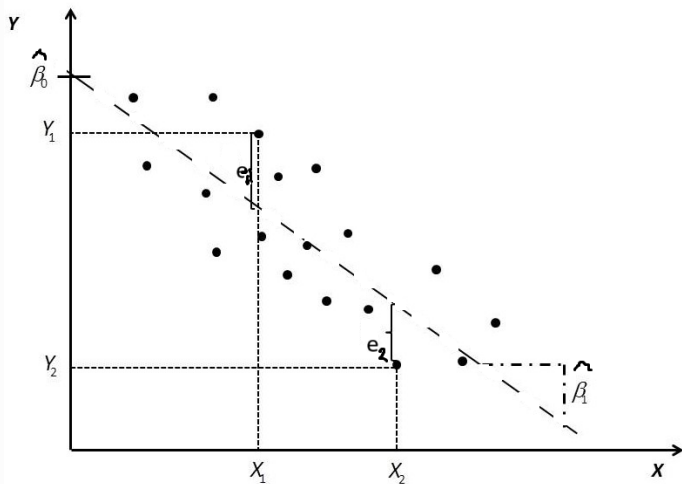
$$S_{XX} = \sum_{i=1}^n (X_i - \bar{X})^2 = (n-1) s_X^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2$$

$$S_{XY} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n (X_i - \bar{X})Y_i = \sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}.$$

Éléments du modèle de RLS

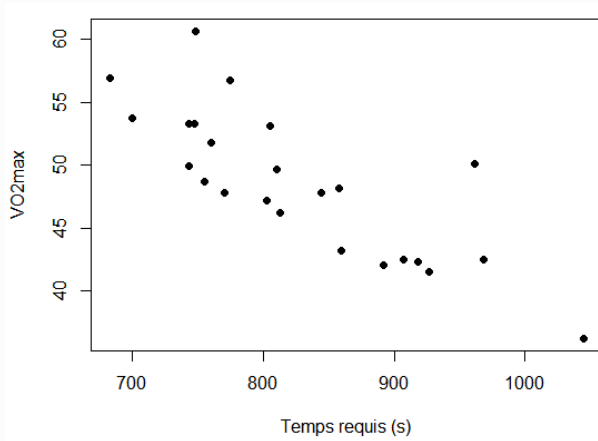


Éléments du modèle de RLS



Exemple 1 : Absorption d'oxygène chez 24 hommes qui courent 3,2 km

SUJET	VOLUME MAX. D'O ₂ (Y)	TEMPS EN SECONDES (X)
1	42.33	918
2	53.10	805
3	42.08	892
4	50.06	962
5	42.45	968
6	42.46	907
7	47.82	770
8	49.92	743
9	36.23	1045
10	49.66	810
11	41.49	927
12	46.17	813
13	48.18	858
14	43.21	860
15	51.81	760
16	53.28	747
17	53.29	743
18	47.18	803
19	56.91	683
20	47.80	844
21	48.65	755
22	53.69	700
23	60.62	748
24	56.73	775



Exemple 1 : Estimation des paramètres

On a effectué les calculs suivants :

$$\bar{x} = 826,5$$

$$s_X^2 = 8593,8$$

$$\sum_{i=1}^{24} x_i y_i = 952\,885$$

$$\bar{y} = 48,55$$

$$s_Y^2 = 34,1$$

Quelle est l'équation de la droite de régression estimée ?

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

QUIZ!

- Jacques a couru les 3,2 km en 10 secondes de plus que Marc, mais ils n'ont pas mesuré leur VO_2max . Quelle différence de VO_2max le modèle ajusté prédit-il entre les deux hommes?
- Vrai ou Faux : Il est raisonnable de dire que les hommes qui courent le 3,2 km en 0 seconde ont un VO_2max moyen de 90,7.

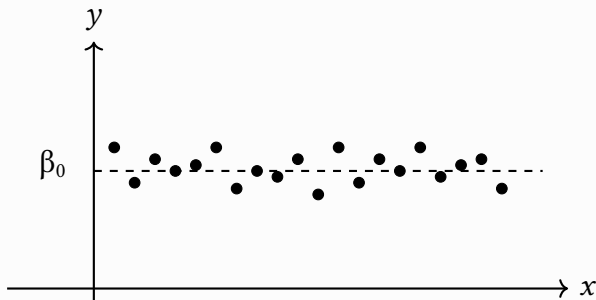
Pour votre curiosité : ChatGPT nous informe que des VO_2max de 90,7 ne sont pas complètement impossibles, mais absolument exceptionnels et n'ont été atteints que par quelques fondeurs ou cyclistes d'exception (p.ex., on évalue celui de Pogacar à 89,4).

Et un temps de 0 seconde est très certainement plus que farfelu! Jakob Ingebrigtsen détient la meilleure performance mondiale sur 2 miles ($\approx 3,218$ km) avec un temps de 474 secondes ... son VO_2max est estimé à 70-80.

Relation non significative entre Y et X

Une pente nulle pour la droite de régression ($\beta_1 = 0$) signifie qu'il n'y a pas d'association entre les variables Y et X.

⇔ La loi de Y ne change pas avec la valeur de X (la valeur espérée de Y est égale à β_0 , peu importe la valeur de X.)



Valeur de $\hat{\beta}_1$

ATTENTION!!! Il faut résister à l'envie de dire que si une valeur observée de $\hat{\beta}_1$ est “près de zéro”, on a une faible association entre Y et X.

P.ex. si on obtient $\hat{\beta}_1 = 2,35$ avec X mesuré en mètres, si on refait exactement la même analyse avec les mêmes données (donc association entre Y et X inchangée), mais avec X mesuré en mm, on obtiendra $\hat{\beta}_1 = 0,00235$.

La solution : Faire des inférences statistiques formelles (tests d'hypothèses, intervalles de confiance) sur la valeur de β_1 .

Tests sur la significativité de la relation linéaire

Il existe deux approches équivalentes pour tester

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0.$$

1) Analyse de la variance (Test F)

2) Test t sur β_1 ou IC(β_1)

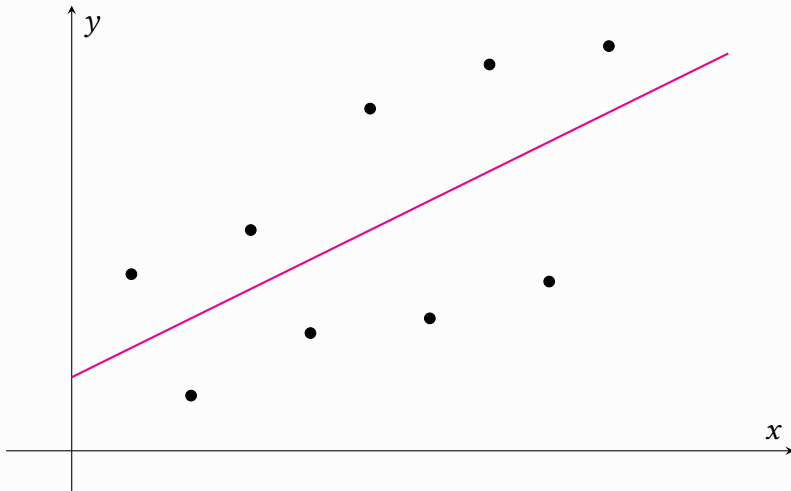
1) Approche basée sur l'analyse de la variance

On décompose la variabilité des $Y_1, \dots, Y_n$ en deux sources :

$$\underbrace{\sum_{i=1}^n (Y_i - \bar{Y})^2}_{SS_T} = \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}_{SS_R} + \underbrace{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}_{SS_E}$$

Variabilité totale = Variabilité due au modèle + Variabilité résiduelle.
(variabilité expliquée par X)

Illustration de la décomposition de la variabilité



Trucs calculatoires

$$SS_T = \sum_{i=1}^n (Y_i - \bar{Y})^2 = S_{YY} = (n-1)s_Y^2 = \sum_{i=1}^n Y_i^2 - n\bar{Y}^2$$

$$SS_R = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_1 S_{XY} = \frac{(S_{XY})^2}{S_{XX}} = \frac{\left(\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y} \right)^2}{\sum_{i=1}^n X_i^2 - n\bar{X}^2}$$

$$SS_E = SS_T - SS_R$$

Tableau d'analyse de la variance

On résume cette décomposition dans une table d'anova :

Source de variabilité	Somme des carrés	Degrés de liberté	Carrés moyens	F_0	Valeur P
Régression	SS_R	1	$MS_R = \frac{SS_R}{1}$	$F_0 = \frac{MS_R}{MS_E}$	$P(F_{1, n-2} > F_0)$
Erreur	SS_E	$n - 2$	$MS_E = \frac{SS_E}{n - 2}$		
Totale	SS_T	$n - 1$			

Règle de décision du test

Si $H_0 : \beta_1 = 0$ est vraie (pas d'association entre X et Y) :

- la majeure partie de la variabilité dans les Y_i sera expliquée par la fluctuation aléatoire
- $SS_T \approx SS_E$, et MS_R prendra une petite valeur
- $F_0 = \frac{MS_R}{MS_E} = \frac{SS_R/1}{SS_E/(n-2)} \sim F_{1, n-2}$

Si $H_1 : \beta_1 \neq 0$ est vraie (relation linéaire positive ou négative entre X et Y) :

- une bonne partie de la variabilité des Y_i sera expliquée par X
- MS_R prendra une valeur assez supérieure à MS_E (F_0 sera grand)

Ainsi on rejette $H_0 : \beta_1 = 0$ au seuil α lorsque $f_0 > F_{\alpha;1, n-2}$.

Illustration

QUIZ!

Vrai ou faux ?

- Quand $H_0 : \beta_1 = 0$ est vraie, les valeurs prédites de Y sont “proches” de la moyenne $\bar{y}$.
- Quand $H_0 : \beta_1 = 0$ est vraie, l'estimation de la pente ($\hat{\beta}_1$) vaut exactement 0.

2) Approche basée sur la loi t

On peut utiliser le fait que

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = \sum_{i=1}^n \left[\frac{(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] Y_i$$

est une combinaison linéaire des $Y_i \Rightarrow$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right) \Leftrightarrow \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{XX}}}} \sim N(0, 1)$$

et ensuite

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_E}{S_{XX}}}} \sim t_{n-2}.$$

Règles de décision

On peut facilement déduire les tests d'hypothèses de niveau α suivants pour $H_0 : \beta_1 = c$.

$$\text{Soit } T_0 = \frac{\hat{\beta}_1 - c}{\sqrt{MS_E/S_{XX}}}.$$

$$H_1 : \beta_1 \neq c \Rightarrow \text{rejeter } H_0 \text{ si } |t_0| > t_{\alpha/2; n-2}$$

$$H_1 : \beta_1 > c \Rightarrow \text{rejeter } H_0 \text{ si } t_0 > t_{\alpha; n-2}$$

$$H_1 : \beta_1 < c \Rightarrow \text{rejeter } H_0 \text{ si } t_0 < -t_{\alpha; n-2}$$

Règles de décision

Le cas qui nous intéressera le plus souvent est $H_0 : \beta_1 = 0$.

$$\text{Donc } T_0 = \frac{\hat{\beta}_1}{\sqrt{MS_E/S_{XX}}}.$$

$H_1 : \beta_1 \neq 0 \Rightarrow$ rejeter H_0 si $|t_0| > t_{\alpha/2;n-2}$ équivalent à l'ANOVA

$H_1 : \beta_1 > 0 \Rightarrow$ rejeter H_0 si $t_0 > t_{\alpha;n-2}$

$H_1 : \beta_1 < 0 \Rightarrow$ rejeter H_0 si $t_0 < -t_{\alpha;n-2}$

Exemple 2 : significativité de la relation linéaire

R Commander permet d'effectuer facilement le test F et le test t de

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

On peut aussi les faire “à la main”.

1) ANOVA

$$SS_T = (n - 1)s_y^2$$

$$SS_R = \frac{S_{XY}^2}{S_{XX}}$$

$$SS_E = SS_T - SS_R$$

Exemple 2

Source de variabilité	Somme des carrés	Degrés de liberté	Carrés moyens	F_0	Valeur P
Régression					
Erreur					
Totale					

Example 2

2) Test t

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Intervalles de confiance et de prédiction

Estimateurs de β_0 et β_1

$$Y_i = \beta_0 + \beta_1 X_i + E_i, \quad E_i \sim N(0, \sigma^2)$$

Paramètre	Pente	Ordonnée à l'origine
Estimateur	$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
Espérance	$E(\hat{\beta}_1) = \beta_1$	$E(\hat{\beta}_0) = \beta_0$
Variance	$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{XX}}$	$V(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{X}^2}{S_{XX}}$
Stat. inférence	$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_E}{S_{XX}}}} \sim t_{n-2}$	$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{MS_E \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)}} \sim t_{n-2}$

Intervalle de confiance pour β_1

On a que

$$P\left(-t_{\alpha/2;n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MS_E/S_{XX}}} \leq t_{\alpha/2;n-2}\right) = 1 - \alpha.$$

En isolant β_1 au centre, on obtient l'intervalle de confiance de niveau $1 - \alpha$ suivant pour β_1 :

$$\hat{\beta}_1 \pm t_{\alpha/2;n-2} \sqrt{\frac{MS_E}{S_{XX}}}.$$

Intervalle de confiance pour β_0

$\hat{\beta}_0$ est aussi une combinaison linéaire des Y_i , donc

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\text{MS}_E \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)}} \sim t_{n-2}.$$

On obtient l'intervalle de confiance de niveau $1 - \alpha$ pour β_0 :

$$\hat{\beta}_0 \pm t_{\alpha/2; n-2} \sqrt{\text{MS}_E \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)}.$$

Exemple

```
lm(formula = VO2max ~ Temps, data = Oxygene)

Residuals:
    Min       1Q   Median       3Q      Max
-3.6272 -2.6463 -0.6485  1.0615  8.4277

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  90.721717   6.546373   13.858 2.39e-12 ***
Temps        -0.051028   0.007873   -6.481 1.61e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.5 on 22 degrees of freedom
Multiple R-squared:  0.6563, Adjusted R-squared:  0.6407
F-statistic: 42.01 on 1 and 22 DF,  p-value: 1.608e-06

> Confint(RegModel.1, level=0.95)
              Estimate      2.5 %      97.5 %
(Intercept)  90.72171668  77.14537040 104.29806296
Temps        -0.05102849 -0.06735667 -0.03470032
```

Estimation de $E(Y|X = x_0)$

- Pour un X fixé à x_0 , la valeur de Y tourne autour de $\beta_0 + \beta_1 x_0$.
- On veut faire de l'inférence sur la valeur moyenne de Y à ce niveau $X = x_0$, c'est-à-dire sur $E(Y|X = x_0)$.
- Un estimateur sans biais de la vraie valeur moyenne $\beta_0 + \beta_1 x_0$ est

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Intervalle de confiance pour la moyenne de Y en x_0

On peut montrer que

$$V(\hat{Y}_0) = V(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{XX}} \right]$$

et que

$$\frac{\hat{Y}_0 - (\beta_0 + \beta_1 x_0)}{\sqrt{MS_E \left[\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{XX}} \right]}} \sim t_{n-2}$$

et donc un intervalle de confiance de niveau $1 - \alpha$ pour $E(Y|X = x_0)$ est donné par

$$\hat{Y}_0 \pm t_{\alpha/2; n-2} \sqrt{MS_E \left[\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{XX}} \right]}.$$

Une question un peu différente

Au lieu d'estimer la valeur moyenne de Y en x_0 , supposons qu'on veut **prédire la valeur d'une observation individuelle de Y** pour un X fixé.

Mathématiquement, on cherche à prédire la valeur de $\beta_0 + \beta_1 x_0 + E_0$.

L'estimateur ponctuel d'une observation individuelle est

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 + 0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

(le même que pour la valeur moyenne de Y ... mais sa variance sera plus grande.)

Intervalle de prévision de Y quand $X = x_0$

On peut montrer que

$$\frac{\hat{Y}_0 - (\beta_0 + \beta_1 x_0 + E_0)}{\sqrt{MS_E \left[\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{XX}} \right] + MS_E}} \sim t_{n-2},$$

ce qui mène à l'intervalle de prévision pour $\beta_0 + \beta_1 x_0 + E_0$

$$\hat{Y}_0 \pm t_{\alpha/2; n-2} \sqrt{MS_E \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{XX}} \right]}.$$

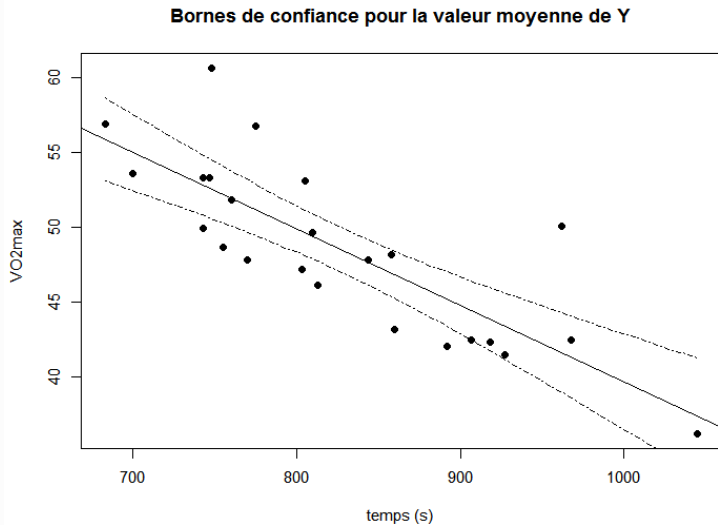
Exemple 3

- (i) Calculons une prévision ponctuelle et un intervalle de confiance à 95% pour la valeur moyenne de l'absorption maximale d'oxygène (VO_{2max}) chez les hommes qui courent les 3,2 km en 15 minutes et 25 secondes;

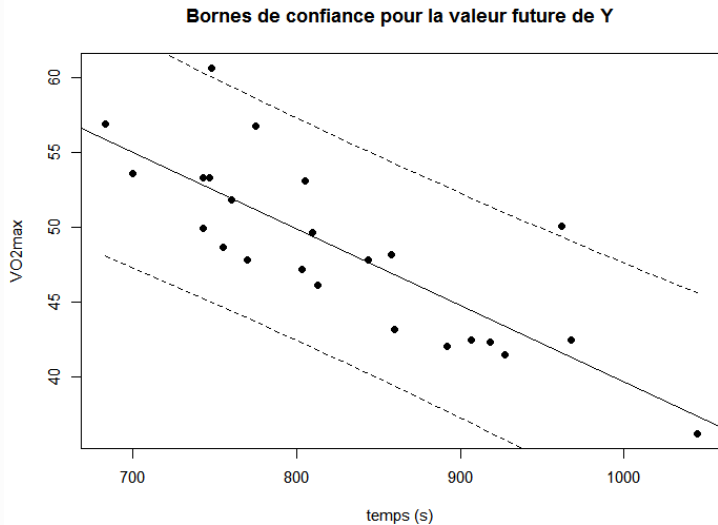
Exemple 3

- (ii) Calculons une prévision ponctuelle et un intervalle de prévision à 95% pour le VO_2max d'un homme qui court 3,2 km en 15 min 25 s.

Exemple 3 (suite)



Exemple 3 (suite)



Corrélation et qualité de l'ajustement

Corrélation entre les valeurs de Y et X

Considérons à nouveau les X_i comme des variables aléatoires. Si nous retournons aux nuages de points du module 4 sur les lois conjointes, on se doute qu'il devrait y avoir un fort lien entre la valeur du coefficient de régression β_1 et la valeur du coefficient de corrélation $\rho = \text{corr}(X, Y)$.

Illustration

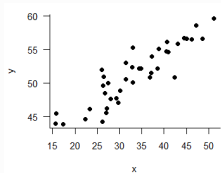
Estimation de ρ

On peut montrer que

$$\rho = \frac{\beta_1 \sigma_X}{\sqrt{\beta_1^2 \sigma_X^2 + \sigma^2}} = \beta_1 \frac{\sigma_X}{\sigma_Y}.$$

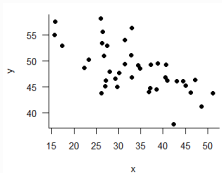
ρ est estimé par le coefficient de corrélation échantillonnal r :

$$\hat{\rho} = r = \hat{\beta}_1 \frac{s_X}{s_Y} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}.$$



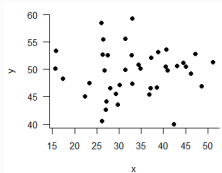
$\rho_{\text{simulé}} = 0,9$

$r = 0,90$



$-0,6$

$-0,640$



$0,1$

$0,041$

Le coefficient de détermination R^2

Est-ce que X "explique bien" Y?

Le coefficient de détermination (R^2), soit **la proportion de la variabilité dans les Y_i qui est expliquée par le modèle de régression** (donc par X), est une quantité très utilisée pour juger de la qualité d'un modèle de régression linéaire.

$$\begin{aligned} R^2 &= \frac{SS_R}{SS_T} \\ &= \frac{S_{XY}^2 / S_{XX}}{S_{YY}} = \frac{S_{XY}}{S_{XX}} \frac{S_{XY}}{S_{YY}} \\ &= \hat{\beta}_1 \frac{S_{XY}}{S_{YY}} = \hat{\beta}_1^2 \frac{S_{XX}}{S_{YY}} \\ &= r^2. \end{aligned}$$

Exemple 4 sur le VO_2max

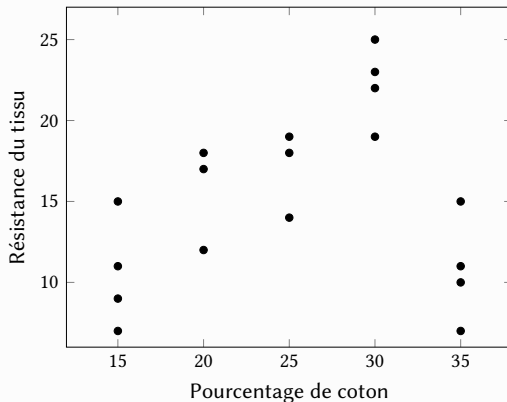
Coefficient de détermination :

Coefficient de corrélation échantillonnal :

Validation du modèle : Analyse des résidus

Une question avant de débiter

Si on reprend l'exemple de la résistance du tissu du module 9, aurait-on pu utiliser la régression linéaire pour répondre à la question posée ?



Validation du modèle : Analyse des résidus

La validité des conclusions de l'analyse statistique du modèle de régression dépend du respect des postulats (P1) à (P4).

P1 **linéarité** : l'espérance de Y_i est linéaire en X_i ;

P2 **homoscédasticité** : $V(E_i) = V(Y_i) = \sigma^2$ (ne dépend pas de X_i);

P3 **indépendance** : $E_1, \dots, E_n$ indépendants (idem pour Y_i);

P4 **normalité** : $E_i \sim$ loi normale autour de la droite.

Comme ces postulats portent tous sur le comportement aléatoire des termes d'erreur E_i , nous les vérifierons à l'aide des résidus

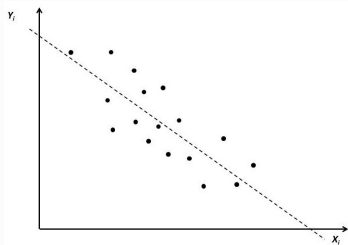
$$e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) = Y_i - \hat{Y}_i.$$

(P1) Linéarité : Validation graphique

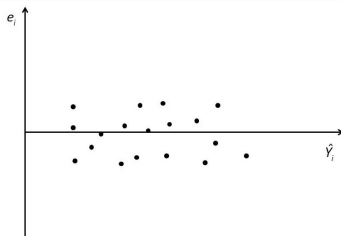
- Un graphique de dispersion des paires (x_i, y_i) : devrait montrer un nuage de points distribués de façon symétrique autour d'une ligne droite.
- Un graphique des paires $(\hat{Y}_i, e_i)$: devrait montrer un nuage de points distribués symétriquement autour de l'axe des abscisses.

Exemples de nuages de points

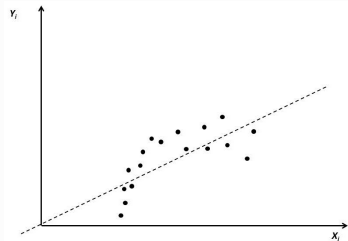
Linéarité raisonnable



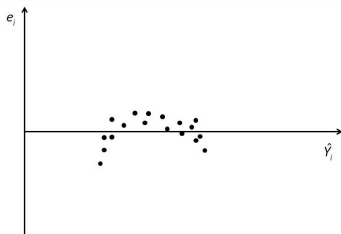
Linéarité raisonnable



Linéarité déraisonnable



Linéarité déraisonnable

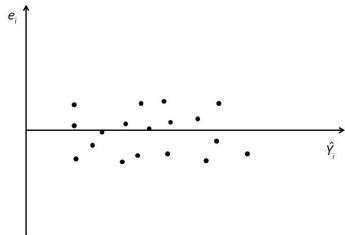


(P2) Homoscédasticité

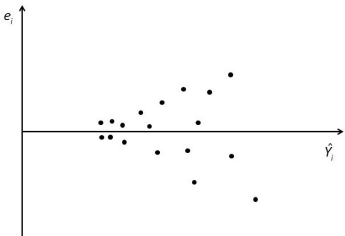
- Un graphique des e_i en fonction des $\hat{Y}_i$: devrait montrer que la variabilité des e_i demeure constante pour toute valeur de $\hat{Y}_i$.

Exemples de nuages de points

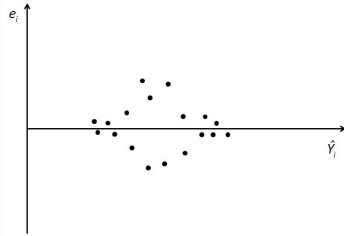
Homoscédasticité raisonnable



Homoscédasticité déraisonnable



Homoscédasticité déraisonnable



(P3) Indépendance : Difficile à vérifier !

C'est la collecte des données qui en est garante. Il faut s'assurer qu'**aucune variable à part X_i n'ait pu avoir une influence sur les valeurs observées.**

Exemple d'une mauvaise collecte :

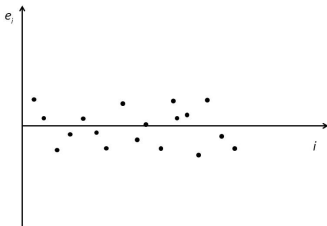
- les observations 1, 2, 3, 4 sont prises par l'employé A qui a tendance à surestimer Y;
- les mesures 5, 6, 7 et 8 sont prises par l'employé B qui a tendance à sous-estimer Y;
- Problème : les observations prises par un même employé seront corrélées.

Données chronologiques

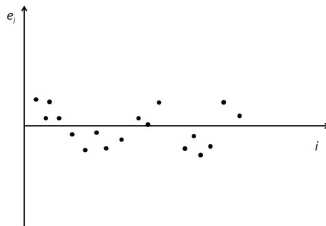
Il peut aussi y avoir de la corrélation lorsque les observations sont prises de façon trop rapprochées dans le temps (ou l'espace).

Si les valeurs de $i = 1, \dots, n$ correspondent à un ordre chronologique, le **graphique des e_i en fonction de i** pourrait révéler un problème de corrélation positive entre les résidus rapprochés (“effet de grappe”).

Pas d'effet chronologique



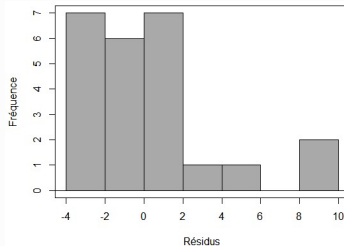
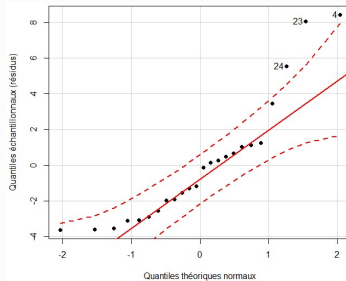
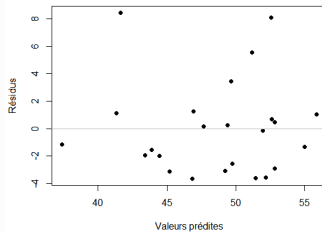
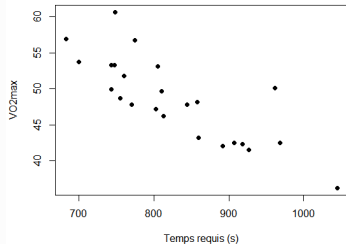
Effet de grappe



(P4) Normalité

- histogramme des résidus
- graphique quantile-quantile normal des résidus
- diagramme en boîte des résidus

Exemple 5 sur l'absorption d'oxygène



Que faire si les postulats ne sont pas respectés ?

Lorsque les postulats de linéarité, d'homoscédasticité ou de normalité ne sont pas respectés, on tente de **transformer Y ou X** pour régler le problème.

- Problème de non-linéarité : on peut s'inspirer de la forme du nuage de points pour trouver une relation plus représentative de la tendance. Par exemple :

$$\blacksquare X_i^* = X_i^2 \quad \Rightarrow \quad Y_i = \beta_0 + \beta_1 X_i^2 + E_i$$

$$\blacksquare X_i^* = \ln(X_i) \quad \Rightarrow \quad Y_i = \beta_0 + \beta_1 \ln(X_i) + E_i$$

$$\blacksquare X_i^* = 1/X_i \quad \Rightarrow \quad Y_i = \beta_0 + \beta_1 [1/X_i] + E_i$$

$$\blacksquare Y_i^* = \sqrt{Y_i} \quad \Rightarrow \quad \sqrt{Y_i} = \beta_0 + \beta_1 X_i + E_i$$

$$\blacksquare Y_i^* = \ln(Y_i) \quad \Rightarrow \quad \ln(Y_i) = \beta_0 + \beta_1 X_i + E_i$$

- Parfois, il faut transformer et **X et Y** pour parvenir à une relation linéaire. On peut aussi ajuster un modèle polynomial avec les outils de la régression multiple.

Que faire si les postulats ne sont pas respectés ?

- Problème d'hétéroscédasticité ou de non-normalité : un changement dans l'échelle de Y créé par **une transformation de Y** pourra parfois stabiliser la variance ou améliorer la normalité. Par exemple :
 - $Y_i^* = \sqrt{Y_i} \Rightarrow \sqrt{Y_i} = \beta_0 + \beta_1 X_i + E_i$
 - $Y_i^* = \ln(Y_i) \Rightarrow \ln(Y_i) = \beta_0 + \beta_1 X_i + E_i$
 - $Y_i^* = 1/Y_i \Rightarrow 1/Y_i = \beta_0 + \beta_1 X_i + E_i$
- Problème de dépendance entre les observations : refaire l'échantillonnage (!) ou utiliser des modèles statistiques qui prennent en compte des structures de dépendance entre les données (non couverts dans ce cours).

Réponses aux exemples

1. $\hat{Y} = 90,7 - 0,051X$

2. Anova : Valeur $P = P(F_{1,22} > 42) \approx 0$. On rejette $H_0 : \beta_1 = 0$, le lien linéaire est significatif.

Test t : Puisque $|t_{obs}| = |-6,48| > t_{0,025;22} = 2,074$, on rejette H_0 (même valeur P et conclusion que test F).

3. $43,5 \pm 2,19$

$43,5 \pm 7,58$

4. $R^2 = 0,6563$, $r = -0,81$

5. linéarité du lien et homoscedasticité respectées, normalité des erreurs problématique