

Module 10 - Régression linéaire simple

STT-1000 Probabilités et statistique

Département de mathématiques et de statistique



UNIVERSITÉ
LAVAL

Introduction

But d'une analyse de régression

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

- Quel est l'effet de la température sur la durée de vie de la peinture sur l'asphalte?
- Quel est l'effet d'un hiver très froid sur le niveau des réservoirs d'eau d'une ville en été?
- Quels sont les effets des heures de sommeil, de sport et d'étude sur la note moyenne d'un étudiant?

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 la vitesse de 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** : *Sachant que la luminosité L d'une étoile d'une certaine galaxie varie en fonction de sa distance d selon $L = K_1 + K_2 d + \sigma \epsilon$, quelles sont les valeurs des paramètres K_1 , K_2 et σ ?*

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



variable explicative
variables indépendantes
variables exogènes
facteurs
covariables

(Ce sont les termes **variable réponse** et **variable explicative** qui seront réutilisés dans la suite du cours.)

Exemple d'un modèle à une variable

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

masse d'un
homme de
30 ans (kg)

taille de l'homme
(cm)

(explique une
partie de la
masse)

deux hommes de 30 ans de
même taille peuvent avoir
une masse différente

(variation de la masse non
expliquée par la taille)

Modèle linéaire

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

$$Y = \beta_0 + \beta_1 f_1(X_1) + \beta_2 f_2(X_2) + \dots + \beta_k f_k(X_k) + E,$$

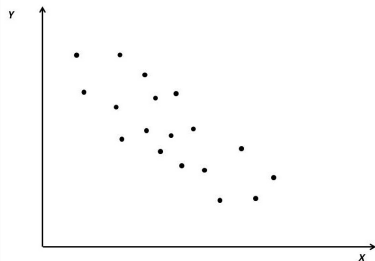
- Y : variable réponse
- $X_1, \dots, X_k$: variables explicatives
- $f_1, \dots, f_k$: transformations connues des variables explicatives
- $\beta_0, \dots, \beta_k$: paramètres de valeurs inconnues à estimer à l'aide d'observations.
- E : erreur aléatoire

La régression linéaire simple

Modèle de régression linéaire simple (RLS)

Une seule variable explicative X , liée à Y par une droite.

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 :

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

équivalent à

$E_i \sim N(0, \sigma^2)$ indép.

Les postulats du modèle

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

P1 **linéarité** : la moyenne de Y_i est linéaire par rapport à X_i ;

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

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

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

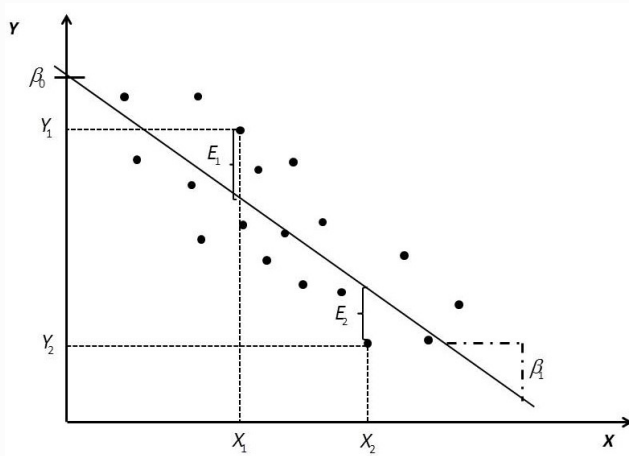
P3 **indépendance** : les observations sont indépendantes;

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

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

$$\Rightarrow \text{les } E_i \text{ sont de loi normale autour de } 0$$

Éléments du modèle de RLS



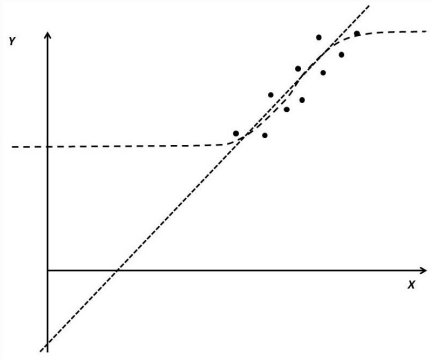
Interprétation des paramètres β_0 et β_1

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 $X = 0$ est peu plausible en pratique, β_0 peut prendre une valeur farfelue, mais le modèle de régression linéaire simple reste adéquat pour les valeurs de X près de celles observées.

Extrapolation dangereuse!

Soit Y , l'exactitude (*accuracy*, en %) d'un modèle d'apprentissage automatique, et X , la taille (en milliers d'exemples) de l'ensemble d'entraînement, mesurées pour X entre 1 et 50. Dans cet intervalle, la relation semble linéaire. On ignore sa forme à l'extérieur de cet intervalle : extrapoler linéairement serait dangereux, car l'exactitude ne peut pas dépasser 100 %, alors que le modèle linéaire prédirait une hausse indéfinie.



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)
- $e_i = Y_i - \hat{Y}_i$ est le i^e **résidu** (la partie de Y_i inexpliquée par le modèle).

Problème de minimisation

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

$$\begin{aligned}\hat{\beta}_1 &= \frac{S_{XY}}{S_{XX}} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}\end{aligned}$$

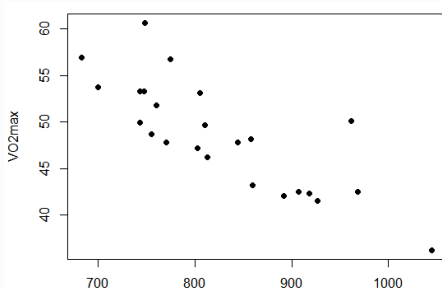
où

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

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

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 :

$$\begin{array}{lll} \bar{x} = 826,5 & s_X^2 = 8593,8 & \sum_{i=1}^{24} x_i y_i = 952\,885 \\ \bar{y} = 48,55 & s_Y^2 = 34,1 & \end{array}$$

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 : interprétation du modèle

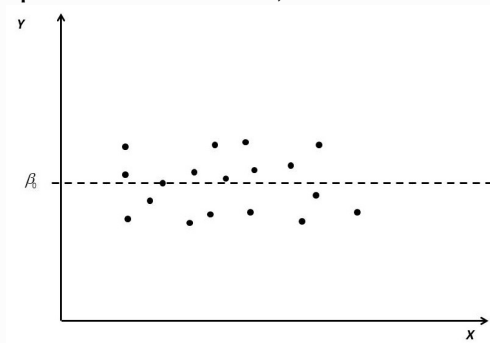
QUIZ!

- 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 : on peut dire que les hommes qui courent le 3,2 km en 0 seconde ont un VO_2max moyen de 90,7.

Relation non significative entre Y et X

Si le postulat de linéarité est respecté, alors **une pente nulle pour la droite de régression ($\beta_1 = 0$) signifie qu'il n'y pas d'association entre les variables Y et X.**

Dans ce cas, la loi de Y ne change pas avec la valeur de X (la valeur moyenne de Y est égale à β_0 peu importe la valeur de X).



Tests sur la significativité de la relation linéaire

Il existe deux approches équivalentes pour tester :

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

Il s'agit du test F, basé sur l'analyse de la variance et du test t sur β_1 .

Dans ce cours, nous ne verrons que le test t .

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 $Y_i \Rightarrow$

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

Par contre, nous ne connaissons pas σ^2 . Il faut donc l'estimer.

Estimation de σ^2

Comme σ^2 est la variance des erreurs du modèle $E_i = Y_i - \beta_0 - \beta_1 X_i$ que nous estimons par $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$, nous utiliserons

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-2}.$$

Nous en déduisons donc

$$T_0 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_E}{S_{XX}}}} \sim t_{n-2}.$$

Intervalles de confiance et de prévision

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	β_1	β_0
Variance	$\frac{\sigma^2}{S_{XX}}$	$\frac{\sigma^2}{n} + \frac{\sigma^2 \bar{X}^2}{S_{XX}}$
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 :

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

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 $\beta_0 + \beta_1 x_0 + E_0$

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

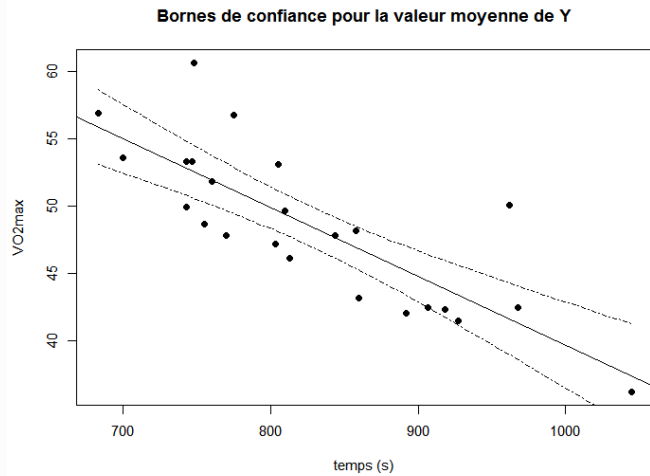
Exemple 2 (i)

- (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 min 25 s;

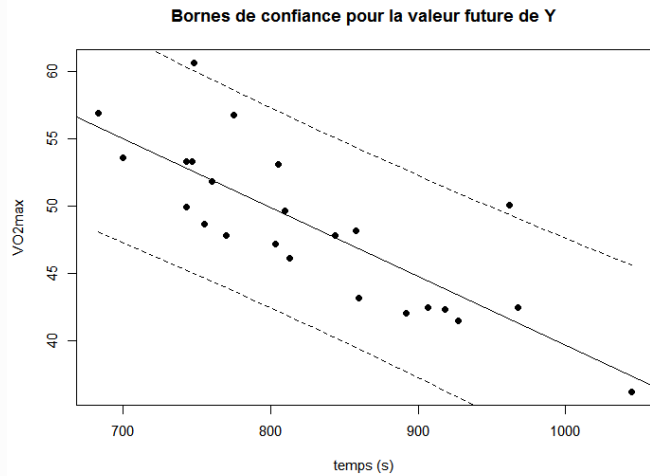
Exemple 2 (ii)

- (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 2 (i) (suite)



Exemple 2 (ii) (suite)



Qualité de l'ajustement

Rappel de la corrélation

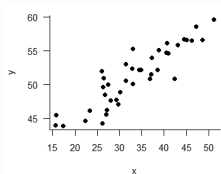
On pourrait montrer que $\beta_1 = 0 \Leftrightarrow \rho = 0$,
alors dans une relation linéaire, il n'y a corrélation que si la pente est différente
de zéro.

Estimation de ρ

On a $\rho = \beta_1 \frac{\sigma_X}{\sigma_Y}$, alors il y a un lien entre les estimations de β_1 et de ρ .

ρ est estimé par le coefficient de corrélation échantillonnal r
(lui aussi compris entre -1 et 1) :

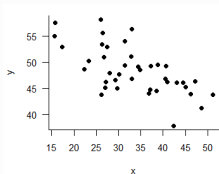
$$\hat{\rho} = r = \hat{\beta}_1 \frac{s_X}{s_Y} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}.$$



$\rho_{\text{simulé}} = 0,9$

$r = 0,90$

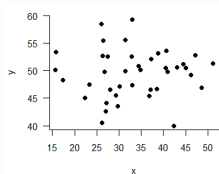
$R^2 = 0,809$



$-0,6$

$-0,640$

$0,409$



$0,1$

$0,041$

$0,0017$

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 3 sur le VO_2max

Coefficient de détermination :

Coefficient de corrélation échantillonnal :

Validation du modèle : Analyse des résidus

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

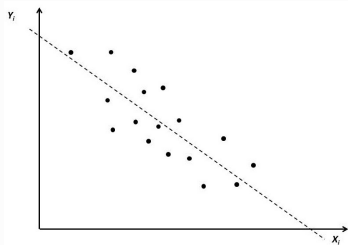
$$e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) = Y_i - \hat{Y}_i.$$

(P1) Linéarité : Validation graphique

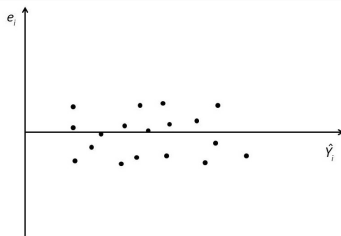
- Un graphique de dispersion des paires (x_i, y_i) : devrait montrer un nuage de points distribués de façon symétrique autour d'une ligne droite.
- Un graphique des paires $(\hat{Y}_i, e_i)$: devrait montrer un nuage de points distribués symétriquement autour de l'axe des abscisses.

Exemples de nuages de points

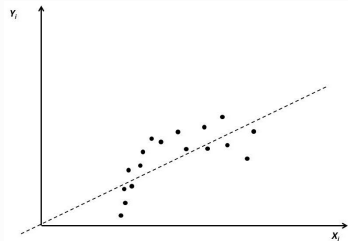
Linéarité raisonnable



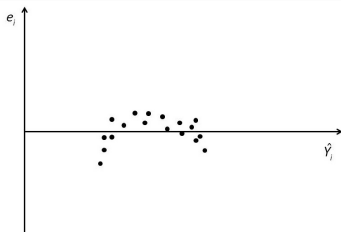
Linéarité raisonnable



Linéarité déraisonnable



Linéarité déraisonnable

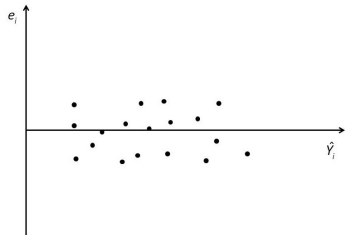


(P2) Homoscédasticité

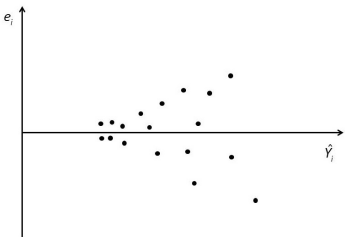
- Un graphique des e_i en fonction des $\hat{Y}_i$: devrait montrer que la variabilité des e_i demeure constante pour toute valeur de $\hat{Y}_i$.

Exemples de nuages de points

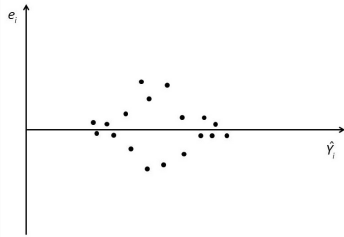
Homoscédasticité raisonnable



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 :

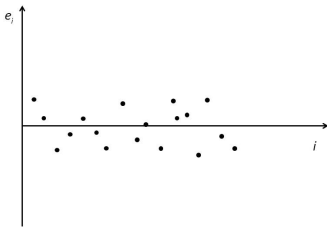
- les observations 1, 2, 3, 4 sont prises par l'employé A qui a tendance à surestimer Y;
- les observations 5, 6, 7 et 8 sont prises par l'employé B qui a tendance à sous-estimer Y;
- Problème : les observations prises par un même employé seront corrélées.

Données chronologiques

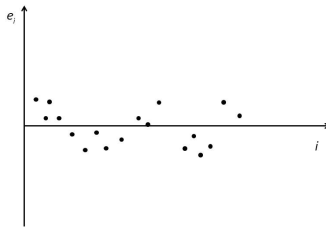
Il peut aussi y avoir de la corrélation lorsque les observations sont prises de façon trop rapprochées dans le temps (ou l'espace).

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



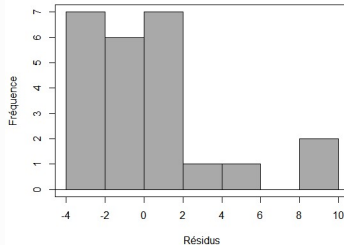
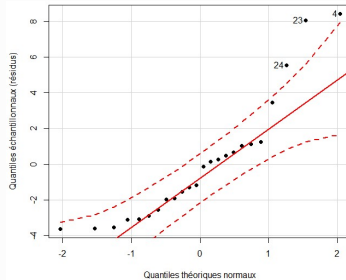
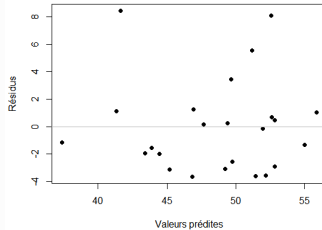
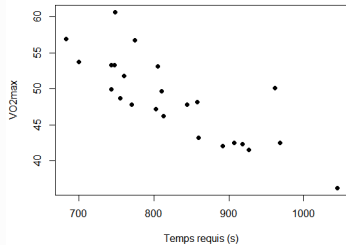
Effet de grappe



(P4) Normalité

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

Exemple 4 sur l'absorption d'oxygène



Réponses des exemples

Réponses des exemples

- **Exemple 1 (estimation des paramètres) :** $\hat{\beta}_1 = S_{XY}/S_{XX} \approx -0,0510$; $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} \approx 90,72$; **droite estimée :** $\hat{Y} = 90,72 - 0,0510 X$.
- **Exemple 2 (i) :** à $x_0 = 925$ s (15 min 25 s), **prévision ponctuelle** $\hat{Y}_0 \approx 43,52$; **IC à 95 % pour** $E(Y|X = x_0) : (41,33 ; 45,71)$. **(ii) :** même **prévision ponctuelle** $\hat{Y}_0 \approx 43,52$; **intervalle de prévision à 95 % pour une observation individuelle :** $(35,94 ; 51,10)$.
- **Exemple 3 :** $R^2 \approx 0,656$ (65,6 % de la variabilité de Y expliquée par le modèle); $r \approx -0,810$.
- **Exemple 4 :** les graphiques suggèrent que l'homoscédasticité (P2) semble raisonnable, mais la normalité (P4) est douteuse (résidus asymétriques, quelques valeurs extrêmes positives).

Réponses des quiz

Réponses des quiz

- **Quiz (interprétation du modèle) : différence prédite entre Jacques (10 s plus lent) et Marc : $10\hat{\beta}_1 \approx -0,51$ (le VO_{2max} prédit de Jacques est environ 0,51 unité plus bas); Faux — $X = 0$ est loin à l'extérieur du domaine observé (683 à 1045 s), il s'agit d'une extrapolation dangereuse (même si numériquement $\hat{\beta}_0 \approx 90,7$).**

Bibliographie

Ce document est rédigé avec Beamer ainsi qu'une classe .cls fournie par Jérôme Soucy. Les diapositives sont adaptées des diapositives créées par Thierry Duchesne et Emmanuelle Reny-Nolin pour le cours STT-1900. Les graphiques proviennent de ces mêmes diapositives.

Pour signaler une erreur dans ce document, veuillez écrire à l'adresse : julien.miron@mat.ulaval.ca.