

Chapitre 6b – Régression linéaire simple (suite)

STT-1920 Méthodes statistiques

Département de mathématiques et de statistique



UNIVERSITÉ
LAVAL

Plan

Prédictions

Qualité de l'ajustement

Vérification des postulats

Prédictions

Intervalle de confiance pour la moyenne de Y en x^* : exemple

On voudrait construire un intervalle de confiance à 95 % pour la **production moyenne** attendue de **tous** les plants ayant reçu 6,5 heures d'ensoleillement par jour.

Rappel des calculs du chapitre 6a : $\bar{x} = 5,5$, $\bar{y} = 3,744$, $S_{XX} = 175$, $S_{XY} = 75,470$, d'où

$$b_1 = \frac{75,470}{175} = 0,4313, \quad b_0 = 3,744 - 0,4313 \times 5,5 = 1,3724, \quad \hat{\sigma}^2 = \text{MSE} = 0,3228.$$

La droite de régression estimée est donc $\hat{y} = 1,3724 + 0,4313 x$.

IC pour la moyenne de Y en x^*

La valeur moyenne de Y lorsque $X = x^*$ est le point sur la vraie droite en x^* : $\beta_0 + \beta_1 x^*$, estimé par $b_0 + b_1 x^*$. On peut montrer que

$$b_0 + b_1 x^* \sim N \left(\beta_0 + \beta_1 x^*, \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}} \right] \right).$$

Intervalle de confiance de niveau $1 - \alpha$ pour $\beta_0 + \beta_1 x^*$

On estime σ^2 par le MSE et on obtient

$$(b_0 + b_1 x^*) \pm t_{n-2; \alpha/2} \sqrt{\text{MSE} \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}} \right]}.$$

Exemple : IC pour la production moyenne à 6,5 h

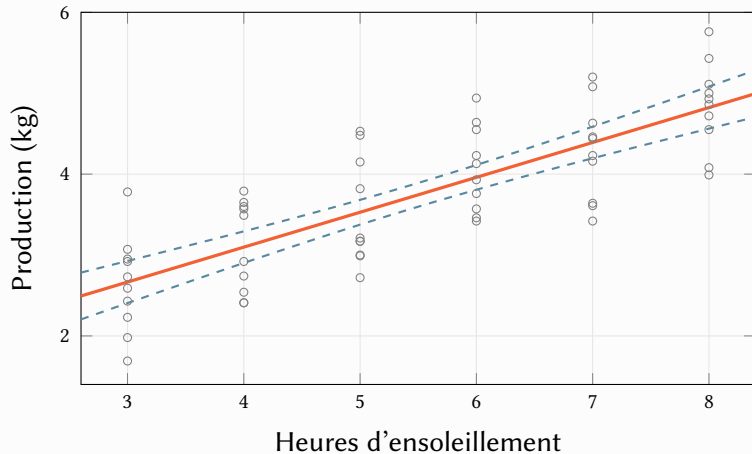
On a $\bar{x} = 5,5$ h et $s_x^2 = 2,966$ (donc $S_{XX} = 59 \times 2,966 = 175$).

Précision optimale?

Question : pour quelle valeur de x^* obtient-on la meilleure précision?

$$(b_0 + b_1 x^*) \pm t_{n-2; \alpha/2} \sqrt{\text{MSE} \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}} \right]}$$

Bornes de confiance à 95 % pour la valeur moyenne de Y



Intervalle de prédiction pour une observation future

Le producteur veut vendre des semis de cette variété de tomates, et indiquer des productions attendues dans des conditions similaires (terreau, arrosage, engrais, etc.) pour diverses durées d'ensoleillement quotidien.

- Selon notre modèle, le plant recevant 6,5 heures d'ensoleillement produira combien de kilos de tomates ?
- Quelle est l'erreur-type associée à cette estimation ? Sera-t-elle plus grande ou plus petite que celle de la moyenne attendue avec 6,5 h de soleil ?

Intervalle de prédiction

Ici, on n'estime plus une constante (la moyenne de la population au point x^*), mais une **variable aléatoire** (une nouvelle observation). Sa variance sera donc plus grande que celle de la moyenne.

Intervalle de prédiction de niveau $1 - \alpha$ pour une valeur future de Y

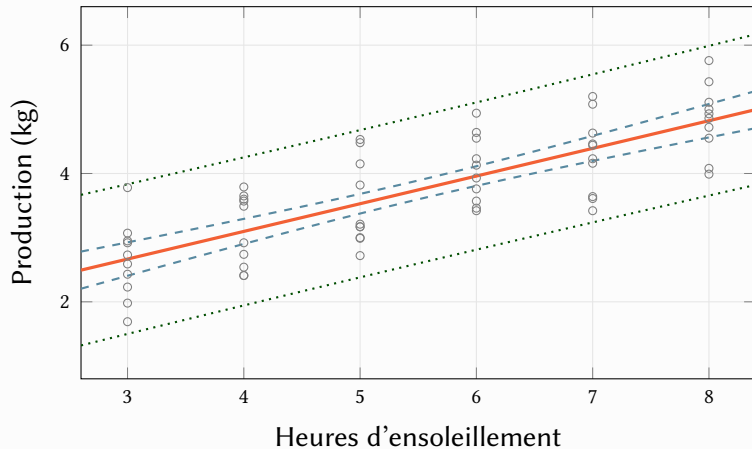
En estimant σ^2 par le MSE :

$$(b_0 + b_1 x^*) \pm t_{n-2; \alpha/2} \sqrt{\text{MSE} \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}} \right]}.$$

Exemple

Construire un intervalle de prédiction à 95 % pour la production de tomates du plant ayant reçu 6,5 heures d'ensoleillement.

Bornes de prédiction à 95 % pour une valeur future de Y



Tirets : bornes de confiance pour la moyenne; pointillés : bornes de prédiction.

En R : `predict()`

```
> nouveau <- data.frame(hrs_soleil=6.5)
> predict(RegModel.1, newdata=nouveau, interval="confidence")
      fit      lwr      upr
1 4.17559 4.005455 4.345725
> predict(RegModel.1, newdata=nouveau, interval="prediction")
      fit      lwr      upr
1 4.17559 3.025689 5.325492
```

- Même estimation ponctuelle (`fit`) dans les deux cas : 4,18 kg.
- L'intervalle de prédiction est environ **7 fois plus large** : la variabilité d'un plant individuel (σ^2) domine.

IC ou intervalle de prédiction ?

	IC pour la moyenne	Intervalle de prédiction
Cible	$E(Y) = \beta_0 + \beta_1 x^*$ (une constante)	une nouvelle observation Y (aléatoire)
Question type	« En moyenne, les plants à 6,5 h... »	« Ce plant-ci, à 6,5 h... »
Sous la racine	$\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}$	$1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}$
Si $n \rightarrow \infty$	largeur $\rightarrow 0$	largeur $\rightarrow 2 \times 1,96 \sigma$

- Les deux sont plus étroits en $x^* = \bar{x}$ et s'élargissent en s'éloignant de $\bar{x}$.
- **Extrapolation dangereuse** (chapitre 6a) : ne pas prédire hors du domaine observé (ici, de 3 à 8 h).

QUIZ!

IC pour la moyenne ou intervalle de prédiction ?

- Un agronome veut estimer la production moyenne de l'ensemble des plants de cette variété cultivés avec 7 h de soleil par jour.
- Une jardinière vient d'acheter un plant qu'elle placera 7 h par jour au soleil; elle veut savoir combien il produira.
- Vrai ou faux : avec un très grand nombre de plants ($n \rightarrow \infty$), la largeur de l'intervalle de prédiction tend vers 0.
- Vrai ou faux : à 7 h de soleil, l'IC pour la moyenne est plus étroit qu'à 6,5 h.

Qualité de l'ajustement

Coefficient de détermination R^2

On sait que $SSR + SSE = SST$.

Coefficient de détermination

R^2 = proportion de la variabilité totale de Y expliquée par le modèle (par la variable X)

$$R^2 =$$

Propriétés :

- $0 \leq R^2 \leq 1$;
- $R^2 = 0$ si...
- $R^2 = 1$ si...

Exemple

	Sum Sq	Df	F value	Pr(>F)	
hrs_soleil	32.547	1	100.83	2.661e-14	***
Residuals	18.721	58			

Calculez et interprétez le R^2 .

Coefficient de corrélation r

Le coefficient de corrélation mesure le **degré d'association linéaire** entre X et Y.

Coefficient de corrélation

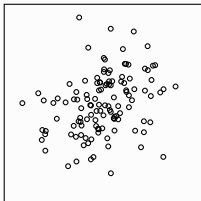
$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = (\text{signe de } b_1) \sqrt{R^2}$$

Propriétés de r

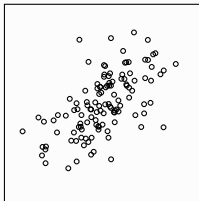
- $-1 \leq r \leq 1$;
- $r > 0$ si
- $r < 0$ si
- $r = 1$ ou -1 si tous les points sont alignés;
- $r = 0$ si la pente de la droite estimée est nulle;
- plus r est proche de 1 ou de -1 , plus la relation linéaire est forte;
- r ne mesure **pas** la causalité de la relation.

Quelques nuages de points et leur corrélation

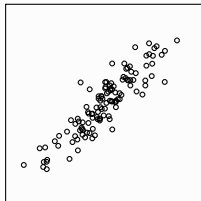
$r = 0.3$



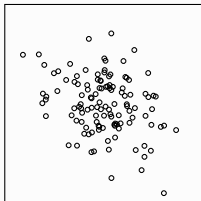
$r = 0.6$



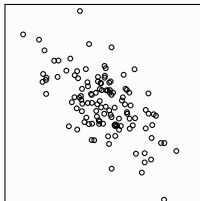
$r = 0.9$



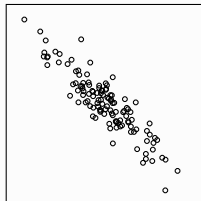
$r = -0.3$



$r = -0.6$



$r = -0.9$



Exemple

Quel est le coefficient de corrélation entre la production de tomates et le temps d'ensoleillement ?

```
> cor(tomates[,c("hrs_soleil", "production")])
```

	hrs_soleil	production
hrs_soleil	1.0000000	0.7967679
production	0.7967679	1.0000000

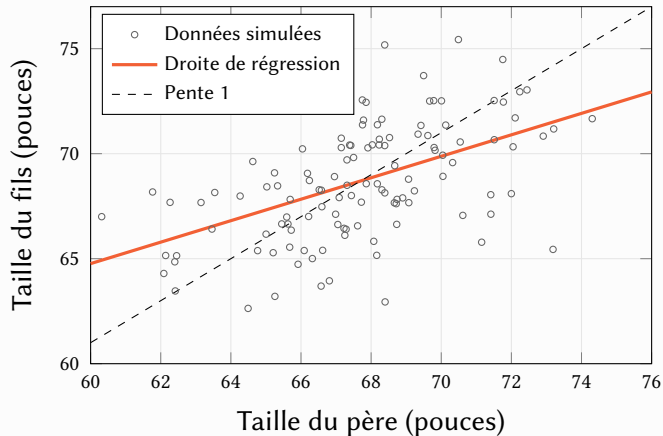
QUIZ!

- Pour les hirondelles (chapitre 6a), on obtient $R^2 = 0,792$ et une pente positive. Que vaut r ? Interprétez R^2 .
- Vrai ou faux : $r = -0,9$ indique un lien linéaire plus faible que $r = 0,5$.
- Vrai ou faux : si $r = 0$, il n'y a aucune relation entre X et Y.
- Vrai ou faux : un R^2 élevé garantit que les postulats du modèle sont respectés.

Pourquoi « régression » ?

- Le terme vient de Francis Galton, *Regression towards mediocrity in hereditary stature* (Journal of the Anthropological Institute, 1886).
- Galton a comparé la taille d'enfants à la taille moyenne de leurs parents : les enfants de parents très grands sont en moyenne grands, mais **moins** que leurs parents ; ceux de parents très petits sont petits, mais moins que leurs parents.
- Les tailles « régressent » vers la moyenne : la pente de la droite est inférieure à 1. C'est la **régression vers la moyenne**.
- Karl Pearson (1903) a observé le même phénomène sur plus de 1000 paires père-fils.

Régression vers la moyenne



Données simulées selon les caractéristiques des données père-fils de Pearson (1903) : corrélation d'environ 0,5.

Vérification des postulats

Les postulats

Tests et intervalles reposent sur les postulats suivants, **en ordre d'importance** :

- (P1) **Linéarité** : la moyenne de Y est une fonction linéaire de X;
- (P2) **Homoscédasticité** : la variance autour de la droite est la même pour toutes les valeurs de X;
- (P3) **Indépendance** des observations;
- (P4) **Normalité** des observations autour de la droite.

Outils de vérification

- (P1) nuage (x_i, y_i) ; résidus en fonction des valeurs prédites $\hat{y}_i$
- (P2) résidus en fonction des valeurs prédites $\hat{y}_i$
- (P3) mode de collecte des données; résidus selon l'ordre de collecte
- (P4) histogramme, diagramme quantile-quantile, test de Shapiro-Wilk

Analyse des résidus

Rappel du modèle : $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$.

Puisque $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$, il suit que $\varepsilon_i \sim N(0, \sigma^2)$.

Résidus

Les erreurs échantillonales ε_i sont estimées par les **résidus** du modèle :

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i).$$

Au lieu d'analyser directement les observations, nous analyserons les résidus.

Calcul des résidus

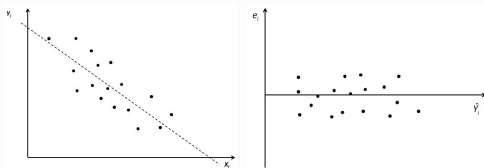
Dans notre exemple, $\hat{y} = 1,372 + 0,431 x$. Voici quelques calculs de résidus :

Donnée i	1	12	24	51
Heures de soleil x_i	3	4	5	8
Production observée y_i	2,92	2,92	3,21	4,08
Production prédite $b_0 + b_1 x_i$	2,67	3,10	3,53	
Résidu $\hat{\varepsilon}_i$	0,25	-0,18	-0,32	

(P1) Linéarité

Le nuage (x_i, y_i) doit être réparti symétriquement autour d'une droite, et le graphique des résidus e_i en fonction de $\hat{y}_i$, autour de 0, **sans tendance**.

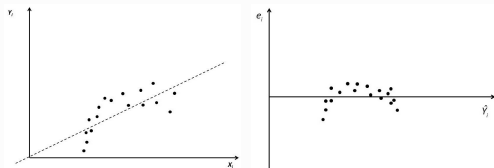
Linéarité raisonnable



y_i selon x_i

e_i selon $\hat{y}_i$

Linéarité déraisonnable



y_i selon x_i

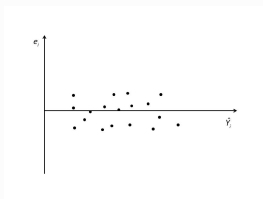
e_i selon $\hat{y}_i$

À droite : la tendance est courbe ; les résidus forment un « arc » (négatifs aux extrémités, positifs au centre).

(P2) Homoscédasticité

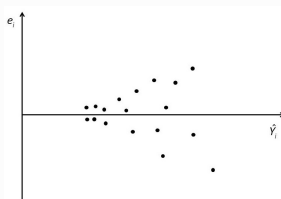
Dans le graphique des résidus e_i en fonction des valeurs prédites $\hat{y}_i$, la **dispersion verticale** doit rester à peu près **constante**.

Raisonné



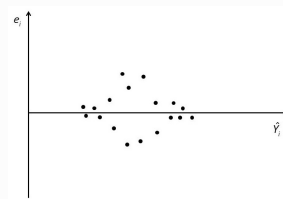
bande horizontale

Déraisonné



« entonnoir » : la variance
augmente avec $\hat{y}$

Déraisonné

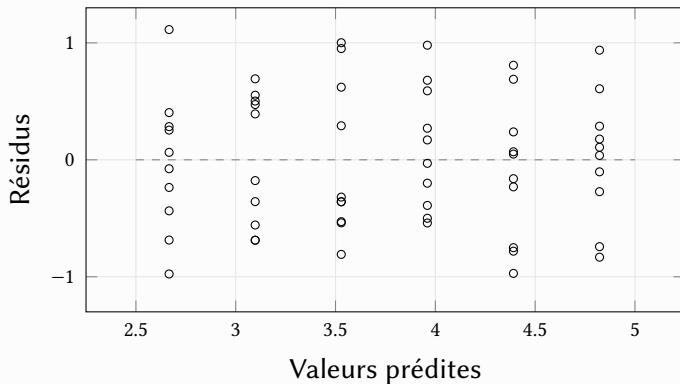


variance plus grande
au centre

En biologie, l'entonnoir est fréquent : les grandes valeurs (concentrations, biomasses, dénombrements) varient souvent davantage.

Exemple 1 : résidus en fonction des valeurs prédites

Diagramme de dispersion des résidus en fonction des valeurs prédites :



Les postulats (P1) et (P2) semblent-ils respectés ?

(P3) Indépendance : difficile à vérifier!

C'est surtout la **collecte des données** qui en est garante : aucune autre variable que X ne doit lier certaines observations entre elles.

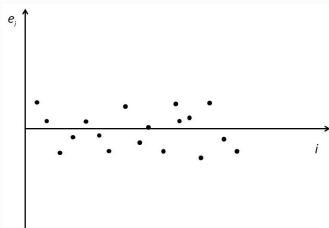
Exemples de collectes problématiques :

- plusieurs mesures prises sur le **même individu** (même souris, même plant) comptées comme des observations distinctes;
- des plants cultivés dans le même pot ou sur la même table de la serre, qui partagent lumière, arrosage et température;
- les mesures 1 à 30 prises par une technicienne qui a tendance à surestimer Y, les mesures 31 à 60 par un technicien qui le sous-estime;
- des **données chronologiques** ou spatiales : mesures quotidiennes de la concentration d'un polluant dans une rivière, parcelles voisines d'un champ... les observations rapprochées se ressemblent.

(P3) Données chronologiques

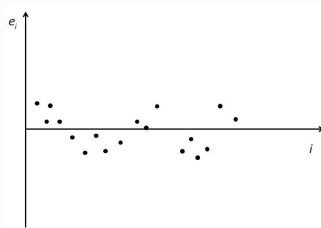
Si les indices $i = 1, \dots, n$ correspondent à l'ordre de collecte, le graphique des résidus e_i en fonction de i peut révéler une corrélation entre résidus rapprochés (« effet de grappe »).

Pas d'effet chronologique



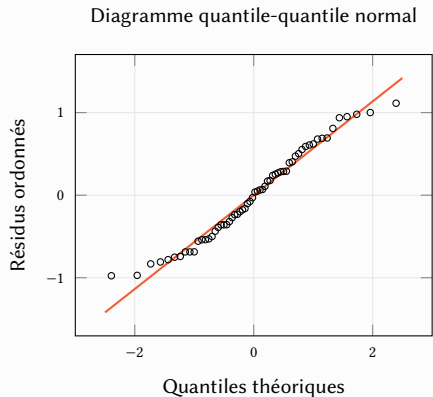
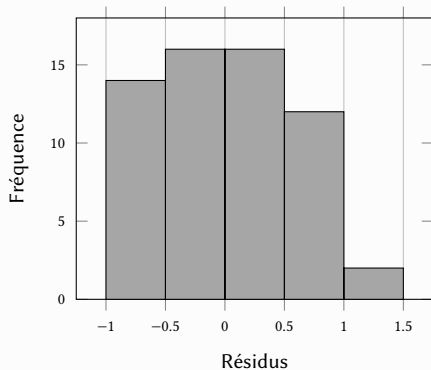
signes + et - bien mélangés

Effet de grappe



résidus de même signe regroupés

(P4) Normalité : histogramme et diagramme quantile-quantile



(P4) Test de normalité sur les résidus

Rappel (chapitre 5a) : le test de Shapiro-Wilk teste H_0 : les erreurs suivent une loi normale, contre H_1 : elles ne suivent pas une loi normale.

```
> shapiro.test(residuals(RegModel.1))
```

Shapiro-Wilk normality test

```
data:  residuals(RegModel.1)  
W = 0.96941, p-value = 0.1365
```

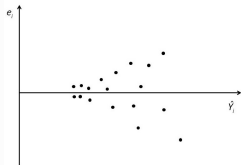
Conclusion?

Rappel : avec un grand n , de petits écarts à la normalité deviennent significatifs; avec un petit n , le test a peu de puissance. Les graphiques restent essentiels.

QUIZ!

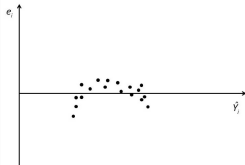
Pour chaque graphique de résidus, quel postulat semble violé (s'il y en a un)?

A



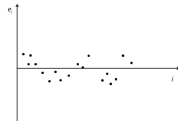
e_i selon $\hat{y}_i$

B



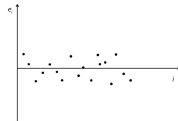
e_i selon $\hat{y}_i$

C



e_i selon l'ordre i

D



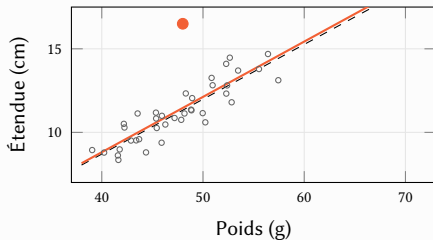
e_i selon l'ordre i

- Pour C et D, les données ont été recueillies une par jour, dans l'ordre i . Pourquoi l'ordre de collecte importe-t-il?
- Vrai ou faux : si le nuage (x_i, y_i) est bien linéaire, il est inutile de regarder les résidus.

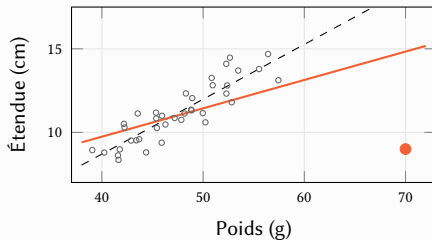
Valeurs aberrantes et influentes

Un oiseau fictif (en orange) ajouté aux hirondelles du chapitre 6a. Tirets : droite sans cet oiseau ; trait plein : droite avec cet oiseau.

Valeur **aberrante**: (48; 16,5)



Valeur **influyente**: (70; 9)



- **Aberrante** : très grand résidu ; la pente change peu ($0,328 \rightarrow 0,331$), mais R^2 chute ($0,79 \rightarrow 0,65$).
- **Influente** : x loin des autres ; elle « tire » la droite vers elle (pente $0,328 \rightarrow 0,170$; R^2 : $0,79 \rightarrow 0,33$).

Valeurs aberrantes et influentes : que faire ?

- Toujours **regarder les graphiques** : un seul point peut changer toute la conclusion.
- Vérifier la donnée : erreur de saisie (virgule mal placée, g au lieu de kg)? erreur de mesure? individu d'une autre population (autre espèce, animal malade)?
- Si c'est une erreur, on la corrige (ou on retire la donnée) en le mentionnant.
- Sinon, on **ne retire pas** une donnée simplement parce qu'elle dérange! On peut présenter l'analyse **avec et sans** la donnée et discuter de la différence.

Que faire si les postulats ne sont pas respectés ?

On tente souvent de **transformer** Y et/ou X, puis on refait la régression (et la vérification des postulats!) sur les données transformées.

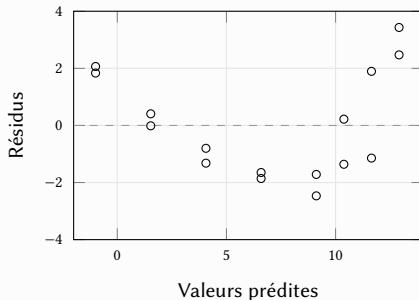
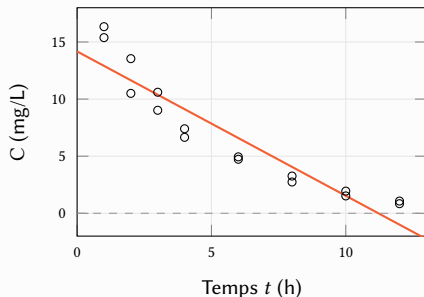
- **Non-linéarité (P1)** : s'inspirer de la forme du nuage, par exemple

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad Y_i = \beta_0 + \beta_1 \ln(X_i) + \varepsilon_i, \quad Y_i = \beta_0 + \beta_1 X_i^2 + \varepsilon_i.$$

- **Hétéroscédasticité (P2) ou non-normalité (P4)** : une transformation de Y ($\ln Y$, $\sqrt{Y}$) stabilise souvent la variance, surtout quand la dispersion augmente avec la moyenne (entonnoir).
- **Dépendance (P3)** : aucune transformation ne règle le problème; il faut revoir la collecte ou utiliser des modèles qui tiennent compte de la dépendance (hors du cours).
- Autres pistes (hors du cours) : régression polynomiale, régression multiple, modèles non linéaires, méthodes non paramétriques.

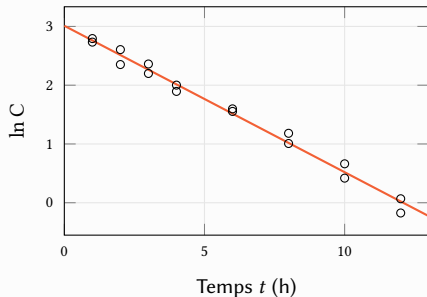
Exemple 3 : concentration d'un médicament

On mesure la concentration C (mg/L) d'un médicament dans le sang de 16 patients (un seul prélèvement chacun), $t = 1$ à 12 h après l'injection (données simulées).



Droite : $\hat{C} = 14,17 - 1,264 t$, $R^2 = 0,875$. Les postulats sont-ils respectés ? Que prédit ce modèle à $t = 12$ h ?

Exemple 3 (suite) : transformation logarithmique



- Droite estimée :
 $\widehat{\ln C} = 3,011 - 0,249 t$, $R^2 = 0,988$;
résidus sans tendance.
- Chaque heure, $\ln C$ diminue de 0,249 : C est **multipliée** par $e^{-0,249} = 0,78$, i.e. elle **diminue d'environ 22 % par heure**.
- Prédiction à 12 h :
 $e^{3,011 - 0,249 \times 12} \approx 1,0$ mg/L (toujours positive).

Les tests et IC portent alors sur le modèle transformé ($\ln C$); on revient à l'échelle d'origine avec e^x .

Réponses aux exemples (1)

- **IC pour la moyenne à 6,5 h** : $\hat{y} = 1,3724 + 0,4313(6,5) = 4,18$; $4,18 \pm 2,00\sqrt{0,3228(1/60 + 1/175)} = 4,18 \pm 0,17$, soit $[4,01; 4,35]$ kg.
- **Précision optimale** en $x^* = \bar{x} = 5,5$ h (le terme $(x^* - \bar{x})^2$ s'annule).
- **Prédiction** : même estimation ponctuelle (4,18 kg), mais erreur-type plus grande; IP à 95 % : $4,18 \pm 2,00\sqrt{0,3228(1 + 1/60 + 1/175)} = 4,18 \pm 1,15$, soit $[3,03; 5,33]$ kg.
- **Quiz (IC ou IP)** : (1) IC pour la moyenne. (2) Intervalle de prédiction. (3) Faux : l'IP tend vers $\pm 1,96\sigma$ (largeur $\approx 2 \times 1,96 \times 0,568 = 2,2$ kg); c'est l'IC qui tend vers 0. (4) Faux : 7 h est plus loin de $\bar{x} = 5,5$ h que 6,5 h, donc l'IC est plus large.
- $R^2 = SSR/SST = 32,547/51,268 = 0,635$: 63,5 % de la variabilité de la production est expliquée par l'ensoleillement. $R^2 = 0$ si $b_1 = 0$ ($SSR = 0$); $R^2 = 1$ si tous les points sont sur la droite ($SSE = 0$).
- r : $r > 0$ si $b_1 > 0$, $r < 0$ si $b_1 < 0$. Exemple : $r = +\sqrt{0,635} = 0,797$.
- **Quiz (R^2 et r)** : (1) $r = +\sqrt{0,792} = 0,89$; 79 % de la variabilité de l'étendue des ailes est expliquée par le poids. (2) Faux : $|-0,9| > 0,5$, le lien est plus fort (mais négatif). (3) Faux : r ne mesure que le lien *linéaire*; une relation en U peut donner $r \approx 0$. (4) Faux : il faut examiner les résidus (voir l'exemple 3 : $R^2 = 0,875$, mais relation courbe).

Réponses aux exemples (2)

- **Résidus** : donnée 51 : $\hat{y} = 4,82$, $\hat{\varepsilon} = -0,74$.
- **Exemple 1, résidus selon $\hat{y}_i$** : bande horizontale autour de 0, sans tendance ni entonnoir : (P1) et (P2) semblent respectés.
- **Normalité** : histogramme et QQ-plot compatibles avec la normalité ; Shapiro-Wilk : valeur-p = 0,137 > 0,05, on ne rejette pas la normalité des erreurs.
- **Quiz (résidus)** : A : (P2), la dispersion augmente avec $\hat{y}$ (entonnoir). B : (P1), les résidus forment un arc (relation courbe). C : (P3), les résidus de même signe sont regroupés (effet de grappe). D : aucun problème apparent. L'ordre importe car des mesures prises à des jours rapprochés peuvent se ressembler (dépendance). Faux : les résidus révèlent mieux une légère courbure, l'hétéroscédasticité ou la non-normalité.
- **Exemple 3** : (P1) n'est pas respecté : résidus en U (positifs aux extrémités, négatifs au centre); (P2) douteux (dispersion plus grande pour les fortes concentrations). À 12 h, $\hat{C} = 14,172 - 1,2639 \times 12 = -1,0$ mg/L : une concentration négative, impossible!