

Chapitre 6a – Régression linéaire simple

STT-1920 Méthodes statistiques

Département de mathématiques et de statistique



UNIVERSITÉ
LAVAL

Plan

Introduction

Modèle de régression linéaire simple

Estimation des paramètres

Test t sur la pente

Test de Fisher pour la régression

Introduction

Objectifs de la régression linéaire simple

- Vérifier si une **relation linéaire** existe entre les variables aléatoires X (variable **explicative**) et Y (variable **réponse**).
- Mesurer la **force** du lien linéaire qui unit les variables.
- Trouver la **meilleure droite** exprimant la relation entre les deux variables à partir de n paires d'observations indépendantes (X_i, Y_i) . Évaluer la précision des estimations de la pente et de l'ordonnée à l'origine.
- **Prédire** la valeur de Y pour un X donné, et évaluer la précision de ces prédictions.

Trois grands usages de la régression

- **Décrire** : quelle est la forme et la force du lien entre deux variables ?
Ex. : comment l'envergure d'une hirondelle varie-t-elle avec son poids ?
- **Prédire** : quelle valeur de Y peut-on attendre pour une valeur de X donnée ?
Ex. : quelle production attendre d'un plant de tomates qui reçoit 6,5 h de soleil par jour ?
- **Contrôler** : quelle valeur de X faut-il fixer pour obtenir, en moyenne, la valeur de Y désirée ?
Ex. : combien d'heures d'ensoleillement faut-il fournir pour qu'un plant produise en moyenne 4 kg de tomates ?

Exemples d'application

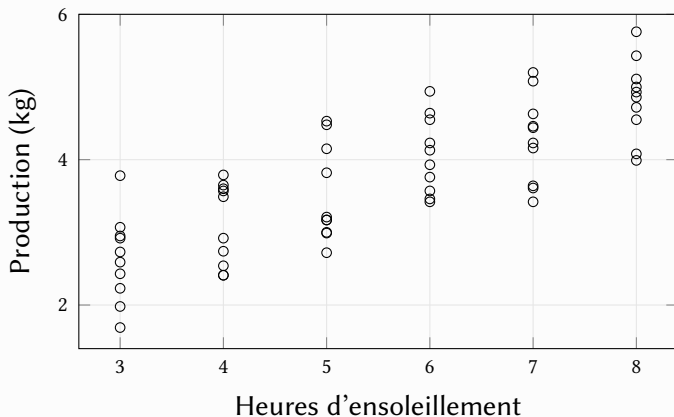
- Un sociologue veut savoir s'il existe un lien linéaire entre le nombre d'années d'études et le salaire à l'embauche.
- Un médecin veut étudier la relation entre l'IMC des individus et le taux de cholestérol dans leur sang.
- Un ingénieur veut savoir si le taux d'humidité dans le papier est directement proportionnel à la température du séchoir.
- Un économiste veut savoir si le lien entre l'offre d'un produit et son prix est linéaire.
- Une écologiste veut savoir si la biomasse d'algues d'un lac augmente linéairement avec la concentration de phosphore dans l'eau.
- Une pharmacologue veut relier la dose d'un antihypertenseur à la baisse de pression artérielle observée chez les patients.

Exemple 1 : ensoleillement et production de tomates

On veut étudier l'effet du nombre d'heures d'ensoleillement quotidien sur la production (kg) de tomates. Dans une serre, on contrôle la quantité de soleil reçue par 60 plants : 10 plants pour chacune des durées 3, 4, 5, 6, 7 et 8 heures par jour. Quatre mois plus tard, on note la production de chaque plant.

Soleil	Production (kg)									
3 h	2,92	2,59	2,73	2,43	3,78	2,23	3,07	1,69	1,98	2,95
4 h	2,54	2,92	3,65	3,57	3,60	2,41	3,79	2,74	2,41	3,49
5 h	3,17	2,72	4,53	3,21	3,00	4,15	2,99	4,48	3,17	3,82
6 h	4,13	4,64	4,55	3,42	4,23	3,76	3,57	4,94	3,93	3,46
7 h	3,42	5,20	4,63	4,16	3,64	3,61	4,44	4,46	5,08	4,23
8 h	4,08	5,43	5,76	3,99	4,55	4,93	5,11	5,00	4,86	4,72

Exemple 1 : les 60 plants de tomates



La relation entre les deux variables semble **linéaire**.

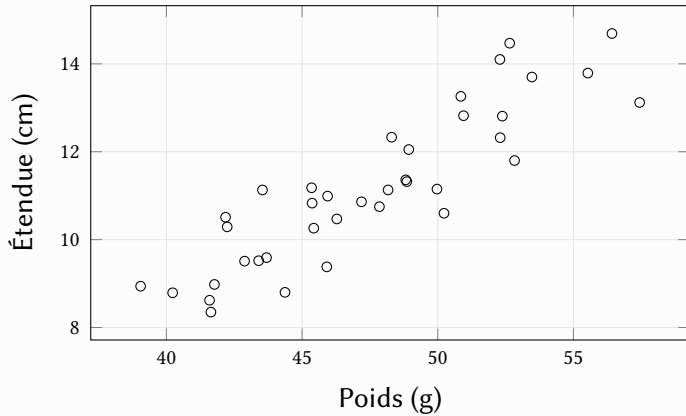
Exemple 2 : poids et envergure des hirondelles

On cherche à démontrer qu'il existe un lien linéaire entre le **poids** d'une hirondelle (X, en g) et l'**étendue de ses ailes** ouvertes (Y, en cm). Pour ce faire, on mesure ces deux quantités sur 50 hirondelles choisies au hasard dans la population d'intérêt.

Oiseau	Poids	Étendue
1	52,29	14,10
2	39,05	8,94
3	40,23	8,79
4	43,69	9,59
5	50,23	10,60
⋮	⋮	⋮

Oiseau	Poids	Étendue
26	48,93	12,05
27	48,17	11,13
28	55,53	13,79
29	50,85	13,26
30	42,88	9,51
⋮	⋮	⋮

Exemple 2 : les hirondelles



Attention : association n'est pas causalité!

- **Tomates** (**expérience contrôlée**) : on fixe l'ensoleillement. Si les plants sont répartis au hasard entre les durées, on peut attribuer à l'ensoleillement les différences de production.
- **Hirondelles** (**étude observationnelle**) : les oiseaux lourds ont de plus grandes ailes, mais engraisser une hirondelle n'allongera pas ses ailes! Les deux variables dépendent de la taille de l'oiseau.
- **Villes** : plus de pharmacies ↔ plus de cas de grippe, car les deux augmentent avec la population (**variable confondante**).

À retenir

Un lien, même très fort, **ne prouve pas** que X cause Y. Seule une expérience contrôlée (répartition aléatoire) permet de conclure à un effet causal.

Modèle de régression linéaire simple

Le modèle

On suppose que X et Y sont reliées par la droite $Y = \beta_0 + \beta_1 X$, et que les observations dévient un peu de ce modèle par une erreur aléatoire ε :

Modèle de régression linéaire simple

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n$$

- $Y_i = i^e$ mesure de la variable Y ;
- $X_i = i^e$ mesure de la variable X ;
- $\beta_0 =$
- $\beta_1 =$
- $\varepsilon_i =$ perturbation due au hasard ou à des variables autres que X (erreur du modèle, différence entre Y_i et la droite).

Les postulats

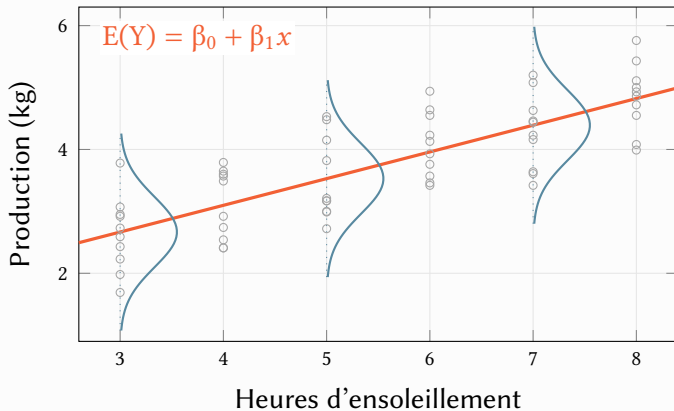
- (P1) **Linéarité** : les variables X et Y sont associées par un lien linéaire, i.e.
 $E(Y_i) = \beta_0 + \beta_1 X_i$.
- (P2) **Homoscédasticité** : la **variance** des Y_i autour de la droite (notée σ^2) est **constante** pour toutes les valeurs de X .
- (P3) **Indépendance** : les observations (X_i, Y_i) sont indépendantes les unes des autres.
- (P4) **Normalité** : les observations Y_i sont issues d'une loi normale pour chaque X_i fixé.

En résumé

$\varepsilon_1, \dots, \varepsilon_n$ sont indépendantes et de loi $N(0, \sigma^2)$.

Ces postulats seront vérifiés au chapitre 6b, à l'aide des résidus.

Illustration du modèle



Pour chaque x fixé, Y suit une loi normale centrée sur la droite, avec la même variance σ^2 (P2 et P4).

Estimation des paramètres

Estimation par les moindres carrés

Pour estimer β_0 et β_1 , on minimise la **somme des carrés des erreurs**, i.e. des écarts verticaux entre les observations Y_i et le point correspondant sur la droite

$\beta_0 + \beta_1 X_i$:

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - [\beta_0 + \beta_1 X_i])^2.$$

Il suffit donc de dériver cette expression successivement par rapport à β_0 et β_1 , d'égaliser les dérivées partielles à 0 et de résoudre le système obtenu pour trouver $\hat{\beta}_0 = b_0$ et $\hat{\beta}_1 = b_1$.

Estimateurs b_0 et b_1

Estimateurs des moindres carrés

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{XY}}{S_{XX}}, \quad b_0 = \bar{y} - b_1 \bar{x}.$$

La valeur de Y prédite par le modèle pour une valeur de X fixée est donc

$$\hat{y}_i = b_0 + b_1 x_i.$$

Notations

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

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

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

Inférence sur les paramètres : distribution de b_0 et b_1

b_0 et b_1 sont des **variables aléatoires** : si on change d'échantillon (X_i, Y_i) , leurs valeurs changent.

Théorème. On peut montrer que

$$E(b_0) = \beta_0, \quad E(b_1) = \beta_1,$$

$$V(b_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right), \quad V(b_1) = \frac{\sigma^2}{S_{XX}},$$

et b_0 et b_1 suivent chacun une loi normale.

Comme on connaît leur distribution, on peut construire des intervalles de confiance et des tests.

Remarques

-

-

-

-

En R : Statistiques → Ajustement de modèle

```
> RegModel.1 <- lm(production ~ hrs_soleil, data=tomates)
> summary(RegModel.1)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.97619	-0.45213	0.00378	0.42100	1.11381

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.37242	0.24733	5.549	7.46e-07 ***
hrs_soleil	0.43126	0.04295	10.042	2.66e-14 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5681 on 58 degrees of freedom

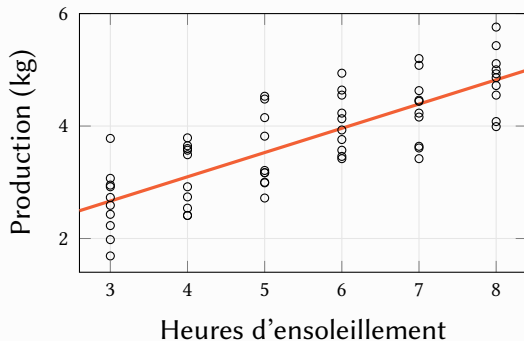
Multiple R-squared: 0.6348, Adjusted R-squared: 0.6285

F-statistic: 100.8 on 1 and 58 DF, p-value: 2.661e-14

Estimation des paramètres

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.37242	0.24733	5.549	7.46e-07	***
hrs_soleil	0.43126	0.04295	10.042	2.66e-14	***

Trouvez et interprétez les estimations des paramètres.



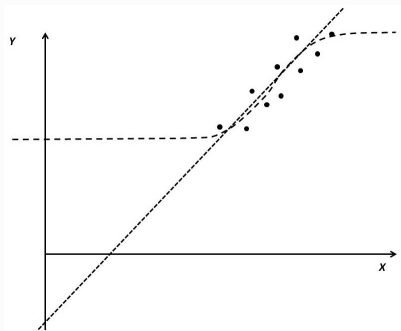
Interprétation des paramètres β_0 et β_1

Si le postulat de linéarité (P1) est vrai

- β_1 : si X augmente d'une unité, la valeur moyenne de Y augmente de β_1 unités.
 - $\beta_1 > 0$: lien positif (grands X associés aux grands Y); $\beta_1 < 0$: lien négatif;
 $\beta_1 = 0$: aucun lien linéaire.
 - β_0 : valeur moyenne de Y lorsque $X = 0$.
-
- **Tomates** : $b_1 = 0,431$ kg/h. Chaque heure supplémentaire d'ensoleillement quotidien augmente la production moyenne d'un plant d'environ 0,43 kg (431 g).
 - $b_0 = 1,37$ kg serait la production moyenne d'un plant sans aucun soleil... Or $x = 0$ est loin des durées observées (3 à 8 h) : b_0 n'a pas d'interprétation réaliste ici. Le modèle reste toutefois utile pour des x proches de ceux observés.

Extrapolation dangereuse!

Attention! On ignore la forme de la relation entre X et Y à l'extérieur de l'intervalle des valeurs de X observées.



- La vraie relation (courbe pointillée) peut être linéaire là où on a des données... et tout autre chose ailleurs.
- Tomates : le modèle prédit 1,37 kg avec 0 h de soleil et $1,372 + 0,431 \times 24 = 11,7$ kg avec 24 h par jour. Rien dans les données ne permet d'y croire!
- En biologie, les réponses **plafonnent** souvent (saturation, toxicité).

QUIZ!

Exemple 1 (tomates) : $\hat{y} = 1,372 + 0,431 x$.

- Le plant A reçoit 2 h de soleil de plus par jour que le plant B. Quelle différence de production moyenne le modèle prédit-il entre les deux ?
- Vrai ou faux : comme $b_1 = 0,43$ est proche de 0, le lien entre l'ensoleillement et la production est faible.
- Vrai ou faux : un plant cultivé à l'obscurité totale produit en moyenne 1,37 kg de tomates.
- Combien d'heures de soleil faut-il fournir pour qu'un plant produise en moyenne 4 kg ? (usage « contrôler »)

QUIZ!

Exemple 2 (hirondelles) : régression de l'étendue des ailes (cm) sur le poids (g).

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-4.42971	1.34034	-3.305	0.00216	**
poids	0.32824	0.02806	11.697	8.05e-14	***

- Écrivez l'équation de la droite estimée. Le lien est-il positif ou négatif?
- Interprétez b_1 dans le contexte.
- Interprétez b_0 . A-t-il un sens biologique?

Intervalles de confiance sur les paramètres

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.37242	0.24733	5.549	7.46e-07	***
hrs_soleil	0.43126	0.04295	10.042	2.66e-14	***

Intervalles de confiance à $(1 - \alpha)100\%$

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

où $\text{MSE} = \text{SSE}/(n - 2)$ estime σ^2 (voir la table d'ANOVA plus loin).

Calculez un intervalle de confiance à 95 % sur la pente et sur l'ordonnée à l'origine.

En R : Modèles → Intervalles de confiance

```
> Confint(RegModel.1, level=0.95)
              Estimate      2.5 %      97.5 %
(Intercept) 1.3724190 0.8773266 1.8675115
hrs_soleil   0.4312571 0.3452894 0.5172249
```

La relation linéaire entre les heures d'ensoleillement et la production est-elle significative, au seuil de 5 % ?

Interpréter l'intervalle de confiance sur la pente

Calcul à la main ($t_{58;0,025} = 2,0017$; l'erreur-type $\sqrt{\text{MSE}/S_{XX}} = 0,04295$ est la colonne Std. Error):

$$0,43126 \pm 2,0017 \times 0,04295 = 0,43126 \pm 0,08597 \Rightarrow [0,345; 0,517].$$

En mots

Avec 95 % de confiance, chaque heure supplémentaire d'ensoleillement quotidien fait augmenter la production **moyenne** d'un plant de **0,35 à 0,52 kg**.

- L'IC donne un ordre de grandeur de l'effet, pas seulement « il y a un effet ».
- Comme 0 n'appartient pas à l'IC, on conclut (au seuil de 5 %) que $\beta_1 \neq 0$: la relation linéaire est significative.
- Attention : l'IC porte sur β_1 (la pente de la population), pas sur la production d'un plant en particulier (voir chapitre 6b).

Test t sur la pente

Test de Student sur la pente

La question principale : **existe-t-il un lien linéaire** entre X et Y? On teste $H_0 : \beta_1 = 0$ (aucun lien linéaire) contre $H_1 : \beta_1 \neq 0$.

Statistique du test

Puisque $b_1 \sim N(\beta_1, \sigma^2/S_{XX})$, on a, sous H_0 ,

$$T = \frac{b_1 - 0}{\sqrt{MSE/S_{XX}}} \sim t_{n-2}.$$

Même logique qu'au chapitre 3 : (estimation – valeur sous H_0) / erreur-type.

H_1	Rejeter H_0 si...	H_1	Rejeter H_0 si...
$\beta_1 \neq 0$	$ t_{obs} > t_{n-2; \alpha/2}$	$\beta_1 > 0$	$t_{obs} > t_{n-2; \alpha}$
		$\beta_1 < 0$	$t_{obs} < -t_{n-2; \alpha}$

Exemple 1 (suite) : la pente est-elle nulle ?

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.37242	0.24733	5.549	7.46e-07	***
hrs_soleil	0.43126	0.04295	10.042	2.66e-14	***

- Statistique observée : $t_{obs} = \frac{0,43126}{0,04295} = 10,04$ (colonne `t value`).
- Valeur critique : $t_{58;0,025} = 2,00$. Comme $|t_{obs}| = 10,04 > 2,00$, on **rejette** H_0 .
- Valeur-p = $P(|T| > 10,04) = 2,66 \times 10^{-14}$ (colonne `Pr(>|t|)`) : même conclusion.
- **Conclusion** : au seuil de 5 %, il existe un lien linéaire significatif entre l'ensoleillement et la production de tomates; la production moyenne **augmente** avec l'ensoleillement.
- La ligne (`Intercept`) teste $H_0 : \beta_0 = 0$, ce qui est rarement d'intérêt (ici, $x = 0$ n'a même pas de sens).

QUIZ!

Exemple 2 (hirondelles), $n = 38$:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-4.42971	1.34034	-3.305	0.00216	**
poids	0.32824	0.02806	11.697	8.05e-14	***

- Donnez la statistique et la valeur-p du test de $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$. Que conclut-on au seuil de 5 % ?
- Calculez un IC à 95 % pour β_1 ($t_{36;0,025} = 2,028$) et interprétez-le en mots.
- Une ornithologue affirme que l'étendue des ailes augmente en moyenne de 0,5 cm par gramme. Les données la contredisent-elles ?

Test de Fisher pour la régression

Hypothèses à tester

Faire le test de la régression, c'est vérifier si la relation linéaire est **significative**.
Bref, on veut savoir si la pente est nulle ou non.

$H_0 :$

$H_1 :$

On sait déjà vérifier si 0 est dans l'intervalle de confiance, et on sait interpréter le test de Student sur la pente. En mettant cette statistique au carré, on obtient un **test de Fisher** basé sur la décomposition de la variabilité des Y_i .

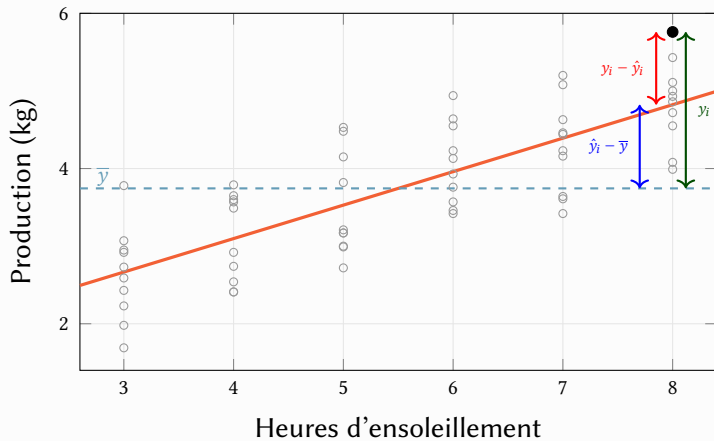
Décomposition de la variance

La variabilité totale dans les données (notée SST) peut être décomposée en deux sources :

- SSR : somme des carrés de la **régression** ;
variabilité entre les valeurs prédites par le modèle linéaire, i.e. la variabilité **expliquée par X** ;
- SSE : somme des carrés des **erreurs** ;
variabilité des données autour de la droite pour un x_i donné (source autre que X, que l'on associe à l'erreur échantillonnale non expliquée par le modèle).

Rappel (chapitre 5a) : c'est la même idée qu'en ANOVA, où $SST = SSR + SSE$.

Décomposition de la variance : illustration



Décomposition de la variance

Décomposition des sommes de carrés

$$SST = SSR + SSE$$

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

Table d'ANOVA

Source de variation	Somme de carrés	d.l.	Carrés moyens	F
Régression				
Erreur				
Total				

- σ^2 est la variance autour de la droite pour un X donné. On l'estime par le MSE.
- La proportion de variabilité des Y_i expliquée par X est $R^2 = \frac{SSR}{SST}$; c'est le carré du coefficient de corrélation.

Table d'ANOVA en R

```
> Anova(RegModel.1, type="II")
```

```
Anova Table (Type II tests)
```

```
Response: production
```

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

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conséquence

- Si H_0 est vraie,
- Si H_0 est fausse (et donc H_1 est vraie), la statistique F_{obs} aura-t-elle tendance à être plus petite ou plus grande que sous H_0 ?

Exemple

- Quel est le seuil observé du test effectué?
- Rejette-t-on H_0 au seuil de 5 %?

Lien entre le test F et le test t

En régression linéaire simple

$$F_{obs} = t_{obs}^2, \quad F_{1, n-2; \alpha} = \left(t_{n-2; \alpha/2} \right)^2,$$

et les deux tests (bilatéraux) donnent **la même valeur-p** : ils sont équivalents.

- Tomates : $t_{obs}^2 = 10,042^2 = 100,8 = F_{obs}$ et $t_{58; 0,025}^2 = 2,0017^2 = 4,007 = F_{1,58; 0,05}$.
- Les deux valeurs-p valent $2,66 \times 10^{-14}$.
- Le test t permet en plus des alternatives **unilatérales** (ex. $H_1 : \beta_1 > 0$), et on peut l'associer à un IC.
- Le test F se généralisera à la régression avec **plusieurs** variables explicatives (toutes les pentes testées à la fois).

QUIZ!

Vrai ou faux ?

- Quand $H_0 : \beta_1 = 0$ est vraie, les valeurs prédites $\hat{y}_i$ sont « proches » de la moyenne $\bar{y}$.
- Quand $H_0 : \beta_1 = 0$ est vraie, l'estimation b_1 vaut exactement 0.
- Dans une régression linéaire simple, on obtient $F_{obs} = 6,25$. Alors $|t_{obs}| = 2,5$ pour la pente.
- Si le test de la pente est très significatif (valeur-p $< 0,001$), alors X cause Y.

Réponses aux exemples (1)

- **Modèle** : β_0 = ordonnée à l'origine (valeur moyenne de Y quand $X = 0$); β_1 = pente (variation moyenne de Y quand X augmente de 1).
- **Remarques** (suggestions) : b_0 et b_1 sont sans biais; la précision de b_1 augmente avec n et avec l'étalement des x_i (S_{XX}); σ^2 est inconnue et estimée par MSE; on utilise alors la loi t_{n-2} .
- **Estimations** : $b_0 = 1,37$ kg (production prédite sans soleil, extrapolation hors du domaine observé); $b_1 = 0,43$: chaque heure de soleil supplémentaire augmente la production moyenne d'environ 0,43 kg.
- **Quiz (tomates)** : (1) $2 \times 0,431 = 0,86$ kg de plus pour le plant A. (2) Faux : la valeur de b_1 dépend des unités (en grammes, $b_1 = 431$ g/h); pour juger du lien, il faut un test ou un IC. (3) Faux : $x = 0$ est hors du domaine observé (3 à 8 h), c'est de l'extrapolation. (4) $1,372 + 0,431 x = 4 \Rightarrow x = (4 - 1,372)/0,431 = 6,1$ h par jour.
- **Quiz (hirondelles, lecture)** : $\hat{y} = -4,43 + 0,328 x$; lien positif. Chaque gramme de plus est associé à une étendue des ailes plus grande de 0,33 cm en moyenne. $b_0 = -4,43$ cm serait l'étendue moyenne d'un oiseau de 0 g : absurde (étendue négative); extrapolation, car les poids observés vont de 39 à 57 g.
- **IC à 95 %** ($t_{58; 0,025} \approx 2,00$) : pente $[0,345; 0,517]$, ordonnée $[0,877; 1,868]$. Comme 0 n'est pas dans l'IC de la pente, la relation est significative au seuil de 5 %.

Réponses aux exemples (2)

- **Quiz (hirondelles, test t)** : $t_{obs} = 11,70$, valeur- $p = 8,05 \times 10^{-14} < 0,05$: on rejette H_0 ; le lien linéaire (positif) est significatif. IC : $0,328 \pm 2,028 \times 0,02806 = 0,328 \pm 0,057$, soit $[0,271; 0,385]$: avec 95 % de confiance, chaque gramme de plus est associé à une étendue moyenne plus grande de 0,27 à 0,39 cm. Comme 0,5 n'est pas dans l'IC, les données contredisent l'ornithologue (au seuil de 5 %).
- **Hypothèses** : $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$.
- **Table** : SSR = 32,547 (1 d.l.), SSE = 18,721 (58 d.l.), SST = 51,268 (59 d.l.); MSR = 32,547, MSE = 0,323, $F = 100,8 = 10,042^2$; $R^2 = 0,635$.
- **Conséquence** : sous H_0 , $F = \text{MSR}/\text{MSE} \sim F_{1, n-2}$; sous H_1 , F_{obs} tend à être plus grande.
- **Exemple** : valeur- $p = 2,7 \times 10^{-14} < 0,05$: on rejette H_0 (de même, $F_{obs} = 100,8 > F_{1,58; 0,05} = 4,01$).
- **Quiz (vrai ou faux)** : (1) Vrai : si $\beta_1 = 0$, b_1 est près de 0, donc $\hat{y}_i = \bar{y} + b_1(x_i - \bar{x}) \approx \bar{y}$ et SSR est petite. (2) Faux : b_1 varie d'un échantillon à l'autre autour de 0. (3) Vrai : $\sqrt{6,25} = 2,5$. (4) Faux : association n'est pas causalité ; seule une expérience contrôlée permet de conclure à un effet causal.