Corrélation linéaire et régression linéaire simple

Ségolen Geffray

IUT Carquefou

Année 2008-2009

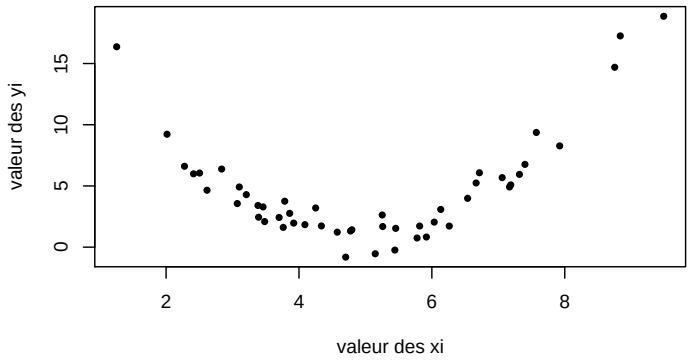
segolen.geffray@univ-nantes.fr

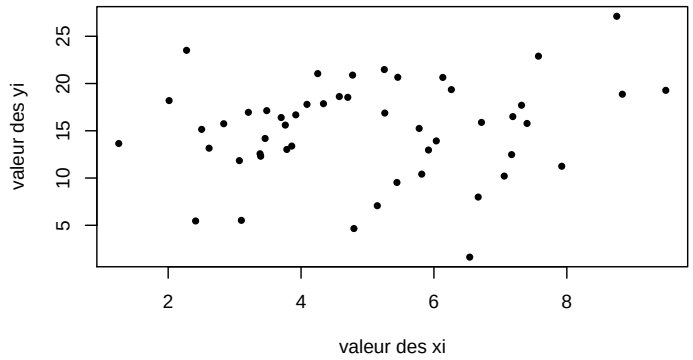
Notion de corrélation

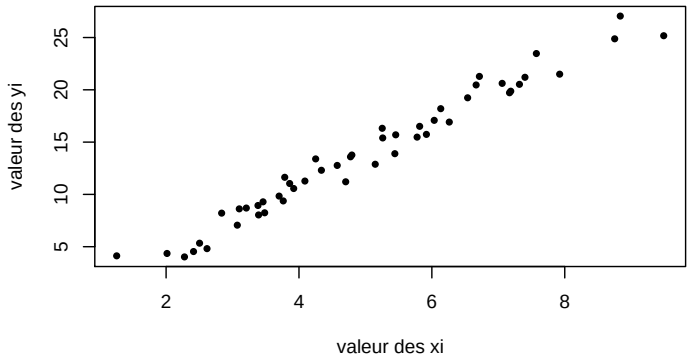
- Contexte : on soupçonne qu'il existe une liaison entre deux variables X et Y. Par exemple, existe-t-il un lien entre le volume des ventes d'une entreprise et le montant alloué à la publicité ? De même, existe-t-il un lien entre le poids de courrier reçu par une entreprise chaque matin et le nombre de commandes traitées dans la journée ?
- Notion : on dit qu'il y a corrélation entre deux variables lorsqu'elles ont tendance à varier soit toujours dans le même sens (par exemple, si X augmente, Y a tendance à augmenter aussi), soit toujours en sens inverse (par exemple, si X augmente, Y a tendance à diminuer).
- Questions mathématiques :
  - Peut-on quantifier cette liaison ?
  - Peut-on tester si cette liaison est statistiquement significative ?
  - Peut-on utiliser cette liaison à des fins prédictives ?

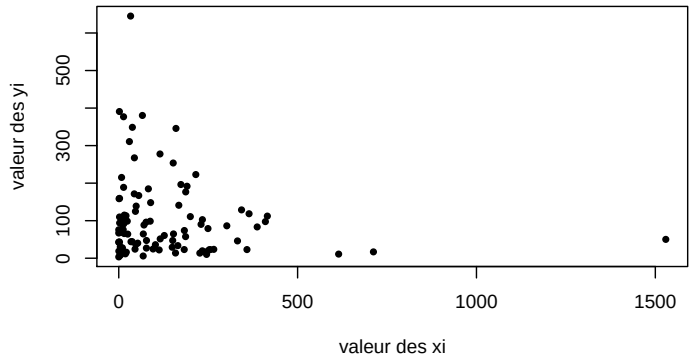
Observations de couples de variables

Regarder ses données ! ! ! Tracer le nuage de points : y a-t-il liaison linéaire, non-linéaire, pas de liaison ? ? ? ?









Coefficient de corrélation linéaire

- Le coefficient de corrélation de Pearson ρ mesure le degré d'association **linéaire** entre X et Y :
$$\rho = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sigma(X)\sigma(Y)}.$$
- ρ est un nombre forcément compris entre -1 et 1.
- Le nombre ρ sert à quantifier l'intensité et le sens de la **dépendance linéaire** entre X et Y.
  - Lorsque $\rho > 0$, cela signifie que lorsqu'une des variables a tendance à augmenter, l'autre aussi.
  - Lorsque $\rho < 0$, cela signifie que lorsqu'une des variables a tendance à augmenter, l'autre a tendance à diminuer.
  - Lorsque $\rho = 0$, on dit que X et Y sont non corrélées : il n'y a pas d'association linéaire entre X et Y.
  - $\rho \pm 1 \Leftrightarrow$ l'une des variables est une fonction affine de l'autre, par exemple Y est une fonction affine de X i.e. $Y = aX + b$ avec b du signe de ρ.
- Lorsque X et Y sont indépendantes, $\rho = 0$ mais la réciproque est fausse ! ! ! Si $\rho = 0$, X et Y ne sont pas forcément indépendantes, par ex : soit X de loi normale et soit $Y = X^2$, alors $\rho = 0$ mais X et Y ne sont pas indépendantes.

Estimation du coefficient de corrélation

- **Les données** : pour chaque individu d'un échantillon de taille n, on relève les valeurs prises par X et Y. On obtient n couples **indépendants** les uns des autres notés (x_i, y_i) pour $i = 1, ..., n$.
- Un estimateur de ρ est :

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right)}}$$

  avec $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.
- r est un nombre compris entre 1 et -1.
- Lorsque les points de coordonnées (x_i, y_i) pour $i = 1, ..., n$ sont parfaitement alignés, alors $r = 1$.
- Lorsqu'on obtient un nuage flou de points, r est proche de 0.
- Plus les points sont étroitement concentrés autour d'une droite, plus r est proche de 1. C'est la concentration des points autour de la droite en question qui indique l'intensité de la liaison tandis que c'est la pente de la droite qui indique le sens de la liaison.

Test de l'hypothèse $H_0 : \rho = \rho_0$

- **Conditions d'application** : $n \geq 30$
- Soit un échantillon $(X_1, Y_1), ..., (X_n, Y_n)$ un échantillon.
- On teste $H_0 : \rho = \rho_0$ contre une hypothèse alternative H_1. La forme de la région de rejet dépend de la forme de H_1.
- L'erreur de 1ère espèce est fixée à α.

H_1	Décision
$\rho \neq \rho_0$	Rejeter H_0 si $\log\left(\frac{1+r}{1-r}\right) > \log\left(\frac{1+\rho_0}{1-\rho_0}\right) + F_N^{-1}(1-\alpha/2)\frac{2}{\sqrt{n-3}}$ ou $\log\left(\frac{1+r}{1-r}\right) < \log\left(\frac{1+\rho_0}{1-\rho_0}\right) - F_N^{-1}(1-\alpha/2)\frac{2}{\sqrt{n-3}}$
$\rho > \rho_0$	Rejeter H_0 si $\log\left(\frac{1+r}{1-r}\right) > \log\left(\frac{1+\rho_0}{1-\rho_0}\right) + F_N^{-1}(1-\alpha)\frac{2}{\sqrt{n-3}}$
$\rho < \rho_0$	Rejeter H_0 si $\log\left(\frac{1+r}{1-r}\right) < \log\left(\frac{1+\rho_0}{1-\rho_0}\right) - F_N^{-1}(1-\alpha)\frac{2}{\sqrt{n-3}}$

Régression linéaire

- **Les données** : on dispose d'un échantillon de n couples (x_i, y_i) pour $i = 1, ..., n$ **indépendants** les uns des autres.
- **Régression linéaire** : on cherche d'une relation **linéaire** entre X=**variable explicative**=**variable de régression** et Y=**variable à expliquer**=**réponse**.
- **Modèle linéaire** : $Y = aX + b + \varepsilon$ où ε est une variable aléatoire appelée **erreur résiduelle** satisfaisant $\mathbb{E}[\varepsilon] = 0$ et $\mathrm{Var}(\varepsilon) = \sigma^2$.
- Droite de régression : $y = ax + b$ à ajuster sur les données au sens des moindres carrés
- Droite de régression estimée (meilleure droite ajustée) : $y = \widehat{a}x + \widehat{b}$ avec $\widehat{a}$=estimateur de a et $\widehat{b}$=estimateur de b :

$$\widehat{a} = \frac{\sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

$$\widehat{b} = \overline{y} - \widehat{a}\,\overline{x}$$

- On appelle **réponse prédite** au point x_i la valeur $\widehat{y}_i = \widehat{a}x_i + \widehat{b}$.
- On appelle **résidu** au point x_i la valeur $e_i = y_i - \widehat{y}_i$ représentant la différence entre la réponse observée y_i et la réponse prédite $\widehat{y}_i$.

Analyse des sources de variabilité

- La variabilité totale des données (correspondant à SCT) se décompose entre la variabilité expliquée par le modèle de régression (correspondant à SCR) et la variabilité résiduelle ou terme d'erreur (correspondant à SCE) de la façon suivante :

$$SCT = SCR + SCE \qquad \text{avec}$$

  - $SCT = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2$ =somme des carrés totale
  - $SCE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ =somme des carrés d'erreur
  - $SCR = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$ =somme des carrés de régression
- On appelle moyenne des carrés de régression le terme MCE défini par :

$$MCE = \frac{SCE}{n-2}.$$

- NB : on peut noter :

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \quad S_{yy} = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 \quad S_{xy} = \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}$$

Intervalle de confiance pour les paramètres

- Un IC bilatéral au niveau de confiance $1-\alpha$ pour le paramètre a est donné par :

$$\mathbb{P}[\widehat{a}_{inf} \leq a \leq \widehat{a}_{sup}] = 1-\alpha$$

avec

$$\widehat{a}_{inf} = \widehat{a} - F^{-1}_{T(n-2)}(1-\alpha/2)\sqrt{\frac{MCE}{S_{xx}}}$$

$$\widehat{a}_{sup} = \widehat{a} + F^{-1}_{T(n-2)}(1-\alpha/2)\sqrt{\frac{MCE}{S_{xx}}}$$

- Un IC bilatéral au niveau de confiance $1-\alpha$ pour le paramètre b est donné par :

$$\mathbb{P}[\widehat{b}_{inf} \leq b \leq \widehat{b}_{sup}] = 1-\alpha$$

avec

$$\widehat{b}_{inf} = \widehat{b} - F^{-1}_{T(n-2)}(1-\alpha/2)\sqrt{MCE\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)}$$

$$\widehat{b}_{sup} = \widehat{b} + F^{-1}_{T(n-2)}(1-\alpha/2)\sqrt{MCE\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)}$$

Test pour le paramètre a ou significativité de la régression

- On teste l'hypothèse nulle $H_0 : a = 0$ contre l'hypothèse alternative H_1 avec un risque de 1ère espèce fixé à α.
- La forme de la région de rejet dépend de la forme de H_1.

H_1	Décision
$a \neq 0$	Rejeter H_0 si $\widehat{a} < -F^{-1}_{T(n-2)}(1-\alpha/2)\sqrt{\frac{MCE}{S_{xx}}}$ ou si $\widehat{a} > F^{-1}_{T(n-2)}(1-\alpha/2)\sqrt{\frac{MCE}{S_{xx}}}$
$a > 0$	Rejeter H_0 si $\widehat{a} > F^{-1}_{T(n-2)}(1-\alpha)\sqrt{\frac{MCE}{S_{xx}}}$
$a < 0$	Rejeter H_0 si $\widehat{a} < -F^{-1}_{T(n-2)}(1-\alpha)\sqrt{\frac{MCE}{S_{xx}}}$

- Si on ne rejette pas H_0, cela signifie que les données ne permettent pas de mettre en évidence une influence linéaire de X sur Y.

Diagnostic de régression

- **Coefficient de détermination** : le coefficient de détermination multiple est le nombre R^2 défini par :

$$R^2 = \frac{SCR}{SCT}$$

Ce coefficient est une mesure de la variabilité expliquée par le modèle de régression linéaire. Il vérifie toujours $0 \leq R^2 \leq 1$. Plus R^2 est proche de 1, plus le modèle choisi semble pertinent.

- **Analyse des résidus** : on appelle résidus studentisés les termes r_i pour $i = 1, ..., n$ définis par :

$$r_i = \frac{e_i}{\sqrt{\left(1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{S_{xx}}\right) MCE}}$$

On analyse les résidus studentisés en disant que si $|r_i| > 2$ alors

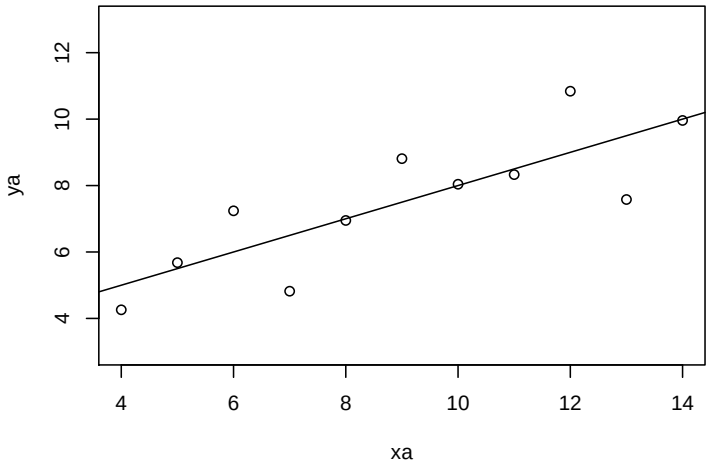
  - soit y_i est une observation aberrante,
  - soit y_i est dans une région où le modèle estimé n'est pas réaliste.

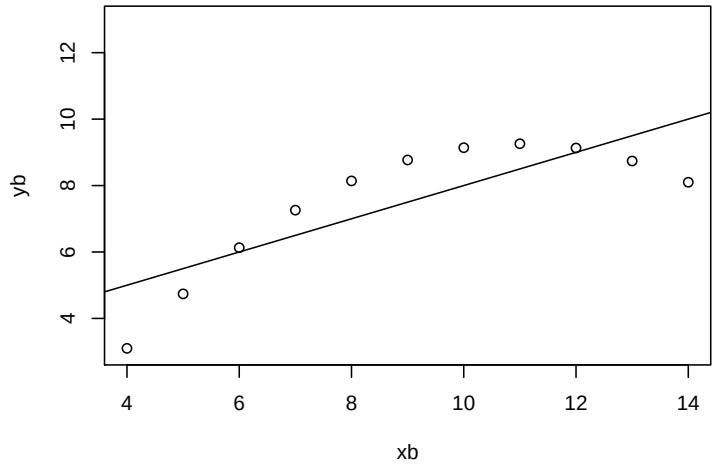
Du bon usage du coefficient de corrélation linéaire

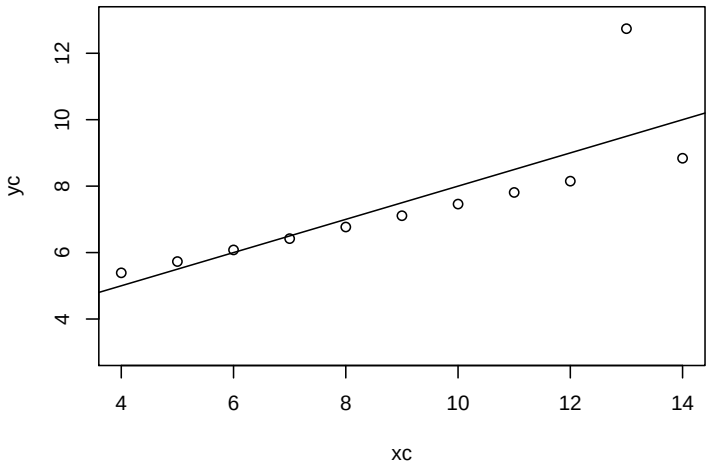
données A		données B		données C		données D	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

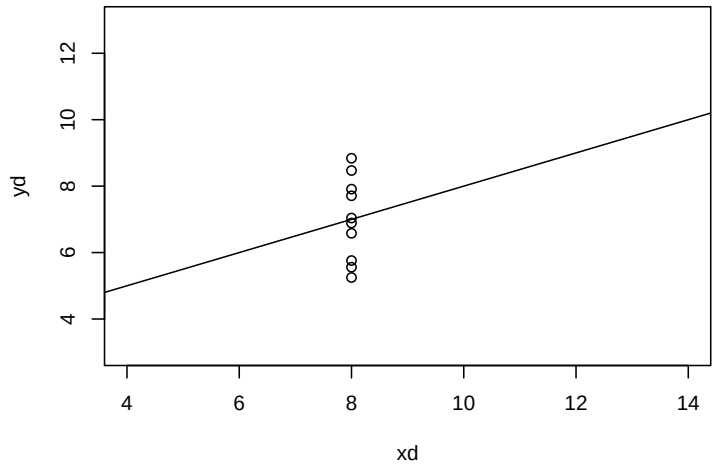
Dans les 4 cas, on a $\overline{x} = 9$, $\overline{y} = 7.5$, $\sum_{i=1}^{n} x_i^2 = 1001$, $\sum_{i=1}^{n} y_i^2 = 660$ et $\sum_{i=1}^{n} x_i y_i = 797.5$. On obtient donc $r = 0.816$ et la droite de régression $y = 0.5x + 3$.

Du bon usage du coefficient de corrélation linéaire (suite)









Exercice : régression linéaire (1)

La gérante d'un commerce veut évaluer l'impact des frais déboursés en publicité par mois (représentés par une variable X exprimée en milliers d'euros) sur le chiffre d'affaires mensuel (représenté par une variable Y exprimée en milliers d'euros). On aimerait évaluer dans quelle mesure une modification du budget publicitaire mensuel affecterait le chiffre d'affaires mensuel. On a donc recueilli sur une période de 10 mois les données du tableau ci-dessous.

chiffre d'affaires	220	280	250	170	150	340	310	210	180	190
frais publicitaires	2.6	2.6	2.4	1.5	0.9	3.0	2.7	2.3	1.7	1.9

1. Tracer le nuage de points et estimer le coefficient de corrélation linéaire.
2. Etablir la droite de régression correspondant à ce problème et tracer cette droite.
3. Déterminer un intervalle de confiance au risque 5% des paramètres de la droite de régression.
4. Tester la significativité de la régression au risque 5%.
5. Calculer le coefficient de détermination.
6. Calculer les résidus studentisés. Y-a-t-il des valeurs aberrantes ou mal expliquées par le modèle ?
7. Quel serait le chiffre d'affaires mensuel prédit par le modèle pour un budget publicitaire mensuel de 400 euros ? de 4000 euros ?

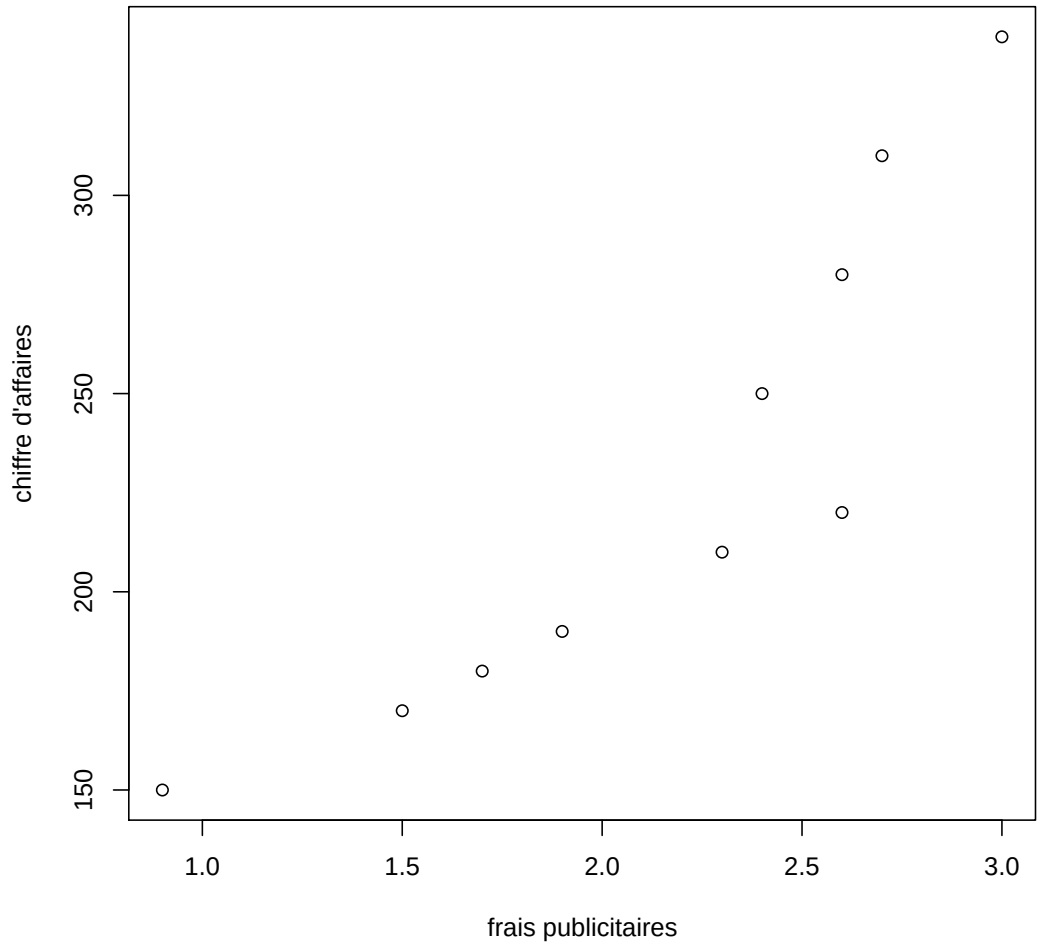
Exercice : régression linéaire (2)

La société Métalex moule des pièces dans un four. L'ingénieur se demande s'il existe un lien entre la température (en degré celsius) à laquelle les pièces sont moulées et leur résistance (en kg/cm^2). Il dispose des données suivantes transmises par l'atelier.

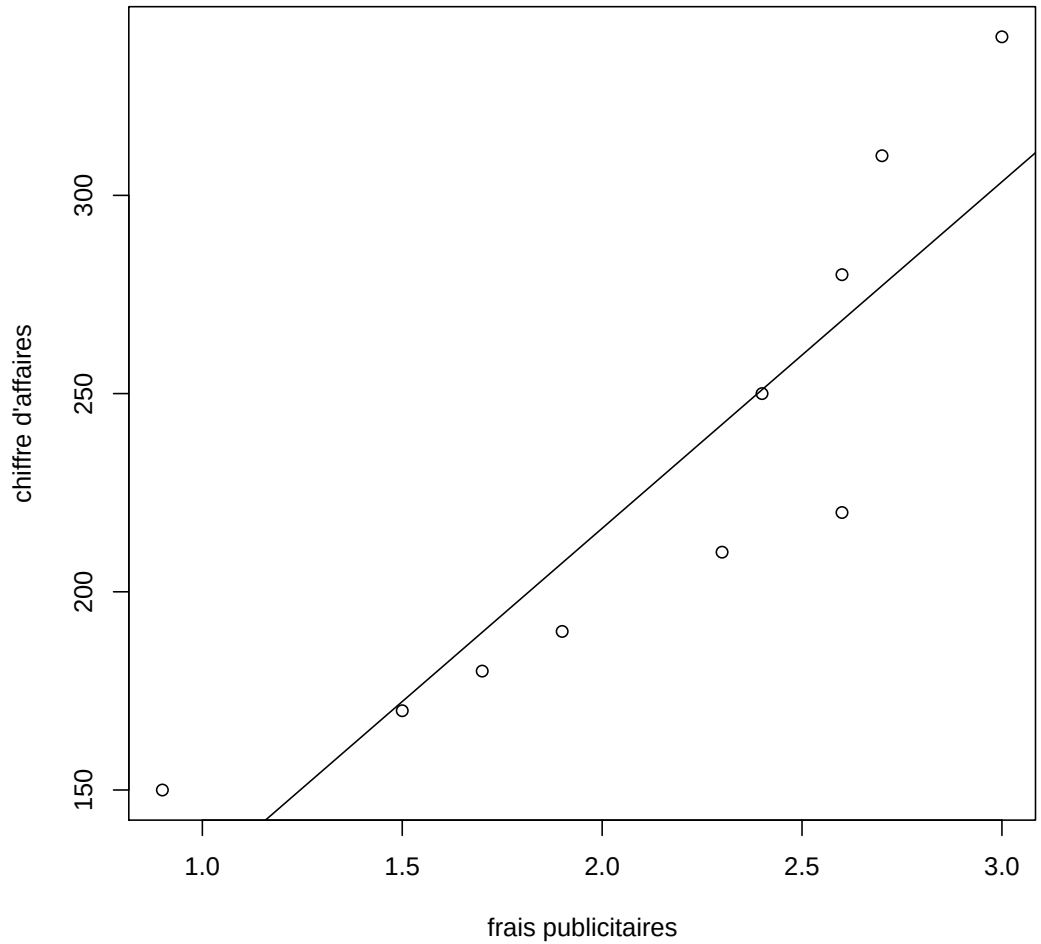
température	100	120	140	160	180	200	220	240	260	280	300
résistance	46	48	49	51	52	53	54	55	56	56	56

1. Tracer le nuage de points et estimer le coefficient de corrélation linéaire.
2. Ajuster un modèle linéaire de la forme $Y = aX + b + \varepsilon$: établir la droite de régression correspondant à ce problème et tracer cette droite.
3. Tester la significativité de la régression au risque 5%.
4. Calculer le coefficient de détermination.
5. Calculer les résidus studentisés. Y-a-t-il des valeurs aberrantes ou mal expliquées par le modèle ?
6. Ajuster un modèle non-linéaire de la forme $Y = a\log(X) + b + \varepsilon$: établir la courbe de régression correspondant à ce problème et tracer cette courbe.
7. Tester la significativité de la régression au risque 5%.
8. Calculer le coefficient de détermination.
9. Calculer les résidus studentisés. Y-a-t-il des valeurs aberrantes ou mal expliquées par le modèle ?

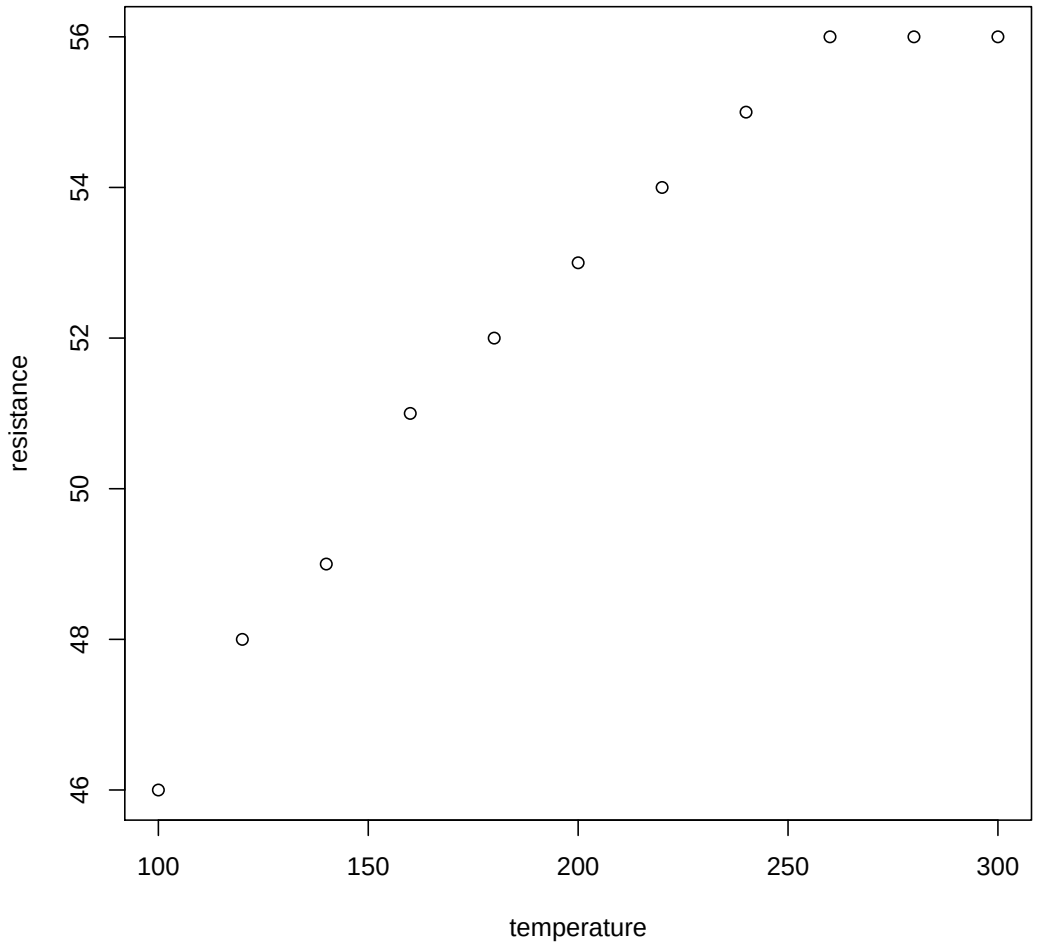
Nuage de points de l'exercice (1)



Modèle ajusté de l'exercice (1)



Nuage de points de l'exercice (2)



Modèles ajustés de l'exercice (2)