Modèle Linéaire Gaussien
Chapitre 1 – Régression linéaire simple

Achille Thin

7 Décembre 2023

Executive Master Statistique et Big Data

Objectifs du Chapitre

Cadre de la régression linéaire simple: expliquer les variations d'**une variable quantitative**, appelée variable réponse, **en fonction d'une variable quantitative**, appelée variable explicative.

Démarche statistique:

1. Écriture du modèle
2. Ajustement (estimation) du modèle grâce aux données
3. Vérification de la validité des hypothèses faites dans le modèle
4. Intervalles de confiance sur les paramètres
5. Test de la pertinence des différents éléments du modèle
6. Critique du modèle
7. Conclusion

Objectifs

Question : le nombre d'années d'étude permet-il de prédire le salaire horaire d'une personne ?

- **Variable réponse:** le salaire.
- **Variable explicative:** le nombre d'années d'étude.

Chargement des données:

```
donnees <- read.table("../data/wagesmicrodata.csv",
                      sep = ";", dec = ",", header = TRUE,
                      colClasses = c(rep("numeric", 2),
                                     "factor",
                                     rep("numeric" , 3),
                                     rep("factor", 6)))
```

Analyse descriptive

Coefficient de corrélation linéaire empirique

On a $n = 534$ observations. On note, pour $1 \leq k \leq n$

- x_k le nombre d'années d'étude de l'individu k. On note $\mathbf{x}$ l'échantillon complet, et $\overline{x}$ la valeur moyenne (`mean`) de l'échantillon ;
- y_k le salaire horaire de l'individu k. On note $\mathbf{y}$ l'échantillon complet, et $\overline{y}$ la valeur moyenne (`mean`) de l'échantillon.

Le corrélation linéaire empirique (`cor`) est donnée par :

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y})}{\sqrt{\sum_{k=1}^{n} (x_k - \overline{x})^2 \sum_{k=1}^{n} (y_k - \overline{y})^2}}$$

Interprétation de $\rho(\mathbf{x}, \mathbf{y})$:

- $-1 \leq \rho(\mathbf{x}, \mathbf{y}) \leq 1$
- Proche de 0 : $\mathbf{x}$ et $\mathbf{y}$ ne sont pas corrélés linéairement
- Proche de 1 : corrélation linéaire positive entre $\mathbf{x}$ et $\mathbf{y}$. Quand $\mathbf{x} \nearrow$, $\mathbf{y} \nearrow$
- Proche de -1 : corrélation linéaire négative entre $\mathbf{x}$ et $\mathbf{y}$. Quand $\mathbf{x} \nearrow$, $\mathbf{y} \searrow$

Avec R : coefficient de corrélation linéaire empirique

```
x_bar <- mean(donnees$EDUCATION)
y_bar <- mean(donnees$WAGE)
rho_xy <- cor(donnees$EDUCATION, donnees$WAGE)
```

Ici, on observe : $\bar{x} = 13.02$, $\bar{y} = 9.02$, $\rho(x, y) = 0.38$.

La fonction `cor.test` réalise un test statistique qui permet de tester si le coefficient de corrélation est nul ou non.

```
cor.test(donnees$EDUCATION, donnees$WAGE)


Pearson's product-moment correlation

data:  donnees$EDUCATION and donnees$WAGE
t = 9.53, df = 532, p-value <2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.30702 0.45212
sample estimates:
    cor
0.38192
```

Écriture du modèle

Modèle de régression linéaire simple

On suppose que y_k est la réalisation d'une variable aléatoire Y_k telle que :

$$Y_k = ax_k + b + \varepsilon_k, \quad 1 \leq k \leq n = 534,$$

- x_k est la valeur de la variable explicative pour l'observation k.
- b est un paramètre inconnu,
- a est un paramètre inconnu (représente l'effet de l'éducation sur le salaire),
- ε_k une variable aléatoire appelée **bruit**, telle que toutes les variables aléatoires $\varepsilon_1, \ldots, \varepsilon_n$ sont **indépendantes**, d'**espérance nulle** et ont la **même variance**, égale à σ^2 (paramètre inconnu).

Cas particulier du modèle linéaire gaussien: les variables aléatoires ε_k sont indépendantes et identiquement distribuées de loi $\mathcal{N}(0, \sigma^2)$.

Écriture matricielle:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \overset{\text{Notation}}{\Rightarrow} \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Ajustement du modèle

Estimations et estimateurs

Objectif: trouver a, b et σ^2 qui s'ajustent le mieux aux $n = 534$ observations dont on dispose.

Les observations y_k sont supposées être des réalisations de variables aléatoires Y_k dont on spécifie (contrôle) la loi de probabilité.

Estimateur: c'est une variable aléatoire qui s'exprime comme une fonction de $Y_1, \ldots, Y_n$ et décrit comment approcher un paramètre inconnu du modèle statistique choisi.

Estimation: c'est une valeur numérique. C'est la réalisation de l'estimateur pour les données observées. Autrement c'est une fonction de $y_1, \ldots, y_n$ obtenues en remplaçant $Y_1, \ldots, Y_n$ par $y_1, \ldots, y_n$ dans l'expression de l'estimateur.

Ainsi, l'estimation fournit une approximation d'un paramètre inconnu calculée à l'aide de nos observations sous l'hypothèse du modèle statistique choisi. L'estimateur permet d'étudier les propriétés probabilistes de l'approximation du paramètre. Il permet notamment de caractériser l'incertitude sur les valeurs estimées si l'on avait un autre jeu d'observations.

Estimateurs de a et b : fonction de perte (*loss function*)

Fonction de perte ℓ: mesure la proximité de la droite au nuage de points. On cherche la droite qui approche le mieux le nuage de points, *i.e.*, a et b tels que

$$\arg\min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} \ell(Y_i - aX_i - b).$$

Exemples de perte:

1. Perte ou coût quadratique : $\ell : u \mapsto u^2$ ($\rightsquigarrow$ facilite les calculs)
2. Perte ou coût absolu·e : $\ell : u \mapsto |u|$ ($\rightsquigarrow$ moins sensible aux valeurs aberrantes)

Estimateurs et estimations de a et b par les moindres carrés

On considère la perte quadratique. Les estimateurs $(\widehat{A}, \widehat{B})$ de (a, b) sont définis par

$$(\widehat{A}, \widehat{B}) = \underset{(a,b)\in\mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} (Y_i - ax_i - b)^2 = \underset{(a,b)\in\mathbb{R}^2}{\arg\min} \|\mathbf{Y} - \mathbf{x}\boldsymbol{\beta}\|^2.$$

Estimateurs (variables aléatoires)

$$\widehat{A} = \frac{\sum_{k=1}^{n}(x_k - \overline{x})(Y_k - \overline{Y})}{\sum_{k=1}^{n}(x_k - \overline{x})^2}, \quad \widehat{B} = \overline{Y} - \widehat{A}\overline{x}.$$

Estimations (valeurs numériques calculée sur les données)

$$\widehat{a} = \frac{\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y})}{\sum_{k=1}^{n}(x_k - \overline{x})^2}, \quad \widehat{b} = \overline{y} - \widehat{a}\overline{x}.$$

Avec R : estimations de a et b

Construction du modèle de régression linéaire simple:

```
reg <- lm(WAGE ~ EDUCATION, data = donnees)
```

Cette fonction fournit une liste de nombreux éléments (que nous apprendrons à utiliser dans la suite du chapitre)

```
names(reg)
```

```
 [1] "coefficients"  "residuals"     "effects"
 [4] "rank"          "fitted.values" "assign"
 [7] "qr"            "df.residual"   "xlevels"
[10] "call"          "terms"         "model"
```

Accès aux estimations de a et b:

```
reg$coefficients  # ou coef(reg)
```

```
(Intercept)   EDUCATION
   -0.74598     0.75046
```

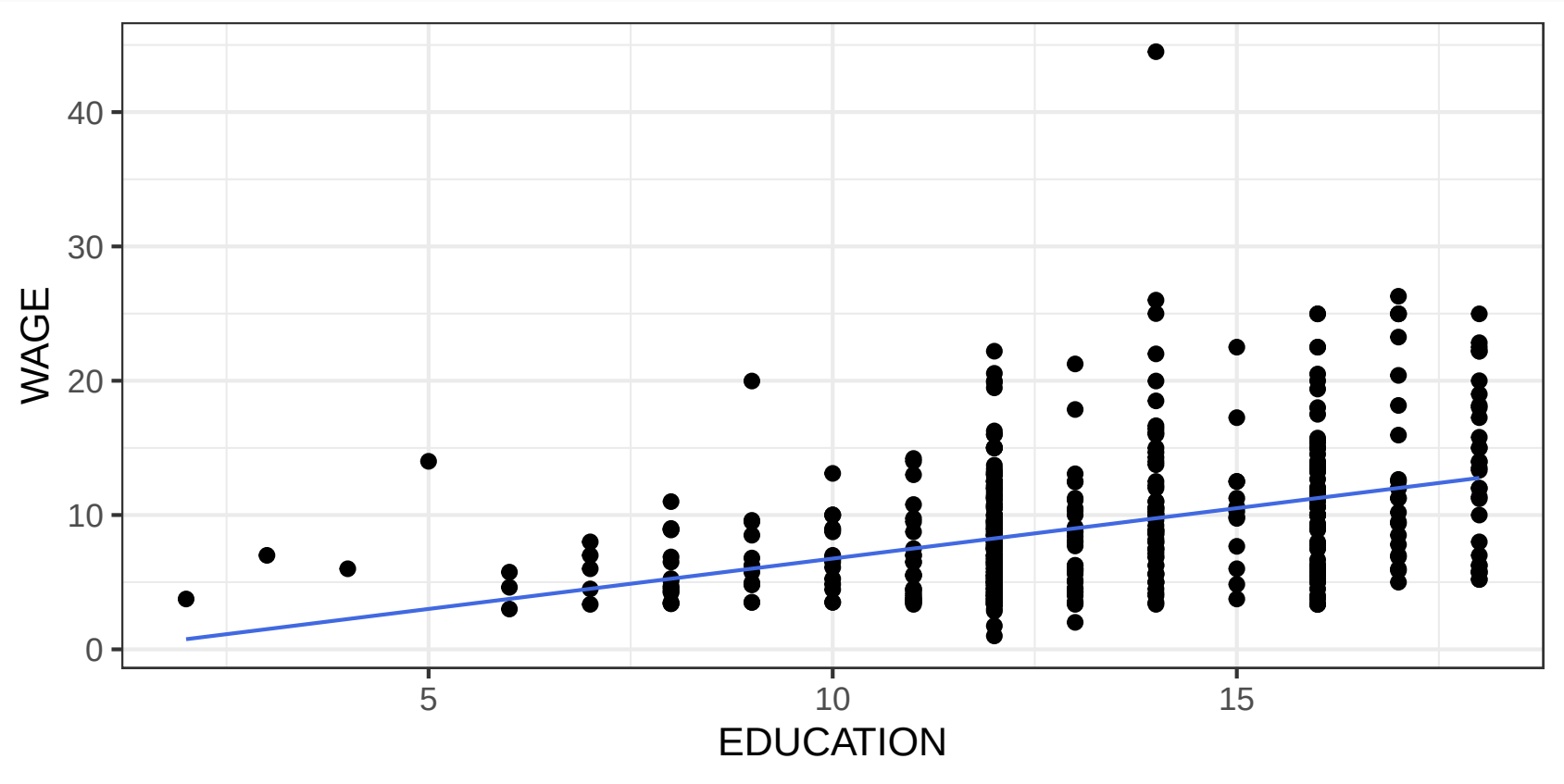
Cette fonction renvoie le vecteur $(\widehat{b}, \widehat{a})$.

Avec R : droite de régression

Accès aux valeurs ajustées: $\widehat{y}_k = \widehat{a}x_k + \widehat{b}$

```
reg$fitted.values   # ou fitted(reg)
```

Représentation de la droite: $y = \widehat{a}x + \widehat{b}$, avec $\widehat{a} = 0.75046$ et $\widehat{b} = -0.74598$.



Propriétés des estimateurs $\widehat{A}$ et $\widehat{B}$

Question: doit-on avoir confiance en notre estimation ? Si on avait un autre échantillon, l'estimation aurait-elle beaucoup varié ?

Loi de l'estimateur des moindres carrés: sous l'hypothèse du modèle linéaire gaussien,

$$\begin{pmatrix} \widehat{A} \\ \widehat{B} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} a \\ b \end{pmatrix}, \sigma^2 V \right), \quad \text{avec} \quad V = \frac{1}{\sum_{k=1}^{n} (x_k - \overline{x})^2} \begin{pmatrix} 1 & -\overline{x} \\ -\overline{x} & \frac{1}{n} \sum_{k=1}^{n} x_k^2 \end{pmatrix}.$$

Conséquences: les estimateurs $\widehat{A}$ et $\widehat{B}$ sont

- sans biais, *i.e.*, $\mathbb{E}\left[\widehat{A}\right] = a$ et $\mathbb{E}\left[\widehat{B}\right] = b$,
- de faible variance lorsque :
  - le bruit (σ^2) est faible ($\rightsquigarrow$ inconnu),
  - $\sum_{k=1}^{n} (x_k - \overline{x})^2$ est grande (valeurs de x dispersées),
  - $\sum_{k=1}^{n} x_k^2$ est petite.

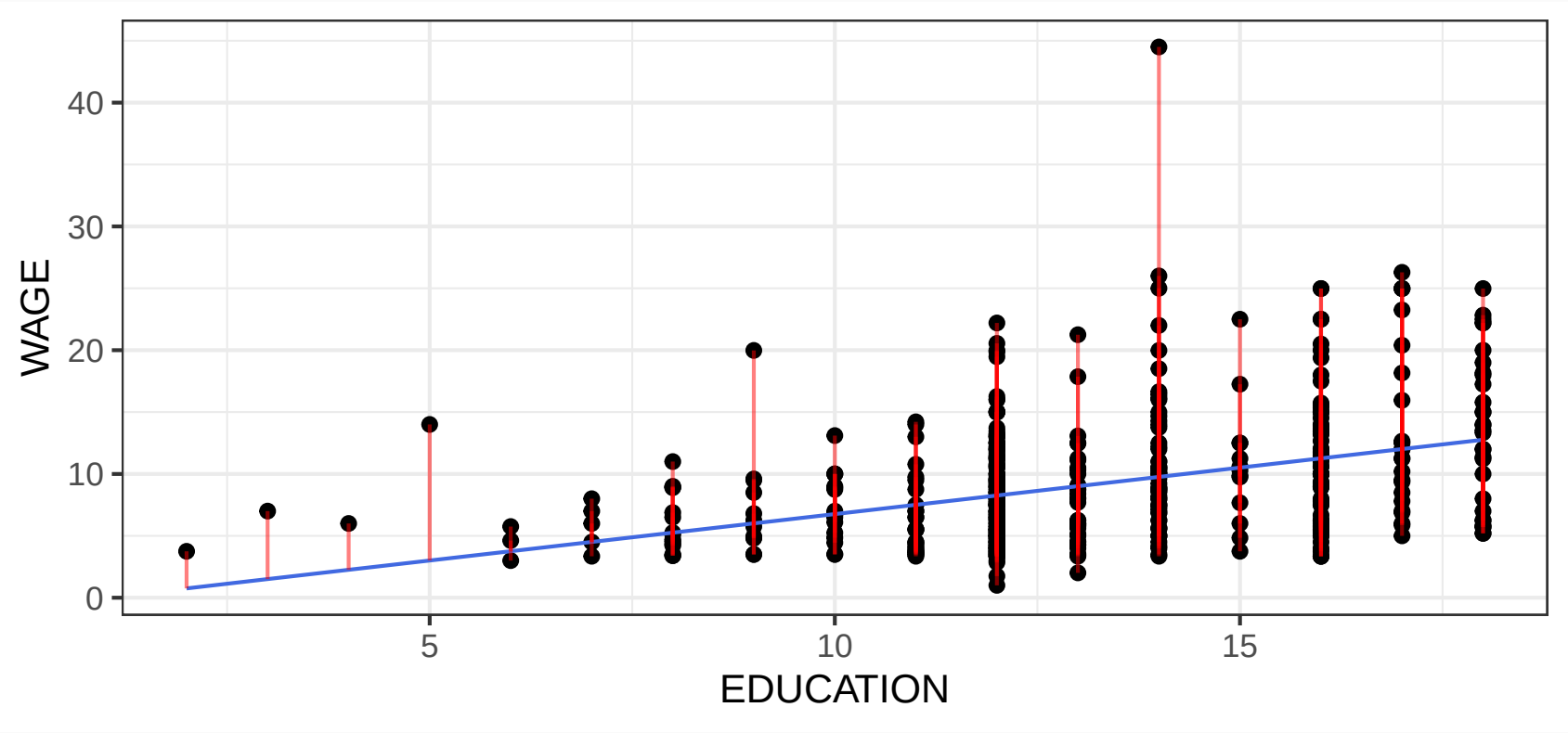
Résidus

Résidus (variables aléatoires) estimateurs des erreurs inconnues ε_k :

$$\widehat{\varepsilon}_k = Y_k - \widehat{Y}_k, \quad \text{où} \quad \widehat{Y}_k = \widehat{A}x_k + \widehat{B} \quad \text{(variable ajustée pour } Y_k\text{)}$$

Résidus observés: $\widehat{e}_k = y_k - \widehat{y}_k$, où $\widehat{y}_k = \widehat{a}x_k + \widehat{b}$ (valeur ajustée).

```
reg$residuals   # ou resid(reg)
```



Propriétés des résidus et estimateurs de σ^2

Propriétés des résidus: $\widehat{\varepsilon} = (\widehat{\varepsilon}_1, \dots, \widehat{\varepsilon}_n)^\intercal$ est un vecteur aléatoire

1. gaussien,
2. indépendant de $(\widehat{A}, \widehat{B})$, et donc indépendant de $(\widehat{Y}_1, \dots, \widehat{Y}_n)^\intercal$.

Estimateur de la variance du bruit (σ^2) (variable aléatoire)

$$S^2 = \frac{1}{n-2} \sum_{k=1}^{n} \widehat{\varepsilon}_k^2.$$

Cet estimateur vérifie

$$(n-2)\frac{S^2}{\sigma^2} = \sum_{k=1}^{n} \frac{\widehat{\varepsilon}_k^2}{\sigma^2} \sim \chi^2(n-2),$$

où $\chi^2(n-2)$ désigne la loi du Khi-deux à $n-2$ degrés de liberté.

Avec R : estimation de σ^2

Estimation de σ^2 (réalisation sur les données)

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{k=1}^{n} \widehat{e}_k^2.$$

Accès à la valeur estimée: on passe par la fonction `summary` qui fournit un résumé de l'ensemble des sorties contenues dans `reg`. La quantité « *residual standard error* » est l'estimation de l'écart-type σ. On peut accéder à $\widehat{\sigma}^2$ comme suit

```
summary(reg)$sigma^2
```

```
[1] 22.6
```

Bilan

Dans cette partie, on a :

1. construit des estimateurs des paramètres inconnus a, b, σ^2 et des erreurs inconnues $\varepsilon_1, \ldots, \varepsilon_n$,
2. établi les propriétés théoriques de ces estimateurs sous les hypothèses du modèle linéaire gaussien,
3. obtenu les estimations à l'aide de la fonction `lm`.

Les résultats théoriques tout comme les valeurs fournies par `lm` supposent que les erreurs $\varepsilon_1, \ldots, \varepsilon_n$ sont *i.i.d.* suivant la loi normale $\mathcal{N}(0, \sigma^2)$. **Mais ce postulat est-il vérifié sur nos données ?**

Validité des hypothèses

Diagnostique des résidus

Les résidus observés permettent de valider les hypothèses du modèle linéaire gaussien, *i.e.*, les variables aléatoires $\varepsilon_1, \dots, \varepsilon_n$

(P1) sont **indépendantes**,

(P2) sont toutes d'**espérance nulle** ($\rightsquigarrow$ la relation entre y et x est bien affine),

(P3) ont la **même variance** σ^2 (homoscédasticité),

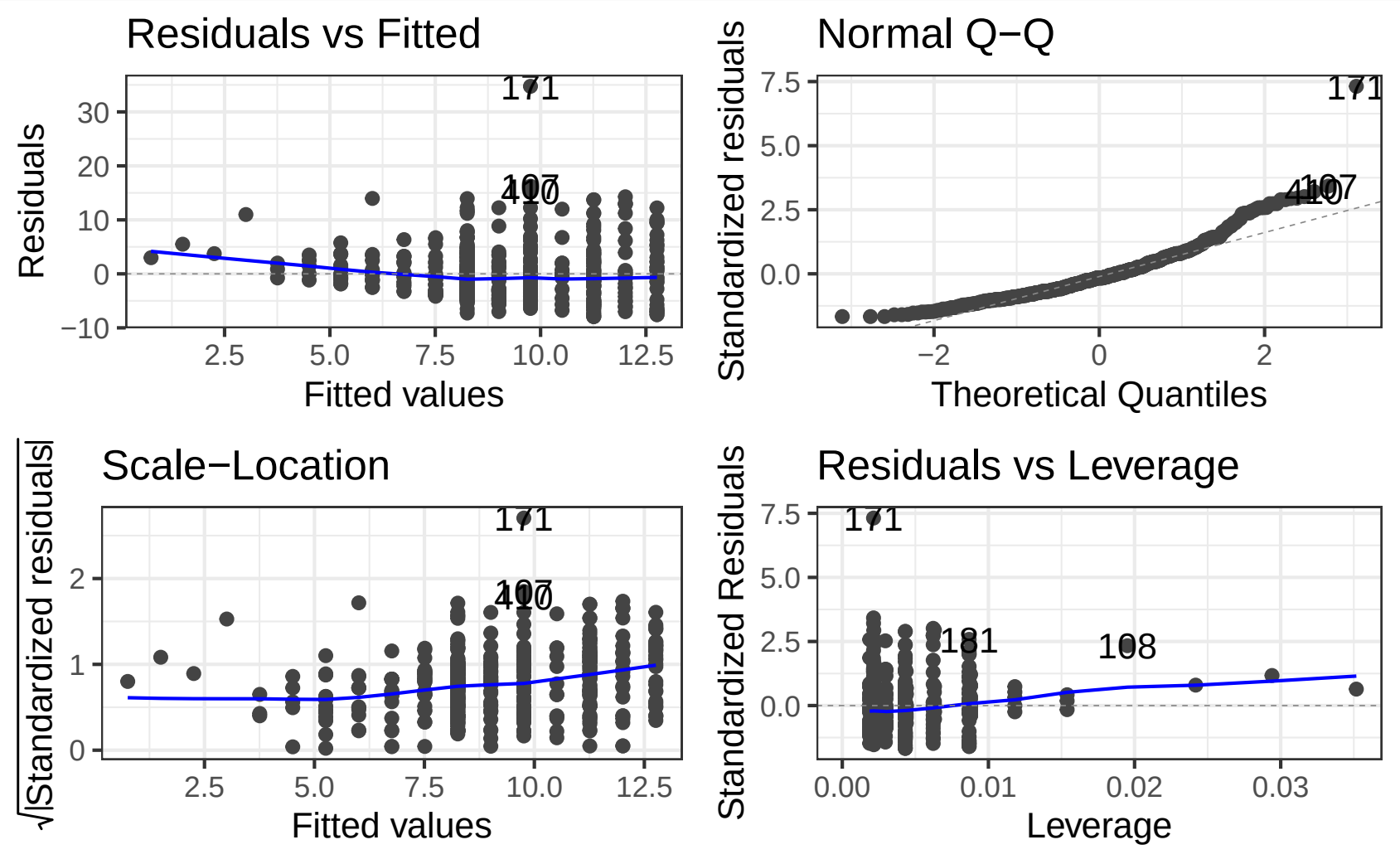
(P4) suivent une **loi normale**.

Validation des hypothèses:

- (P1) : l'indépendance ne peut être assurée que par le protocole expérimentale.
- (P2), (P3), (P4) : on ne peut pas formellement montrer qu'elles sont vérifiées par des tests statistiques car $\varepsilon_1, \dots, \varepsilon_n$ ne sont pas observées. On utilise des outils graphiques sur $\widehat{e}_1, \dots, \widehat{e}_n$ pour les vérifier.

Avec R : 4 graphiques à analyser

```
par(mfrow = c(2, 2))
plot(reg)
```



Residuals vs Fitted

Ce qu'on regarde: les résidus observés $\widehat{e}_k$ en fonction des valeurs ajustées $\widehat{y}_k$.

Pourquoi ? Pour vérifier (P2) et (P3).

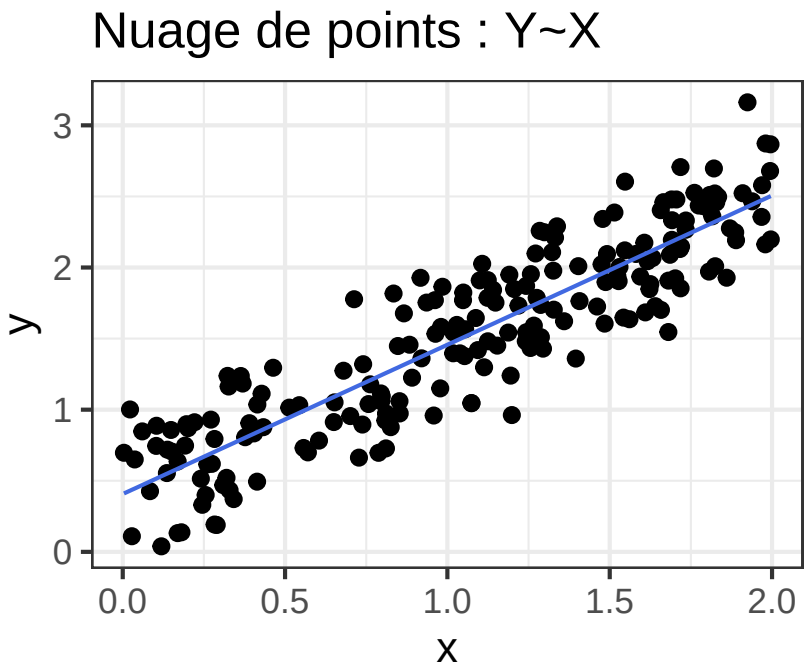
1. Par construction, les $\widehat{e}_k$ sont de moyenne nulle donc on ne peut pas mettre en défaut (P2) en moyennant les résidus estimés.
2. Si les 4 hypothèses sont bien respectées alors $\widehat{\varepsilon}$ et $\widehat{\mathbf{Y}}$ sont indépendantes.

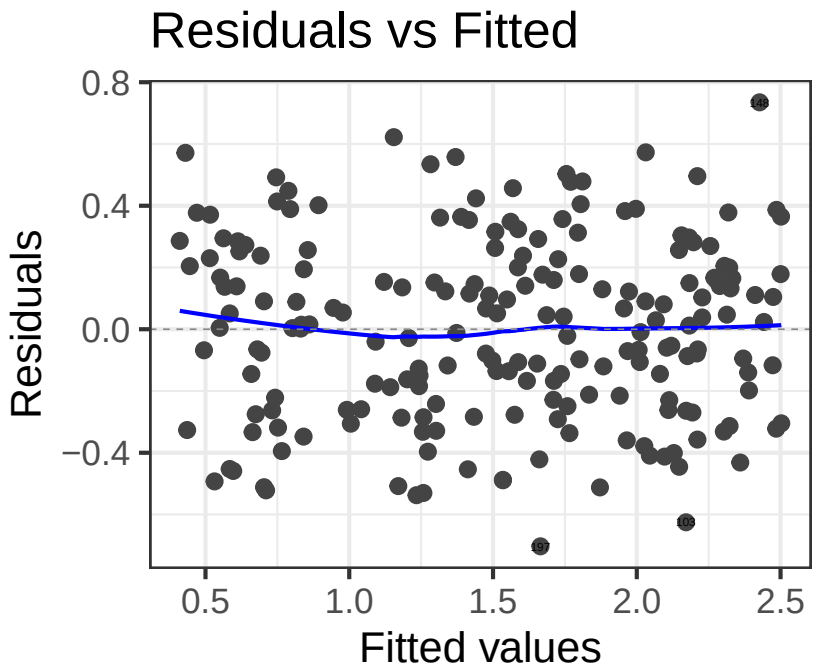
Interprétation:

- (P2) est vérifiée si on observe un nuage de points centré et aligné sans structure particulière.
- (P3) est vérifiée si les points forment une bande horizontale.

Residuals vs Fitted : exemples d'interprétation

Exemple n°1:



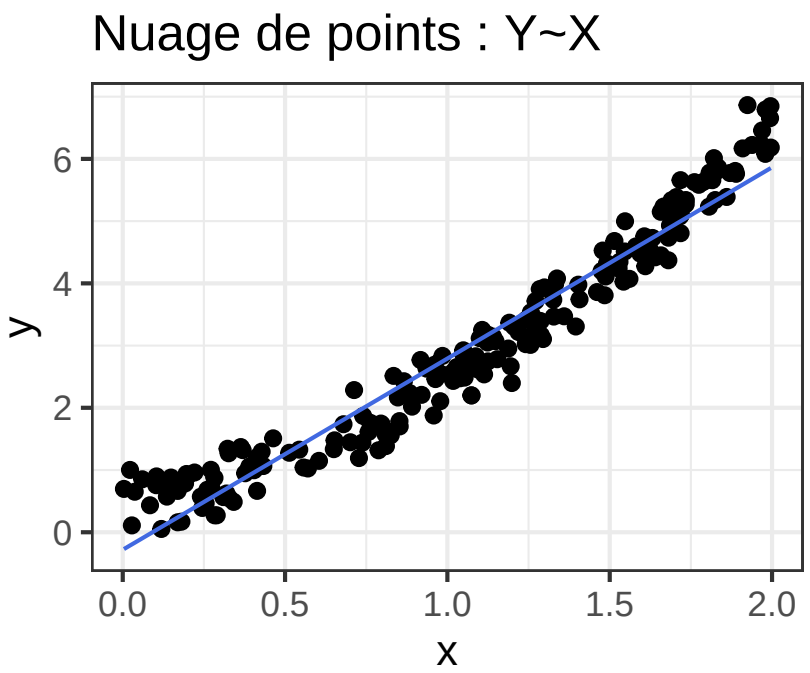


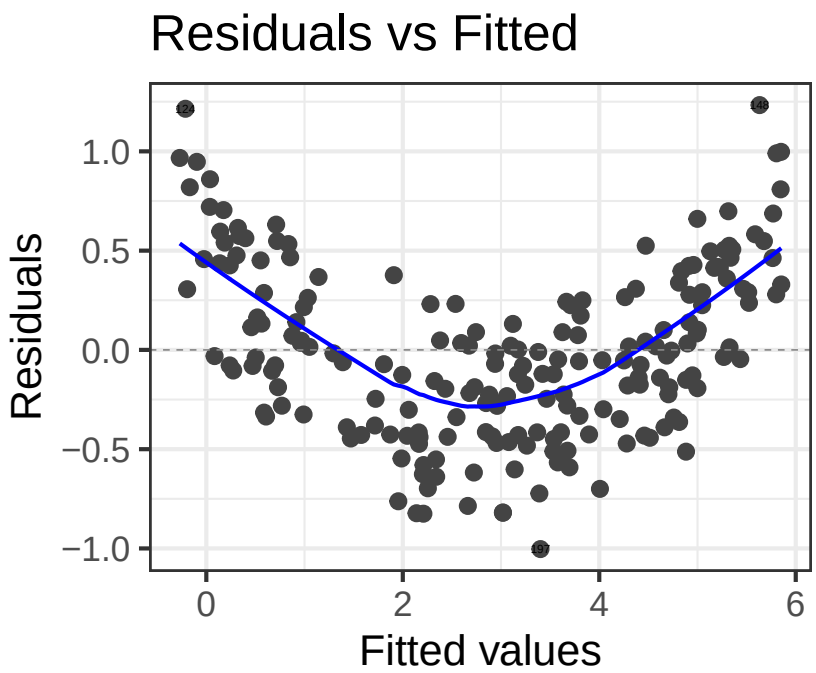
Ce qu'on voit: la valeur des résidus ne semble pas dépendre de la valeur des valeurs ajustées (il ne sont donc pas structurés en fonction de la valeur ajustée). Ils sont globalement identiquement distribués autour de 0.

Ce qu'on conclut: on valide (P2) et (P3).

Residuals vs Fitted : exemples d'interprétation

Exemple n°2:



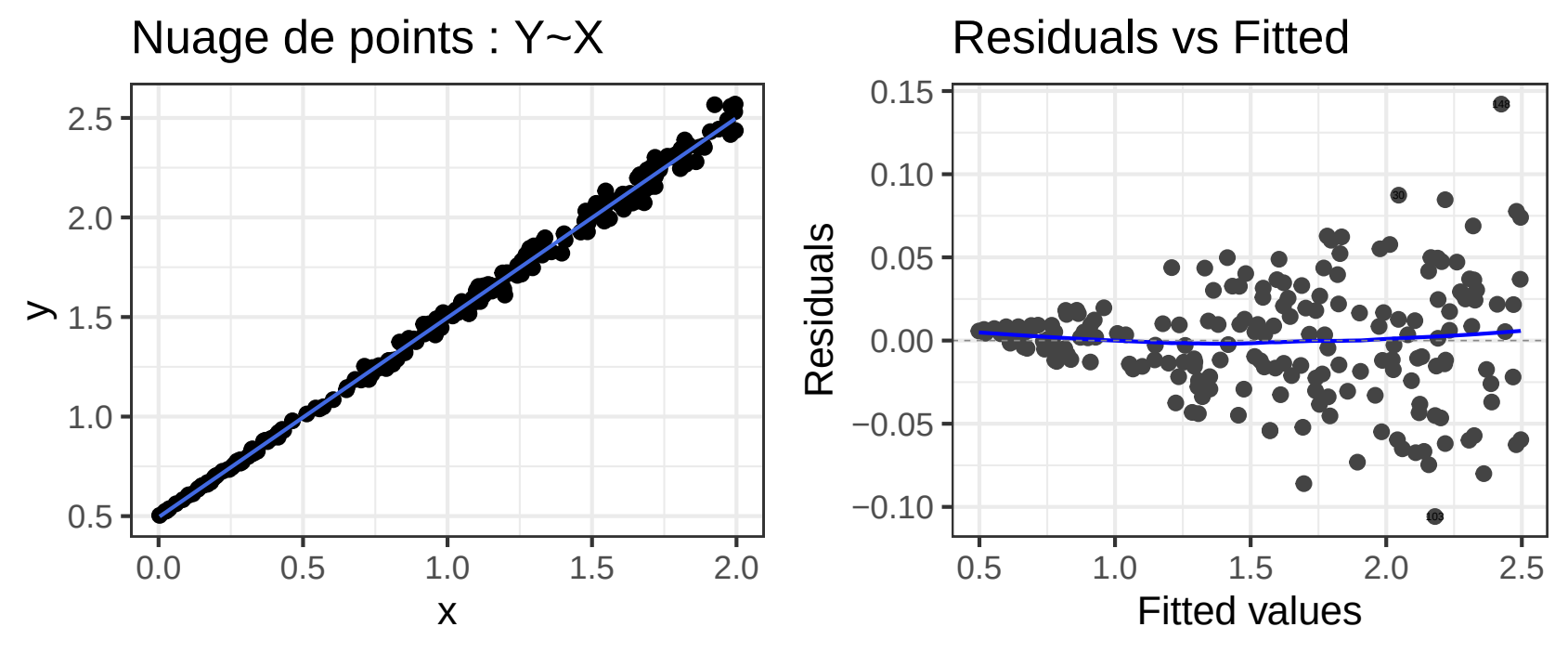


Ce qu'on voit: la valeur des résidus ne semble pas indépendante de la valeur des valeurs ajustées. On observe une structure (croissance des résidus lorsque les valeurs ajustées sont dans un intervalle).

Ce qu'on conclut: modèle non adapté aux données. Il faut proposer un autre modèle de régression ou transformer les données explicatives **X**.

Residuals vs Fitted : exemples d'interprétation

Exemple n°3:



Ce qu'on voit: la dispersion des résidus augmente lorsque les valeurs ajustées croissent.

Ce qu'on conclut: la variance des résidus n'est pas constante (hétéroscédasticité). (P3) n'est pas validée. Une solution à envisager est de transformer la variable à expliquer $\mathbf{Y}$.

Scale–Location

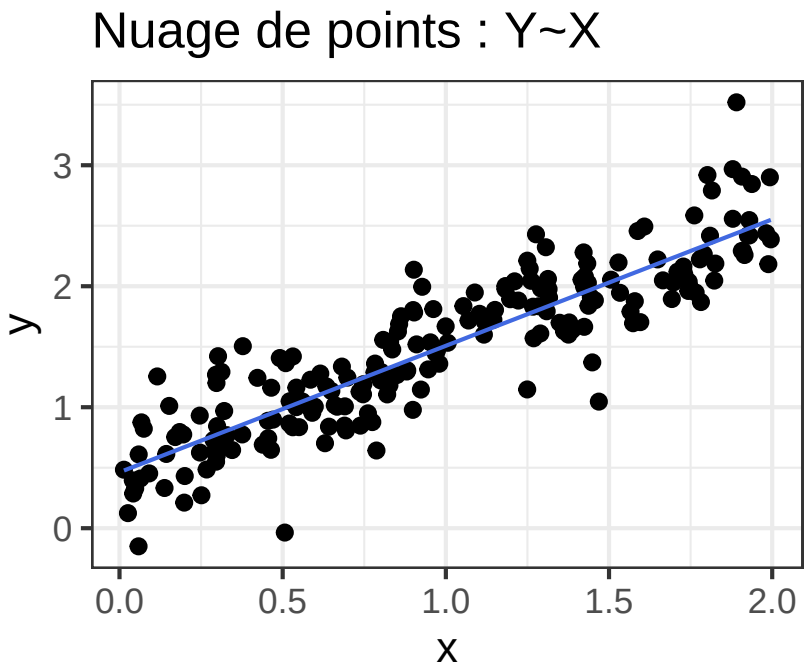
Ce qu'on regarde: la racine carrée de la valeur absolue des résidus standardisés $\widehat{r}_k$ ($\widehat{r}_k$ correspond à $\widehat{e}_k$ divisé par l'écart type estimé de $\widehat{e}_1, \ldots, \widehat{e}_n$) observés en fonction des valeurs ajustées $\hat{y}_k$.

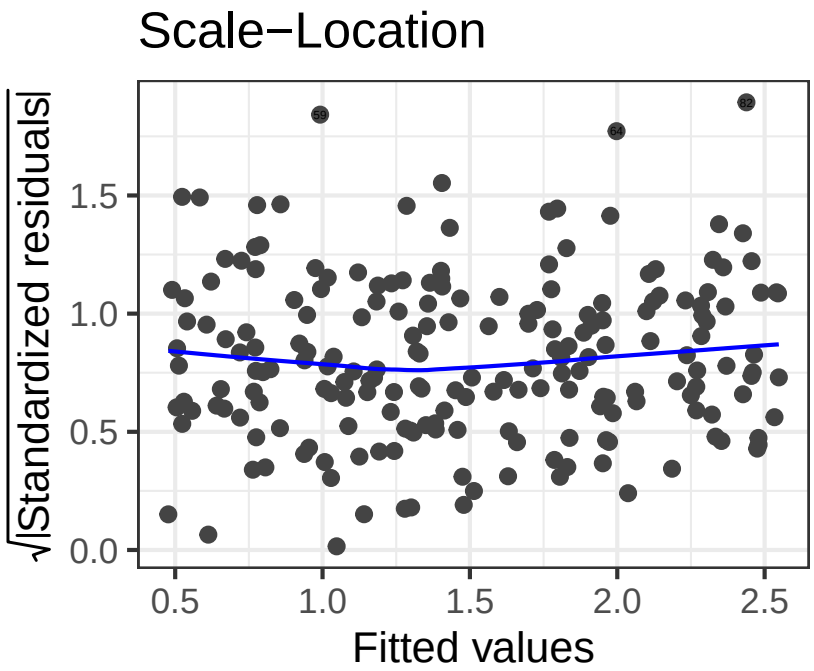
Pourquoi ? C'est un autre graphique permettant de vérifier (P3).

Interprétation: (P3) est vérifiée si on observe un nuage de points centré et aligné sans structure particulière.

Scale–Location : exemples d'interprétation

Exemple n°1:



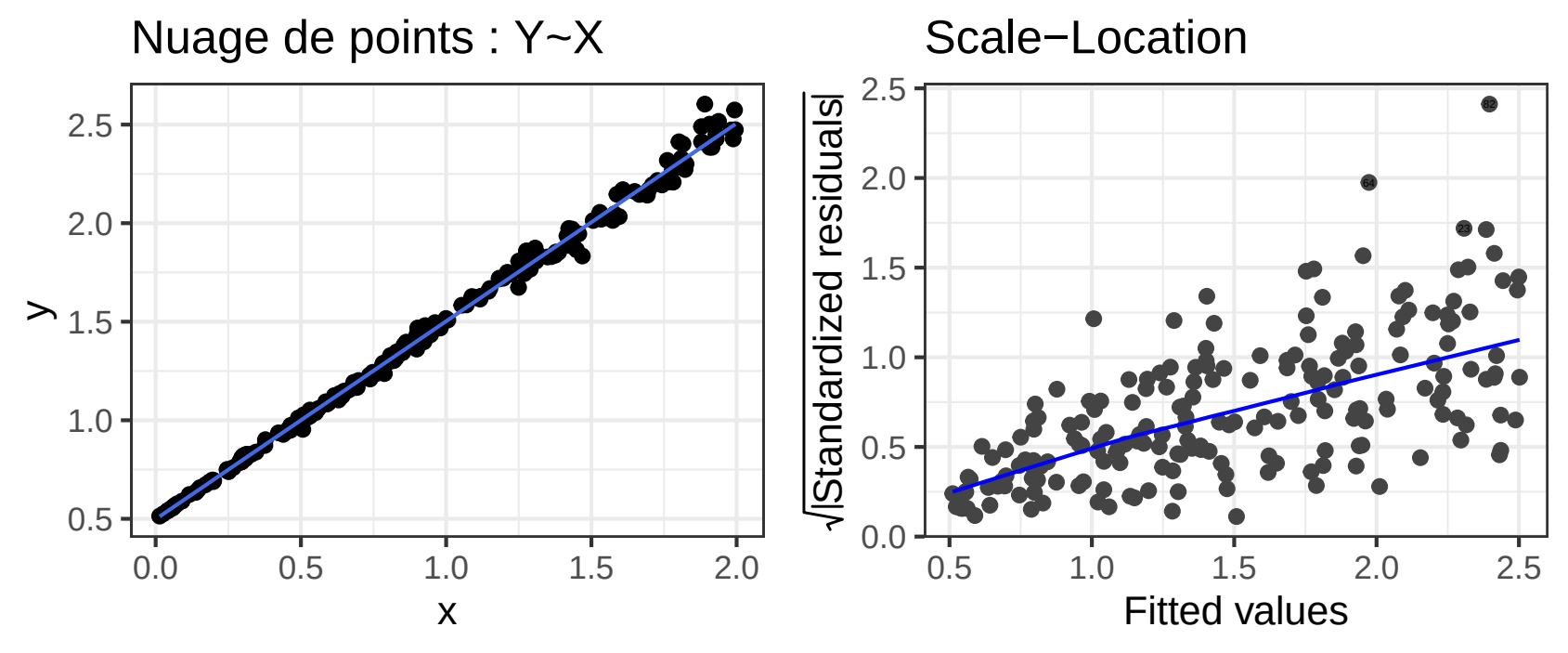


Ce qu'on voit: la valeur des résidus standardisés ne semble pas dépendre de la valeur des valeurs ajustées (il ne sont donc pas structurés en fonction de la valeur ajustée).

Ce qu'on conclut: on valide (P3).

Scale–Location : exemples d'interprétation

Exemple n°3:



Ce qu'on voit: les résidus standardisés, ainsi que leur dispersion, augmentent lorsque les valeurs ajustées croissent.

Ce qu'on conclut: la variance des résidus n'est pas constante (hétéroscédasticité). (P3) n'est pas validée. Une solution à envisager est de transformer la variable à expliquer $\mathbf{Y}$.

Normal Q–Q

Ce qu'on regarde: la valeur des quantiles empiriques des résidus standardisés $\widehat{r}_k$ en fonction de la valeur quantiles théoriques d'une loi normale $\mathcal{N}(0, 1)$.

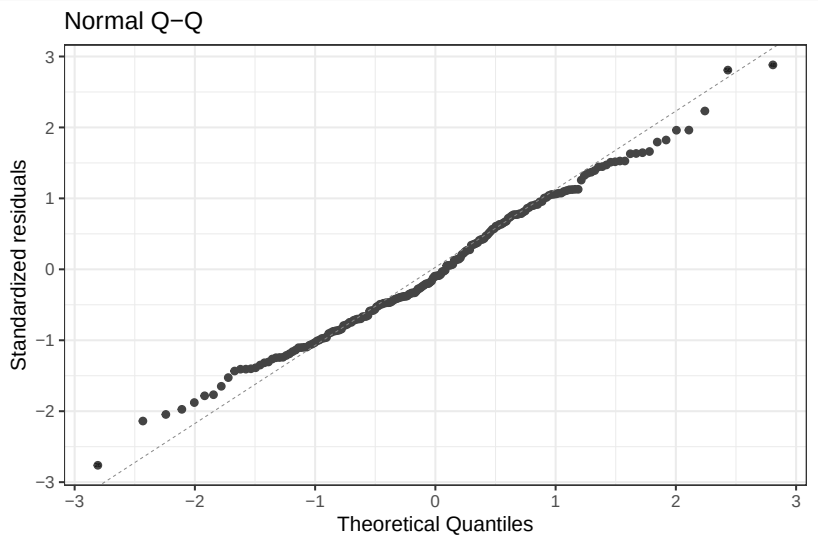
Pourquoi ? Pour valider l'hypothèse de distribution normale des résidus (P4).

Interprétation: un diagramme quantile-quantile permet de tester graphiquement l'adéquation d'un échantillon observé à une loi théorique. Si les points sont globalement alignés sur la droite $y = x$ alors les quantiles empiriques sont à peu près égaux aux quantiles théoriques. Les résidus suivent une loi normale et (P4) est validé.

Remarque: cette hypothèse n'est pas la plus importante à vérifier si le nombre d'observations n est grand. Dans ce cas, on peut obtenir des propriétés asymptotiques de nos estimateurs $\rightsquigarrow$ modèle linéaire généralisé.

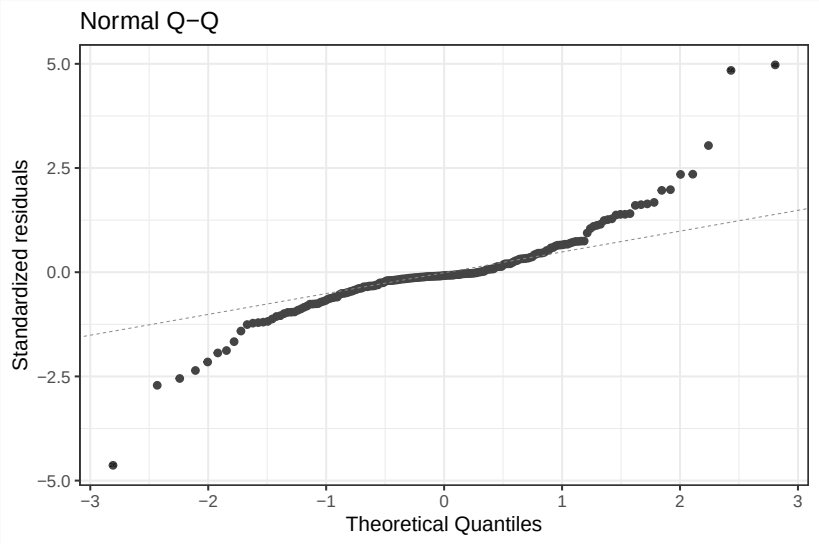
Normal Q–Q : exemples d'interprétation

Exemple n°1:



Conclusion: les résidus sont distribués suivant une loi normale. (P4) est validée.

Exemple n°2:



Conclusion: les résidus ne sont pas distribués suivant une loi normale. (P4) n'est pas validée.

***Leverage*:** on peut écrire

$$\widehat{y} = \mathbf{H}y, \quad \text{avec} \quad \mathbf{H} = (h_{ij}) = \mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal, \quad \rightsquigarrow \quad \widehat{y}_k = \sum_{j=1}^{n} h_{kj} y_j.$$

Le levier de l'observation y_j pour la valeur $\widehat{y}_k$ est h_{kj}.

Ce qu'on regarde: la valeur des résidus standardisés $\widehat{r}_k$ en fonction de h_{kk} (poids d'une observation dans l'estimation de sa prédiction) .

Pourquoi ? Pour vérifier que notre échantillon ne contient pas de points aberrants, c'est-à-dire des observations y_k ayant une grande influence sur l'estimation.

Interprétation: une observation y_k est atypique par rapport aux autres données lorsque le résidu standardisé associé $\widehat{r}_k$ est « grand ». Si de plus :

- h_{kk} est « petit », elle est atypique mais peu influente sur l'estimation des paramètres. Il n'y a pas de problèmes.
- h_{kk} est « grand », elle influe beaucoup sur l'estimation. Cela peut être un problème et on peut décider de refaire l'analyse en enlevant ce point.

Comment mesure-t-on l'effet conjoint de $\widehat{r}_k$ et h_{kk} ? On utilise la distance de Cook, définie pour chaque observation y_k par

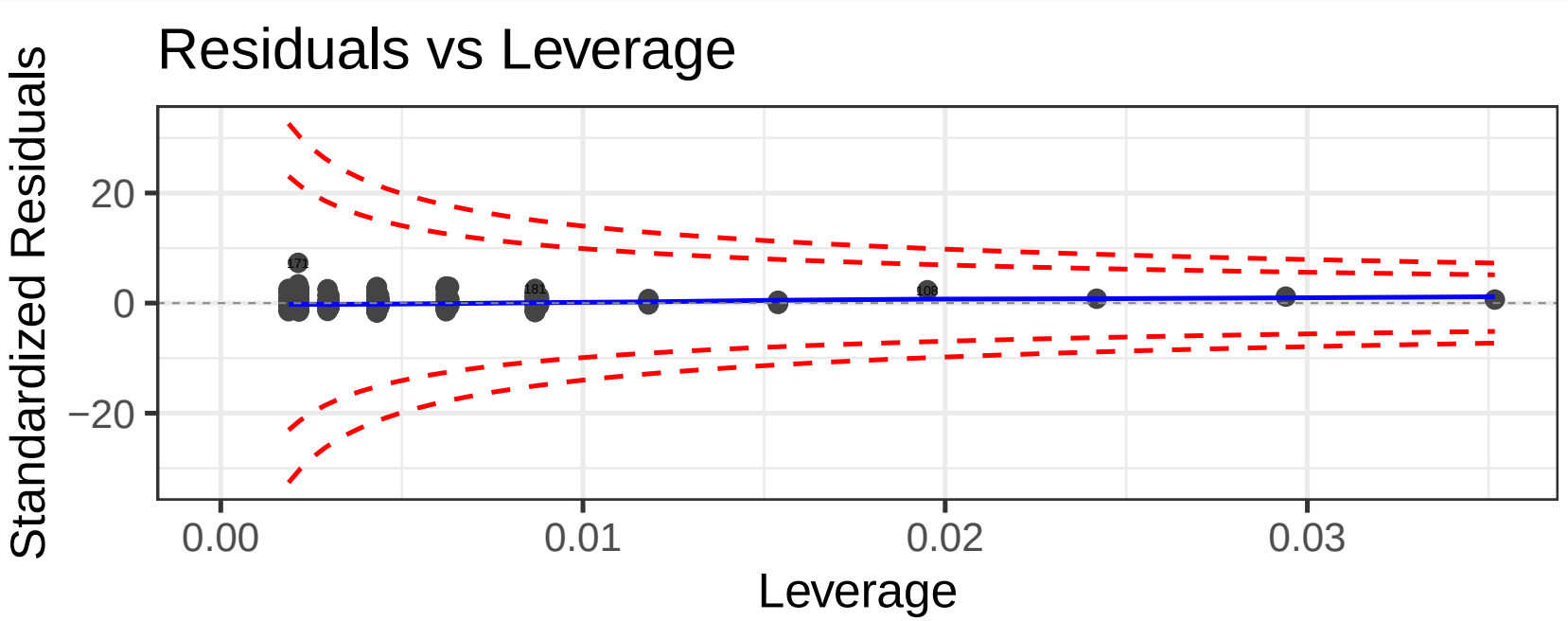
$$d_k = \frac{h_{kk}\widehat{r}_k^2}{(1+p)(1-h_{kk})^2},$$

où p est la dimension de β, *i.e.*, nombre de paramètres de régression estimés (ici $p = 2$).

En pratique: on considère que le point k n'est pas aberrant si sa distance de Cook est inférieure à 1, *i.e.*, le résidu standardisé $\widehat{r}_k$ vérifie

$$-\sqrt{\frac{(1+p)(1-h_{kk})^2}{h_{kk}}} \leq \widehat{r}_k \leq \sqrt{\frac{(1+p)(1-h_{kk})^2}{h_{kk}}}.$$

Par défaut, R trace la bande correspondant à une distance de Cook de 1 et de 0.5. Les points qui sont dans la bande ne sont pas problématiques. On regarde avec soin les autres.

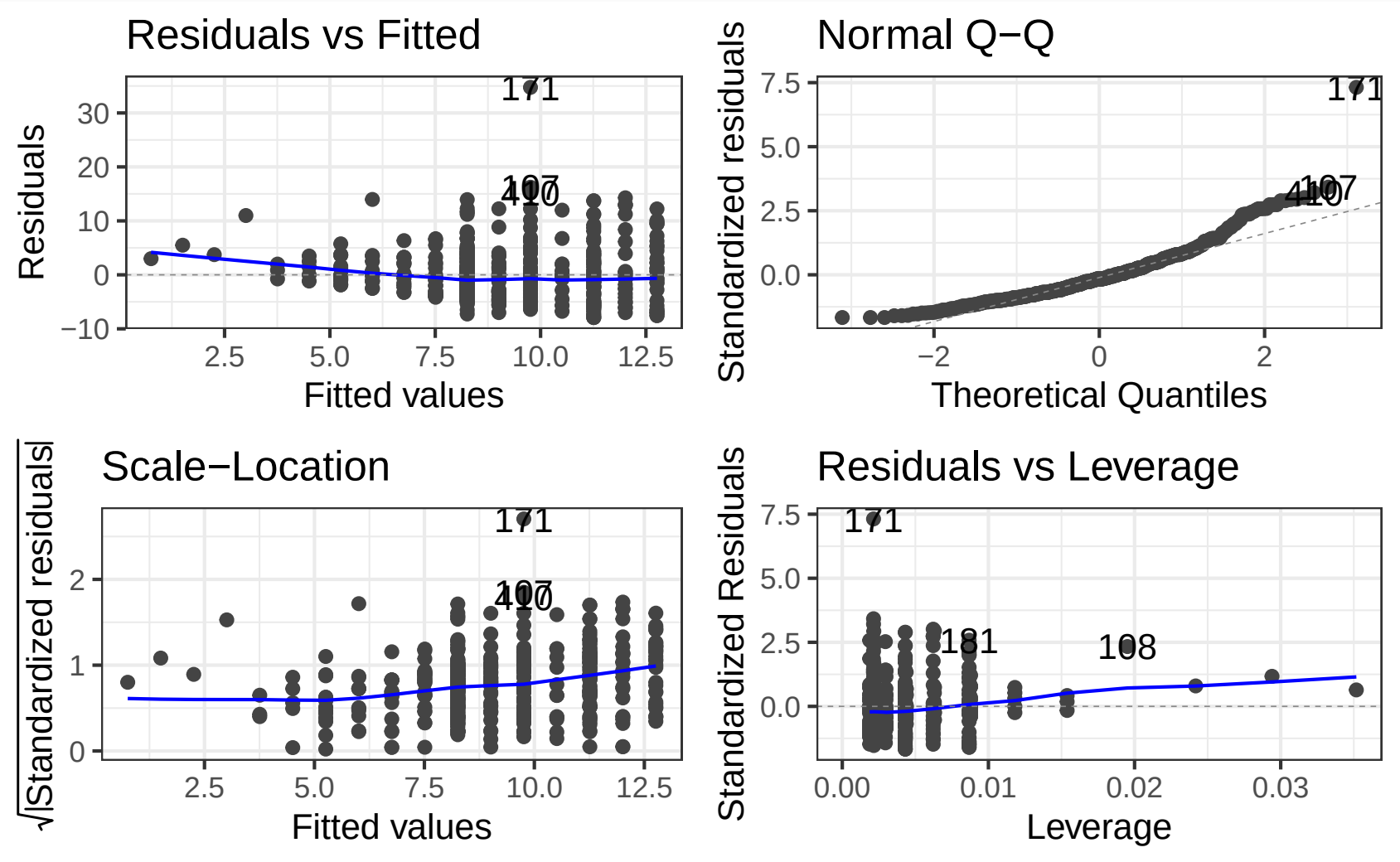


Fin de l'analyse graphique

Que faire si les hypothèses ne sont pas vérifiées ?

- Si le problème semble venir de (P2) on peut essayer de transformer la variable explicative, choisir une autre variable explicative ou un modèle plus complexe.
- S'il y a hétéroscédasticité des résidus, on peut essayer de transformer la variable à expliquer.
- On peut éliminer les éventuels points aberrants qui nuisent à la qualité de l'estimation.

Avec R : analyses de nos 4 graphiques

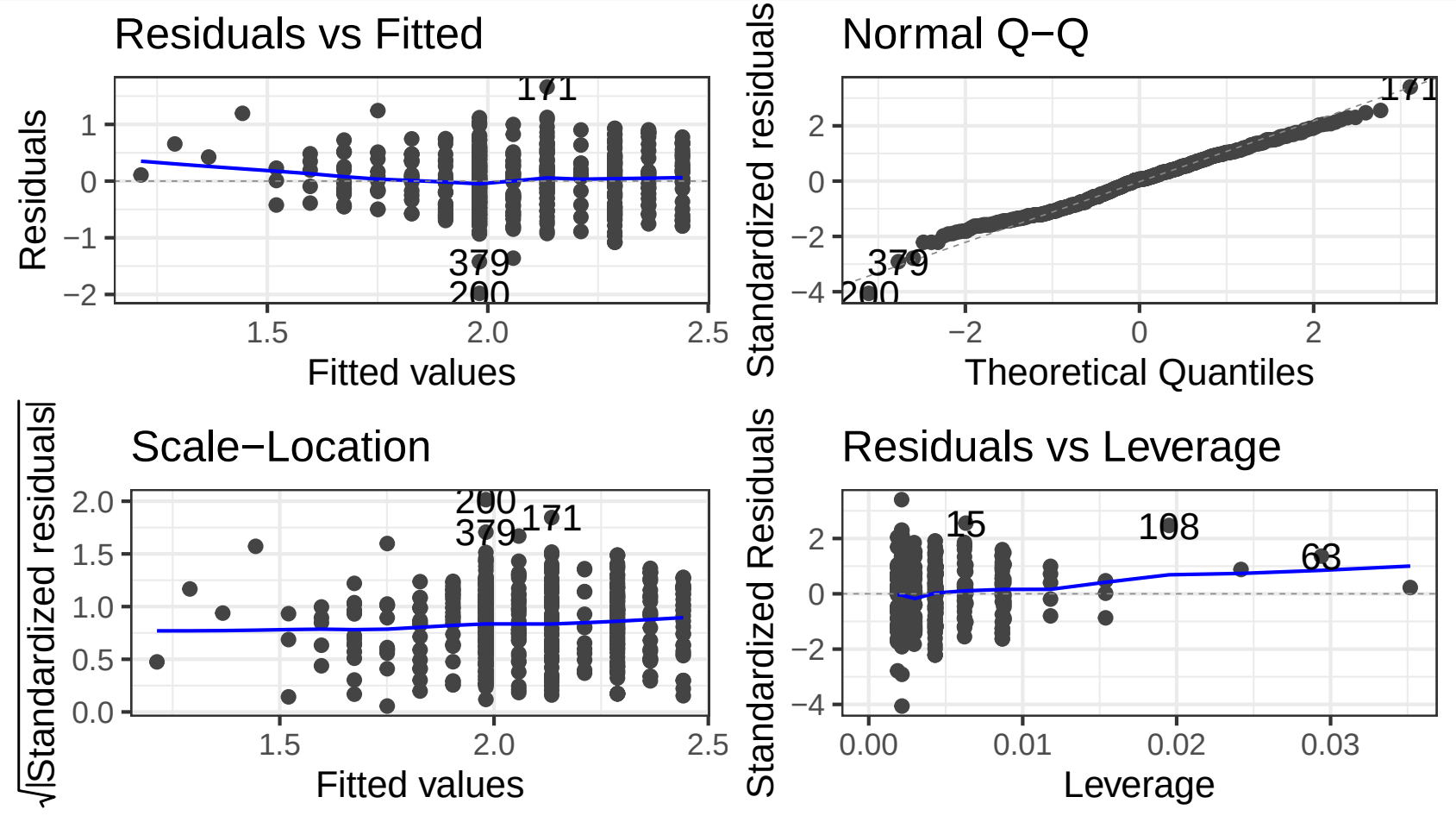


On observe une hétéroscédasticité des résidus. Il n'y a pas de points aberrants dans nos données.

Transformation de la variable réponse

Variable réponse: on considère le log du salaire (*i.e.*, $Y = \log(\texttt{WAGE})$) afin de stabiliser la variance.

```
reg <- lm(log(WAGE) ~ EDUCATION, data = donnees)
```



Transformation de la variable réponse : nouveaux résultats

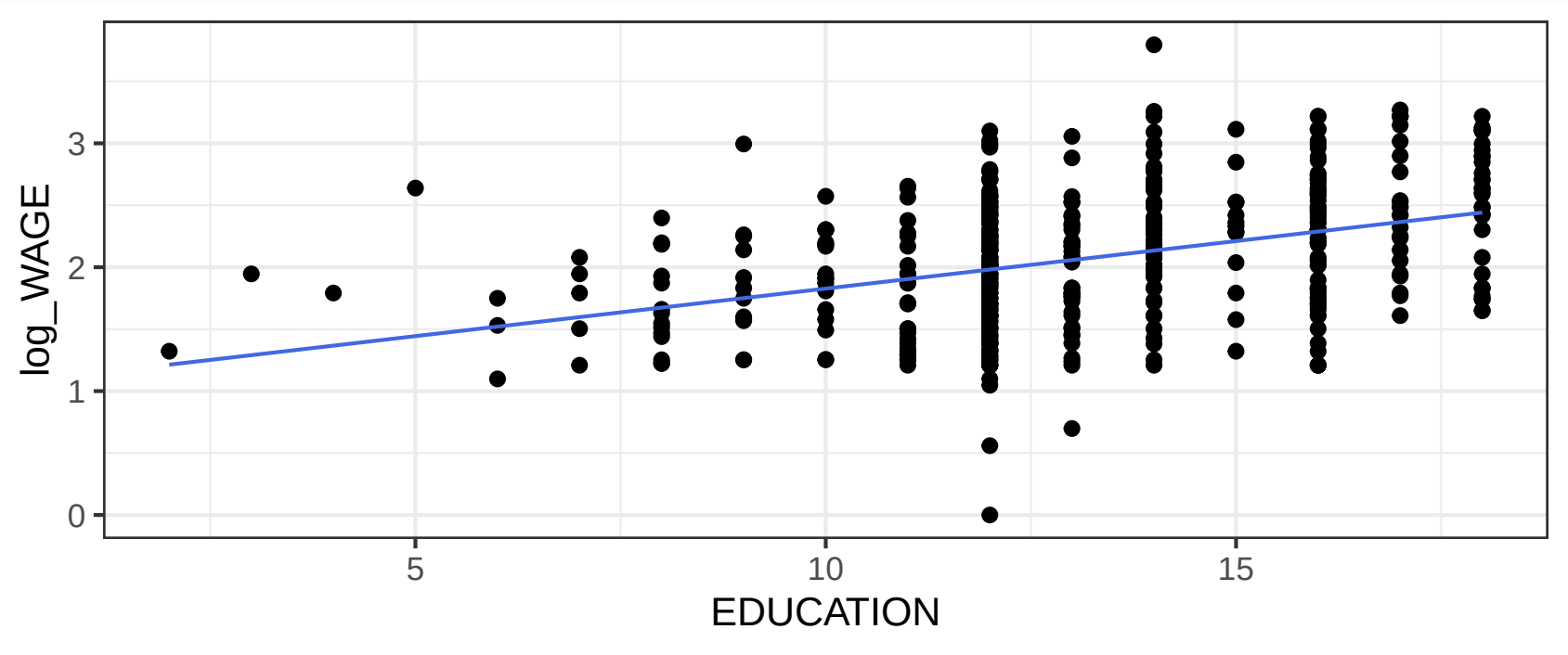
Estimation de a, b et σ^2:

```
coef(reg)
```

```
(Intercept)   EDUCATION
   1.059890    0.076759
```

```
summary(reg)$sigma^2
```

```
[1] 0.23866
```



Intervalles de confiance sur les paramètres

Intervalles de confiance

Motivation: une valeur ponctuelle est (généralement) insuffisante car elle ne tient pas compte de la variabilité de l'estimateur. On construit donc un intervalle qui mesure l'incertitude quant à la valeur estimée sous les hypothèses du modèle choisi.

Définition: soient $I_1(\mathbf{Y})$ et $I_2(\mathbf{Y})$ deux variables aléatoires construites à partir d'un vecteur aléatoire $\mathbf{y}$ dont la loi dépend d'un paramètre inconnu θ. On dit que $[I_1(\mathbf{Y}), I_2(\mathbf{Y})]$ est un intervalle de confiance de niveau $1-\alpha$, $\alpha \in [0,1]$, pour θ lorsque

$$\mathbb{P}[\theta \in [I_1(\mathbf{Y}), I_2(\mathbf{Y})]] = 1-\alpha.$$

Pour une réalisation $\mathbf{y}$ de $\mathbf{Y}$, l'intervalle de confiance (parfois appelé fourchette de confiance) de niveau $1-\alpha$ pour θ est $[I_1(\mathbf{y}), I_2(\mathbf{y})]$.

Intervalles de confiance pour a et b

Notations: $\mathcal{T}(n-2)$ désigne la loi de Student à $n-2$ degrés de liberté et $q_{1-\alpha/2}^{\mathcal{T}(n-2)}$ le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{T}(n-2)$.

Pour le paramètre a, un intervalle de confiance bilatère symétrique de niveau de $1-\alpha$ est

$$\left[\widehat{A} - q_{1-\alpha/2}^{\mathcal{T}(n-2)}\widehat{\sigma}_A\,, \widehat{A} + q_{1-\alpha/2}^{\mathcal{T}(n-2)}\widehat{\sigma}_A\right], \quad \text{avec} \quad \widehat{\sigma}_A = \sqrt{\frac{S^2}{\sum_{k=1}^n (x_k - \overline{x})^2}}.$$

Pour le paramètre b, un intervalle de confiance bilatère symétrique de niveau de $1-\alpha$ est

$$\left[\widehat{B} - q_{1-\alpha/2}^{\mathcal{T}(n-2)}\widehat{\sigma}_B\,, \widehat{B} + q_{1-\alpha/2}^{\mathcal{T}(n-2)}\widehat{\sigma}_B\right] \quad \text{avec} \quad \widehat{\sigma}_B = \sqrt{\frac{S^2 \sum_{k=1}^n x_k^2}{n\sum_{k=1}^n (x_k - \overline{x})^2}}.$$

Avec R : intervalles de confiance pour a et b

Par défaut: $1 - \alpha = 95\%$. On a les intervalles de confiance bilatères symétriques de niveau de 95% :

```
confint(reg)
```

```
                2.5 %    97.5 %
(Intercept) 0.848846 1.270934
EDUCATION   0.060865 0.092652
```

Spécifier $1-\alpha$: on utilise l'argument `level`. Par exemple pour des intervalles de confiance bilatères symétriques de niveau de 97% :

```
confint(reg, level = 0.97)
```

```
                1.5 %    98.5 %
(Intercept) 0.826125 1.293655
EDUCATION   0.059154 0.094364
```

Intervalles de confiance pour σ^2

Notations: $q_{\alpha/2}^{\chi^2(n-2)}$ et $q_{1-\alpha/2}^{\chi^2(n-2)}$ désignent les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi $\chi^2(n-2)$.

Pour le paramètre σ^2, un intervalle de confiance bilatère de niveau de $1-\alpha$ est

$$\left[\frac{(n-2)S^2}{q_{1-\alpha/2}^{\chi^2(n-2)}}, \frac{(n-2)S^2}{q_{\alpha/2}^{\chi^2(n-2)}}\right].$$

Avec R: il faut le faire à la main...

```
alpha <- 0.05
n <- nrow(donnees)
S2 <- summary(reg)$sigma^2
(n - 2) * S2/(qchisq(c(1 - alpha/2, alpha/2), n - 2))
```

```
[1] 0.21239 0.27015
```

Test d'hypothèse : pertinence de la variable explicative

Test statistique : nullité de a

Question: la valeur estimée de a permet-elle de conclure que la variable explicative x (EDUCATION) a une influence sur la variable réponse y ($\log$(WAGE)) ?

Mathématiquement, on souhaite tester l'hypothèse

$$\mathcal{H}_0 \ : \ a = 0 \quad \text{contre} \quad \mathcal{H}_1 \ : \ a \neq 0.$$

On va réaliser un test statistique pour choisir entre ces deux hypothèses.

Remarque:

- $\mathcal{H}_0$ représente le modèle simple (pas d'influence de x).
- $\mathcal{H}_1$ représente le modèle le modèle le plus général (influence de x).

Test statistique : nullité de a

Comment ça marche en général ?

Un test est une fonction des observations qui vaut 0 si on choisit $\mathcal{H}_0$ et 1 si on choisit $\mathcal{H}_1$. Il peut s'écrire sous la forme

$$\phi(\mathbf{Y}) = \mathbb{1}_{\{T(\mathbf{Y})\in\mathcal{R}\}},$$

où T est une fonction appelée statistique de test (elle correspond généralement à une fonction de l'estimateur du paramètre testé) et $\mathcal{R}$ est appelée la zone de rejet du test.

Comment ça marche pour la régression linéaire simple ?

Pour tester $a = 0$, on peut uniquement comparer $\widehat{a}$ à 0. Intuitivement, on va rejeter $\mathcal{H}_0$ dès que $\widehat{a}$ s'écarte « trop » de 0, *i.e.*, lorsque $|\widehat{a}| > s$ où s est un seuil qui tient compte de la variabilité de l'estimateur $\widehat{A}$ et de l'erreur dans la prise de décision que l'on tolère.

Construction de la statistique de test et de la zone de rejet

Comment détermine-t-on s ? On part de l'estimateur $\widehat{A}$ (dont la loi dépend de a et σ^2 inconnues) puis on le transforme de sorte à obtenir une transformation dont la loi ne dépend plus des paramètres inconnus (loi pivotale) :

$$\frac{\widehat{A} - a}{\widehat{\sigma}_A} \sim \mathcal{T}(n-2).$$

Si $\mathcal{H}_0$ est vraie, on a en particulier

$$\frac{\widehat{A}}{\widehat{\sigma}_A} \sim \mathcal{T}(n-2) \quad \rightsquigarrow \quad T(\mathbf{Y}) = \frac{\widehat{A}}{\widehat{\sigma}_A}.$$

On choisit s de sorte que, pour $\alpha \in]0\,,1[$,

$$\mathbb{P}_{\mathcal{H}_0}[|T(\mathbf{Y})| > s] = \alpha \quad \Rightarrow \quad s = q_{1-\alpha/2}^{\mathcal{T}(n-2)}.$$

α représente le risque de première espèce, c'est à dire la probabilité de rejeter $\mathcal{H}_0$ à tort (erreur de décision). Elle est contrôlée par l'utilisateur (usuellement $\alpha = 5\%$).

Test statistique : en pratique

La zone de rejet

$$\mathcal{R} = \left\{ |T(\mathbf{Y})| > q_{1-\alpha/2}^{\mathcal{T}(n-2)} \right\}$$

fournit un test de niveau α de l'hypothèse $\mathcal{H}_0 : a = 0$ contre $\mathcal{H}_1 : a \neq 0$.

p-valeur: les logiciels ne demandent pas de spécifier le niveau souhaité du test. Ils fournissent à la place une p-valeur. C'est le plus petit niveau pour lequel on rejette $\mathcal{H}_0$. Pour toute erreur de première espèce α supérieure à la p-valeur on rejetterait $\mathcal{H}_0$. Pour t_{obs} la valeur observée de $T(\mathbf{Y})$

$$\text{p-valeur} = 2\mathbb{P}[T(\mathbf{Y}) > |t_{obs}|].$$

Interprétation:

- si la p-valeur est inférieure à 5% on rejette $\mathcal{H}_0$ au niveau 5%. On conclut que la variable explicative à une influence sur la variable réponse.
- si la p-valeur est supérieure à 5%, on ne peut rejeter $\mathcal{H}_0$ au niveau 5% (*e.g.*, pas de lien entre les variables ou variabilité d'estimation trop importante).

Avec R : test de nullité de α

```
summary(reg)


Call:
lm(formula = log_WAGE ~ EDUCATION, data = donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9810 -0.3716  0.0339  0.3497  1.6610

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.05989    0.10743    9.87   <2e-16 ***
EDUCATION    0.07676    0.00809    9.49   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.489 on 532 degrees of freedom
Multiple R-squared:  0.145,Adjusted R-squared:  0.143
F-statistic:   90 on 1 and 532 DF,  p-value: <2e-16
```

Conclusion: la p-valeur du test $\alpha = 0$ est inférieure à $2e{-}16$. Donc on rejette $\mathcal{H}_0$ pour tout niveau usuel (en particulier 5%). Le nombre d'années d'études a bien une influence sur le salaire.

Critique du modèle

Décomposition de la variance:

$$\underbrace{\sum_{k=1}^{n}(y_k - \overline{y})^2}_{\text{SCT}} = \underbrace{\sum_{k=1}^{n}(\widehat{y}_k - \overline{y})^2}_{\text{SCM}} + \underbrace{\sum_{k=1}^{n}(y_k - \widehat{y}_k)^2}_{\text{SCR}}$$

- SCT correspond à la variabilité des données.
- SCM correspond la variabilité expliquée par le modèle (*i.e.*, par la variable explicative).
- SCR correspond à la variabilité résiduelle, *i.e.*, la variabilité non-expliquée par le modèle.

Plus SCM est proche de SCT, plus le modèle explique la variabilité des observations.

Coefficient de détermination

Coefficient de détermination: part de variance expliquée par le modèle

$$R^2 = \frac{SCM}{SCT} = 1 - \frac{SCR}{SCT}.$$

Avec R: c'est la sortie `Multiple R-squared` obtenue avec `summary(reg)`. On peut y accéder directement avec

```
summary(reg)$r.squared
```

```
[1] 0.1447
```

Conclusion: le modèle prédit donc mal les données.

Remarque: ce coefficient n'est pas toujours pertinent. Nous reviendrons sur ce point dans le Chapitre 2.

Coefficient de détermination ajusté

Coefficient de détermination ajusté: part de variance expliquée par le modèle prenant en compte le nombre de variables explicatives utilisées

$$R^2_{adj} = 1 - \frac{n-1}{n-2}\frac{SCR}{SCT}.$$

Avec R: c'est la sortie `Adjusted R-squared` obtenue avec `summary(reg)`. On peut y accéder directement avec

```
summary(reg)$adj.r.squared
```

```
[1] 0.1431
```

Conclusion: le modèle prédit donc mal les données.

Remarque: ce coefficient est celui utilisé en pratique. Nous reviendrons sur ce point dans le Chapitre 2.

Predicteur et prédiction

Objectif de la régression: soit x_{n+1} une nouvelle valeur de la variable explicative (nombre d'années d'étude). On souhaite prédire la valeur (inconnue) de Y (salaire horaire) notée y_{n+1}.

Prédicteur (variable aléatoire) : c'est la valeur moyenne attendue Y_k^p pour x_{n+1} sous le modèle ajusté

$$\widehat{Y}_k^p = \widehat{A}x_{n+1} + \widehat{B}.$$

Prédiction (réalisation utilisant les données) :

$$\widehat{y}_k^p = \widehat{a}x_{n+1} + \widehat{b}.$$

Remarque: la valeur x_{n+1} pour laquelle on fait la prédiction n'a pas servi pour estimer a et b.

Avec R : prévisions

Pour calculer les prévisions $(\widehat{y}^p_{n+1}, \ldots, \widehat{y}^p_{n+N})$ à partir de nouvelles valeurs $(x_{n+1}, \ldots, x_{n+N})$ pour la variable explicative :

1. On met les nouvelles valeurs dans un `data.frame` dont le nom de la colonne est le même que le nom de la variable explicative dans les données d'origine (EDUCATION)

```
x_new <- data.frame(EDUCATION = c(3, 5, 12))
```

2. On utilise la fonction `predict`

```
predict(reg, newdata = x_new)
```

```
     1      2      3
1.2902 1.4437 1.9810
```

Avec R : estimation de la variance du prédicteur

Résultat théorique: $\widehat{Y}^p_{n+1}$ est une variable aléatoire gaussienne de variance

$$\mathbb{V}\text{ar}\left[\widehat{Y}^p_{n+1}\right] = \sigma^2\left[\frac{1}{n} + \frac{(x_{n+1}-\overline{x})^2}{\sum_{k=1}^n (x_k-\overline{x})^2}\right]$$

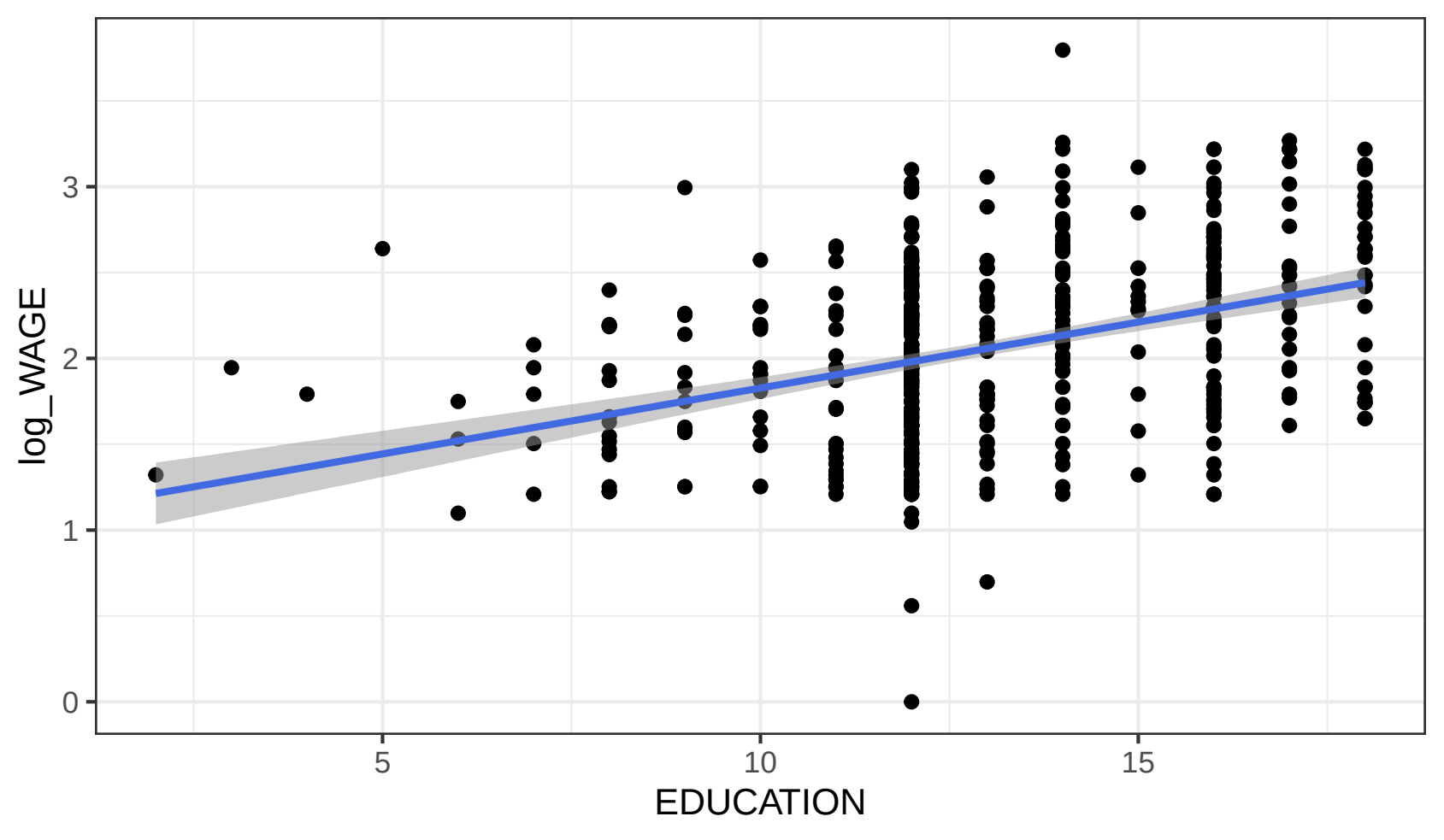
En pratique: comme σ^2 est inconnu, elle est estimée par

$$\widehat{\mathbb{V}\text{ar}}\left[\widehat{Y}^p_{n+1}\right] = \widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x_{n+1}-\overline{x})^2}{\sum_{k=1}^n (x_k-\overline{x})^2}\right]$$

Avec R: l'argument `se.fit` de la fonction `predict` fournit l'écart type associée (*i.e.*, racine carrée de cette variance = *standard error*).

```
predict(reg, newdata = x_new, se.fit = TRUE)
```

Avec R : visualisation de la variance du prédicteur



Erreur de prévision

Erreur de prévision: erreur commise entre la valeur y_{n+1} (inconnue) à prévoir et celle qu'on prédit :

$$\widehat{\varepsilon}^p_{n+1} = y_{n+1} - \widehat{Y}^p_{n+1}.$$

Elle quantifie la capacité du modèle à prévoir (elle tient compte de l'aléa estimé).

Résultat théorique: $\widehat{\varepsilon}^p_{n+1}$ est une variable aléatoire qui vérifie

$$\mathbb{E}\left[\widehat{\varepsilon}^p_{n+1}\right] = 0 \quad \text{et} \quad \mathbb{V}\text{ar}\left[\widehat{\varepsilon}^p_{n+1}\right] = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_{k=1}^n (x_k - \overline{x})^2}\right].$$

Remarque: la variance est d'autant plus grande que x_{n+1} est loin de $\overline{x}$.

Avec R : intervalle de prédiction

Pour y_{n+1}: un intervalle de confiance bilatère symétrique de niveau de $1-\alpha$ est

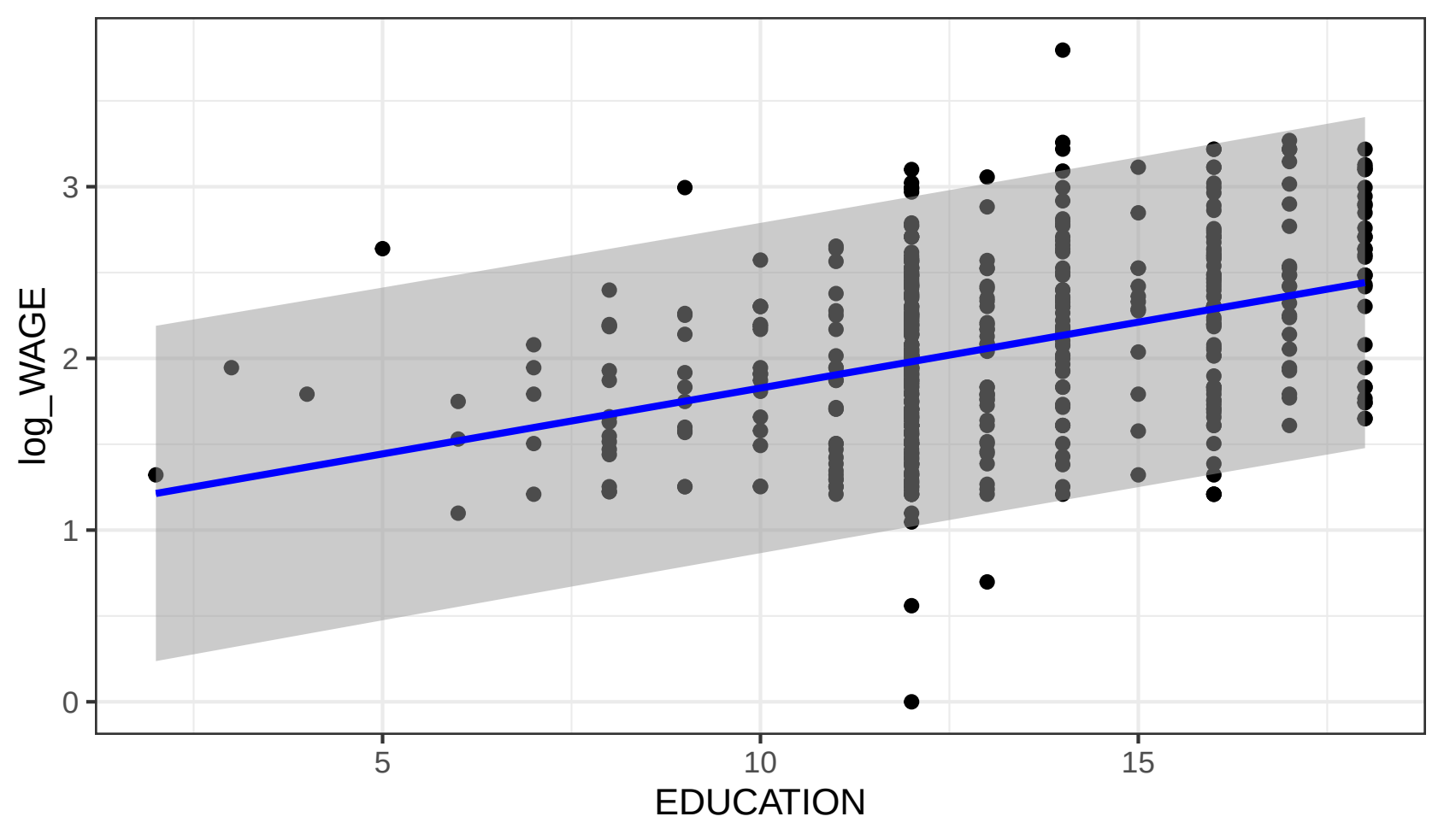
$$\left[\widehat{Y}^p_{n+1} \pm q^{\mathcal{T}(n-2)}_{1-\alpha/2} \sqrt{\widehat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_{k=1}^n (x_k - \overline{x})^2}\right]}\right].$$

Avec R: on utilise les arguments `interval` et `level` (pour $1-\alpha$) de la fonction `predict`

```
predict(reg, newdata = x_new, interval = "prediction",
        level = .95)
```

```
     fit     lwr    upr
1 1.2902 0.31648 2.2639
2 1.4437 0.47469 2.4127
3 1.9810 1.02028 2.9417
```

Avec R : visualisation de l'intervalle de prédiction



Conclusion

Ma feuille de route pour la régression linéaire simple

1. Charger les données, vérifier que les variables sont bien de la nature attendue (quantitatives).
2. Exploration des données et calcul de statistiques descriptives.
3. Tracé du nuage de points (x_k, y_k).
4. Écrire le modèle linéaire. Appliquer la fonction `lm` aux données pour ajuster le modèle.
5. Analyser les graphes de résidus pour valider ou invalider les hypothèses du modèles. Le cas échéant, modifier le modèle pour obtenir des résidus satisfaisants (transformation log de x ou y).
6. Faire le test de nullité de a. Si on ne rejette pas l'hypothèse $\mathcal{H}_0 : a = 0$, on arrête là ! Le modèle linéaire n'est pas adapté.
7. Sinon critiquer le modèle, faire de la prédicition, conclure.