Chapitre 2

La régression linéaire simple

2.1 Le modèle et les postulats

Le modèle de régression linéaire simple constitue un modèle dans lequel on cherche à expliquer une variable endogène Y en fonction d'une seule variable exogène x. Il s'exprime comme suit :

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, ..., n. \tag{2.1}$$

où $Y_1, \ldots, Y_n$ sont n observations de la variable endogène (variable réponse) ;
$x_1, \ldots, x_n$ sont n observations de la variable exogène (variable explicative) ;
β_0 est le paramètre associé à l'ordonnée à l'origine de la droite ;
β_1 est le paramètre associé à la pente de la droite ;
ε est un terme d'erreur ;
et n est le nombre d'observations.

Remarque 2.1. *Il est important de comprendre quelles grandeurs sont supposées constantes ou aléatoires dans le modèle* (2.1)*, à savoir*

- $Y_i, \ldots, Y_n$ *sont les observations d'une variable aléatoire ;*
- $x_1, \ldots, x_n$ *sont considérées comme des valeurs connues et par conséquent non aléatoires ;*
- β_0 *et* β_1*, les paramètres du modèle, associés respectivement à l'ordonnée à l'origine et à la pente de la droite, sont inconnus et il faudra par conséquent les estimer ;*
- $\varepsilon_1, \ldots, \varepsilon_n$ *sont les réalisations inconnues d'une variable aléatoire.*

Pour pouvoir utiliser la régression linéaire, il est nécessaire de formuler certains **postulats**. Les trois principales hypothèses concernent :

1. La linéarité de la relation entre la variable endogène et la variable exogène.
2. L'homoscédasticité (variance constante) des observations de la variable endogène.

3. L'indépendance des observations de la variable endogène.

Comme on le verra à la section 2.10, on vérifiera généralement les postulats à partir d'une estimation de la distribution des termes d'erreur, plutôt qu'à partir des variables elles-mêmes. Pour cette raison, il est possible de reformuler les postulats de la façon suivante :

$\mathcal{H}_1$. $\mathbb{E}[\varepsilon_i] = 0$ pour $1 \leq i \leq n$

$\mathcal{H}_2$. $\mathbb{V}\text{ar}[\varepsilon_i] = \sigma^2$ pour $1 \leq i \leq n$

$\mathcal{H}_3$. $\mathbb{C}\text{ov}(\varepsilon_i, \varepsilon_j) = 0$ pour $i \neq j$

À ces trois postulats, on doit parfois ajouter l'hypothèse de normalité des résidus suivante :

$$\mathcal{H}_4.\ \varepsilon_1, \cdots, \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$$

Il s'agit d'un postulat plus fort généralement utilisé parce qu'il permet de construire des intervalles de confiance et de faire des tests d'hypothèses. On verra que celui-ci est nécessaire par exemple à la mise en œuvre de la méthode du maximum de vraisemblance.

2.2 Estimation de β_0, β_1 et σ^2

2.2.1 Méhode des moindres carrés

On attribue généralement la méthode des moindres carrés au mathématicien français Adrien-Marie Legendre (1752-1833). Ses travaux ont été publiés en 1805. Plusieurs autres mathématiciens de l'époque ont aussi publié des travaux sur le sujet, notamment Carl Friedrich Gauss (1777-1855) et Robert Adrain (1775-1843). Ce dernier était considéré comme le plus brillant mathématicien américain de son époque. Il est surtout connu pour sa formulation de la méthode des moindres carrés, publiée en 1808. R. Adrain ne connaissait probablement pas le travail de C.F. Gauss sur les moindres carrés, mais il est possible qu'il ait eu accès à l'article de A.M. Legendre sur le sujet (publié en 1805). Pour plus de détails sur la biographie de R. Adrain il est possible de consulter l'adresse suivante `http://www-history.mcs.st-andrews.ac.uk/history/Mathematicians/Adrain.html`.

On cherche ici à choisir $\hat{\beta}_0$ et $\hat{\beta}_1$ de façon à minimiser la somme des résidus au carré :

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \Big\{ Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \Big\}^2 \tag{2.2}$$

où $\hat{\varepsilon}_i$ est une estimation du terme d'erreur, aussi appelé résidu, Y_i est la valeur observée de la variable réponse et $\hat{Y}_i$ est la prévision (la position de Y_i s'il était sur la droite). Puisque la fonction à minimiser possède de bonnes propriétés (entre autres lisse et convexe), elle peut être minimisée en prenant les dérivées

de la somme par rapport à $\hat{\beta}_0$ et $\hat{\beta}_1$, puis en posant ces dérivées égales à zéro et finalement en résolvant le système de deux équations à deux inconnues suivant :

$$\left\{ \begin{array}{lcl} \frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 & = & \sum_{i=1}^n 2(-1)(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\ \frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 & = & \sum_{i=1}^n -2x_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \end{array} \right. \tag{2.3}$$

2.2.2 Méthode du maximum de vraisemblance

Si l'on suppose que les observations sont indépendantes, la vraisemblance s'exprime comme suit :

$$L(\beta_0, \beta_1) = \prod_{i=1}^n f(Y_i) \tag{2.4}$$

La mise en œuvre de la méthode du maximum de vraisemblance nécessite la connaissance de la distribution des observations $Y_1, \ldots, Y_n$. On a que $Y_i = (\beta_0 + \beta_1 x_i) + \varepsilon_i$. Étant donné le quatrième postulat de la régression ($\mathcal{H}_4 : \varepsilon_1, ..., \varepsilon_n \tilde{} N(0, \sigma^2)$) et sachant que $(\beta_0 + \beta_1 x_i)$ est une constante, on a que $Y_i \tilde{} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$. Donc,

$$\begin{array}{rcl} L(\beta_0, \beta_1) & = & \prod_{i=1}^n f(Y_i) \\ & = & \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \left(\frac{Y_i - \beta_0 - \beta_1 x_i}{\sigma} \right)^2 \right\}. \end{array}$$

Il reste ensuite à maximiser la fonction $L(\beta_0, \beta_1)$ ou, plus simplement, son logarithme $l(\beta_0, \beta_1)$.

$$\begin{array}{rcl} l(\beta_0, \beta_1) & = & \sum_{i=1}^n \left\{ -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (Y_i - \beta_0 - \beta_1 x_i)^2 \right\} \\ & = & -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2 \end{array} \tag{2.5}$$

On trouve $\hat{\beta}_0$, $\hat{\beta}_1$ et $\hat{\sigma}^2$ qui minimisent $l(\beta_0, \beta_1, \sigma^2)$ en dérivant respectivement par rapport à β_0, β_1 et σ^2 puis en posant la dérivée égale à zéro :

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \tag{2.6}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i Y_i - \bar{Y} \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i} = \frac{S_{xy}}{S_{xx}} \tag{2.7}$$

$$\tag{2.8}$$

où

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2, \tag{2.9}$$

et

$$S_{xy} = \sum_{i=1}^{n} Y_i(x_i - \bar{x}) = \sum_{i=1}^{n} x_i Y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} Y_i\right). \tag{2.10}$$

Il est à noter que S_{xx} et S_{xy} sont appelés respectivement la somme des carrés corrigée de x et la somme des produits croisés corrigée de x et y. Finalement,

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon_i}^2. \end{aligned} \tag{2.11}$$

Toutefois, on peut montrer que ce dernier estimateur est biaisé. On préfère utiliser la version non biaisée :

$$\begin{aligned} s^2 &= \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ &= \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon_i}^2. \end{aligned} \tag{2.12}$$

2.3 Propriétés des estimateurs

On peut montrer que, sous le quatrième postulat (normalité des résidus) :

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right),$$

et

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right).$$

2.4 Intervalles de confiance et tests d'hypothèse pour β_0 et β_1

Afin de construire un intervalle de confiance pour β_0 et β_1 avec la méthode des pivots, la distribution de ces paramètres est requise. Par exemple, pour un intervalle de confiance pour β_1, on centre et réduit la

variable aléatoire pour obtenir

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim \mathcal{N}(0,1).$$

Si σ^2 était connu, un intervalle de confiance de niveau (1-κ)% serait donné par $[\hat{\beta}_1 \pm z_{\kappa/2}\sigma/\sqrt{S_{xx}}]$. Or, σ^2 n'est généralement pas connu et on doit par conséquent l'estimer. L'estimateur du maximum de vraisemblance de cette quantité est $\sum_{i=1}^n \hat{\varepsilon}_i^2/n$ mais ce dernier est biaisé. On utilisera donc plutôt l'estimateur non biaisé défini par $\hat{\sigma}^2 = s^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2/(n-2)$. Dans ce cas toutefois, $\frac{\hat{\beta}_1-\beta_1}{\sqrt{s^2/S_{xx}}}$ ne suit pas une distribution normale. En se rappelant les notions vues en statistique mathématique, on trouve néanmoins facilement cette distribution.

$$\frac{\frac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{S_{xx}}}}{\sqrt{\frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{\sigma^2(n-2)}}} = \frac{U}{\sqrt{V/\nu}}$$

où $U = \frac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{S_{xx}}}$ est distribué selon une loi normale centrée réduite et $V = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{\sigma^2}$ suit une loi du chi-carré à $\nu = (n-2)$ degrés de liberté (comme somme du carré de variables aléatoires normales). On obtient donc que

$$\frac{\frac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{S_{xx}}}}{\sqrt{\frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{\sigma^2(n-2)}}} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{n-2}/(n-2)}} \sim \mathcal{T}_{n-2}$$

Comme il s'agit d'un rapport entre une variable aléatoire normale centrée réduite et la racine carrée d'une variable chi-carré divisée par son degré de liberté $(n-2)$, on obtient la distribution d'une loi $\mathcal{T}$ de Student à $(n-2)$ degrés de liberté ($\mathcal{T}_{n-2}$). De cette dernière on déduit directement les bornes d'un intervalle de confiance de niveau (1-κ)% pour β_1 avec l'équation

$$\left[\hat{\beta}_1 \pm t_{n-2}(\kappa/2)\left(\frac{s}{\sqrt{S_{xx}}}\right)\right], \tag{2.13}$$

où $t_{n-2}(\kappa/2)$ dénote le $(\kappa/2)$-ième quantile supérieur d'une loi $\mathcal{T}$ à $(n-2)$ degrés de liberté, i.e.

$$P[\mathcal{T}_{n-2} > t_{n-2}(\kappa/2)] = \kappa/2.$$

On peut également se baser sur la distribution trouvée pour mettre en œuvre des tests d'hypothèse. L'hypothèse la plus souvent testée concerne la nullité de la pente, à savoir $\mathcal{H}_0 : \beta_1 = 0$. Sous $\mathcal{H}_0$,

$$\frac{\hat{\beta}_1}{s/\sqrt{S_{xx}}} \sim \mathcal{T}_{n-2}$$

et $\mathbb{P}(|t_{n-2}| > |t_{\text{obs}}|)$, où t_{obs} désigne la statistique du test calculée, nous livre la p-valeur du test. Si la p-valeur est plus petite que le seuil κ choisi, on rejette l'hypothèse nulle que $\beta_1 = 0$. Au contraire, si la p-valeur est supérieure au seuil, on ne peut pas rejeter l'hypothèse nulle que $\beta_1 = 0$. Dans ce cas, on en déduit que x ne possède aucun pouvoir explicatif sur Y.

Il est possible de construire de la même manière un intervalle de confiance et un test d'hypothèse pour β_0. L'intervalle de confiance pour l'ordonnée à l'origine obtenu s'exprime comme :

$$\left[\hat{\beta}_0 \pm t_{n-2}(\kappa/2)s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right]. \tag{2.14}$$

Dans le cas où l'hypothèse nulle que l'ordonnée à l'origine $\beta_0 = 0$ ne peut être rejetée, le modèle peut se simplifier de la manière suivante

$$Y_i = \beta_1 x_i + \varepsilon_i.$$

2.5 Prévisions

Voir notes en classe

2.6 Analyse de la variance

L'analyse de la variance est une méthode statistique en soi, mais une petite introduction s'impose pour bien comprendre la régression. On va voir dans cette section comment la variabilité dans les observations de la variable endogène $Y_1, \ldots, Y_n$ peut se décomposer sous le modèle de régression linéaire. Cette décomposition constitue la base des tests d'ajustements de modèle.

On a que

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\
&= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(Y_i - \bar{Y}) \\
&= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2.
\end{aligned}$$

Si on pose la somme des carrés totaux $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$, la somme des carrés de la régression $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ et la somme des carrés résiduelle $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, alors on a que

$$SST = SSR + SSE. \tag{2.15}$$

On comprendra le pourquoi de cette nomenclature pour les diverses sommes des carrés en examinant le tableau 2.1 qui résume l'analyse de la variance.

L'équation (2.15) est prépondérante en régression. La somme des carrés totaux SST quantifie la variabilité dans les Y_i. L'équation (2.15) signifie que cette variabilité se décompose en deux parties : la somme des carrés de la régression (SSR) et la somme des carrés résiduelle (SSE).

La somme des carrés de la régression SSR quantifie la variabilité dans les prévisions $\hat{Y}_i$. Étant donné que les $\hat{Y}_i, i = 1, \ldots, n$ ne varient qu'avec les $x_i, i = 1 \ldots, n$, on a que cette somme de carrés est la partie de la variabilité dans les Y_i expliquée par le fait que toutes les observations n'ont pas la même valeur pour x_i.

La somme des carrés résiduelle SSE mesure la variabilité dans les $(Y_i - \hat{Y}_i)$. Cette variabilité est induite par le fait que la valeur de Y_i n'est pas expliquée parfaitement par le modèle de régression.

Le tableau 2.1 présente les principaux éléments de l'analyse de la variance.

TABLE 2.1: Tableau d'analyse de la variance dans le cas de la régression linéaire simple présentant la source, le nombre de degrés de liberté, la somme des carrés, le carré moyen et la statistique F de Fisher.

Source	Degrés de liberté	Somme des carrés	Carrés moyens	F
Modèle	$ddl_1 = p = 1$	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	$MSR = SSR/ddl_1$	MSR/MSE
Erreur Résiduelle	$ddl_2 = n - 2$	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$MSE = SSE/ddl_2$	
Totale	$n - 1 = n - 2 + p$	$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$		

Remarque 2.2. *Dans le tableau 2.1 d'analyse de la variance, chaque somme des carrés possède un nombre de degrés de liberté qui lui est propre (indiqué dans la deuxième colonne du tableau), à savoir :*

- *SSR possède un seul degré de liberté ($ddl_1 = p = 1$) car dans le modèle de régression linéaire simple il n'existe qu'une seule variable explicative (exogène).*
- *À SSE est associé $(n-2)$ degrés de libertés car on estime deux paramètres dans le modèle de régression (β_0 et β_1), donc on a n observations moins 2 paramètres.*
- *SST possède $(n-1)$ degrés de liberté car il y a n observations moins une moyenne. Il est à souligner que le nombre de degrés de liberté de SST est égal à la somme du nombre de degrés de liberté de SSR et SSE.*

La somme des carrés de la régression (SSR) constitue la portion de la variance expliquée par le modèle. La somme des carrés résiduelle (SSE) représente la partie que le modèle n'explique pas. Intuitivement, on comprend que si la proportion de la variance expliquée par le modèle (SSR/SST) est élevée, le modèle est bon.

Remarque 2.3. *Il est à souligner que le carré moyen de l'erreur résiduelle (MSE) correspond à l'estimation de la variance des termes d'erreur. En effet*

$$MSE = \frac{SSE}{ddl_2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{n-2} = s^2.$$

L'analyse de la variance permet d'effectuer un test global de la régression. On désire tester l'hypothèse nulle $H_0 : Y_i = \beta_0 + \varepsilon_i$ contre l'hypothèse alternative $H_1 : Y_i = \beta_0 + \beta_1 x + \varepsilon_i$. L'idée consiste à comparer SSR à SSE afin d'évaluer si la variable exogène explique une partie significative de SST.

On a déjà mentionné que si la proportion de la variance expliquée par le modèle (SSR/SST) est élevée, le modèle est bon. Pour effectuer le test global, on utilisera plutôt le rapport des carrés moyens (MSR/MSE). Afin de construire un test, il faut en premier lieu trouver la distribution de ce rapport.

$$\begin{aligned}
\frac{MSR}{MSE} &= \frac{SSR/1}{s^2} \\
&= \frac{S_{xy}^2/S_{xx}}{s^2} \\
&= \frac{\hat{\beta}_1^2 S_{xx}}{s^2} \\
&= \left(\frac{\hat{\beta}_1}{s/\sqrt{S_{xx}}} \right)^2 \sim (\mathcal{T}_{n-2})^2 \sim \mathcal{F}_{1,n-2}
\end{aligned}$$

Dans la dernière égalité, nous avons utilisé le fait que si la variable aléatoire T est distribuée selon la loi $\mathcal{T}$ de Student à k degrés de liberté (notée $\mathcal{T}_k$) alors la variable aléatoire T^2 suit la loi de Fisher-Snedecor à 1 et k degrés de liberté (notée $\mathcal{F}_{1,k}$).

On va donc rejeter H_0 au seuil α si $F \geq \mathcal{F}_{1,n-2}(1-\alpha)$, où $\mathcal{F}_{1,n-2}(1-\alpha)$ est tel que $P[\mathcal{F}_{1,n-2} > \mathcal{F}_{1,n-2}(1-\alpha)] = 1-\alpha$. La p-valeur du test de Fisher-Snedecor indique si la variable exogène possède un effet significatif sur la valeur de la variable endogène. Une p-valeur faible (par exemple si la p-valeur $< 0,05$) signifie que l'effet de la variable exogène est significatif. Dans le cadre de la régression linéaire simple, le test de Fisher-Snedecor est exactement égal au test de la nullité de la pente $H_0 : \beta_1 = 0$. On verra que cette remarque ne se généralise pas au cas de la régression linéaire multiple.

2.7 Test F de Fisher pour la validité globale de la régression

Voir notes en classe

2.8 Distribution d'un résidu

Voir notes en classe

2.9 Évaluation de la qualité du modèle

Pour évaluer la qualité du modèle, on regarde les critères suivants :

1. Le test de Fisher et le test de nullité de β_1. Ces deux tests sont équivalents dans le cadre de la régression linéaire simple. Si la conclusion du test est que $\beta_1 = 0$, le modèle est inadapté, c'est-à-dire que la variable explicative (exogène) ne possède pas d'effet significatif sur la variable réponse (endogène).
2. Le coefficient de détermination (R^2).

$$R^2 = \mathbb{C}\text{orr}^2(Y, \hat{Y}) = \left(\frac{\sum_{i=1}^n (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2 \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}} \right)^2 = \frac{SSR}{SST}$$

$$R^2 \in [0, 1]$$

Quand $R^2 = 0$, toute la variabilité est due à l'erreur aléatoire et le modèle n'explique absolument rien de la valeur de Y_i. Quand $R^2 = 1$, tous les points sont alignés sur la droite de régression, c'est-à-dire que l'ajustement du modèle est parfait et que la valeur de Y_i est une fonction exacte de x_i.
Le R^2 s'interprète comme étant le pourcentage de la variance expliquée par le modèle. Une valeur élevée indique que le modèle est bon. Le seuil à partir duquel on considère que R^2 est élevé est très variable et assez subjectif. Il dépend notamment des objectifs de la régression et du domaine d'application.

2.10 Vérification des postulats

Nous avons formulé quatre postulats pour le modèle de régression linéaire simple à la section 2.1, à savoir :

$\mathcal{H}_1$. $\mathbb{E}[\varepsilon_i] = 0$ (linéarité) ;

$\mathcal{H}_2$. $\mathbb{V}\text{ar}[\varepsilon_i] = \sigma^2$ (homoscédasticité) ;

$\mathcal{H}_3$. $\mathbb{C}\text{ov}(\varepsilon_i, \varepsilon_j) = 0$ (non corrélation) ;

$\mathcal{H}_4$. $\varepsilon_1, \ldots, \varepsilon_n$ sont distribués selon une loi normale (normalité).

Une fois le modèle ajusté, il importe de vérifier ces postulats. Si ceux-ci ne sont pas respectés, la régression linéaire n'est pas un modèle approprié et il sera nécessaire par exemple de transformer les variables ou d'utiliser un autre type de modèle. Cette section présente différentes approches permettant de diagnostiquer un problème potentiel de non-respect des postulats.

2.10.1 Divers types de résidus

Comme on l'a vu ci-dessus, les postulats de la régression peuvent tous s'exprimer en fonction des termes d'erreur $\varepsilon_1, \ldots, \varepsilon_n$. Il est donc évident que la vérification de ces postulats sera en grande partie basée sur des estimations de ces termes d'erreur. Ces estimations sont appelées *résidus*. Il est possible de distinguer divers types de résidus. On présente ici deux types de résidus, les *résidus ordinaires* et les *résidus studentisés*. Dans le cadre de la régression multiple, on abordera un troisième type de résidus, les résidus "press".

Les *résidus ordinaires*, $\hat{\varepsilon}_i$: Le $i$ème résidu (ordinaire) est défini par $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$.

- Si l'hypothèse $\mathcal{H}_1$ est vérifiée alors $\mathbb{E}[\hat{\varepsilon}_i] = 0$.
- Si le postulat $\mathcal{H}_2$ est vérifié, alors $\mathbb{V}\text{ar}[\hat{\varepsilon}_i] = (1 - h_{ii})\sigma^2$ où $h_{ii} = 1/n + (x_i - \bar{x})^2/S_{xx}$.
- Si l'hypothèse $\mathcal{H}_3$ est vraie, alors $\mathbb{C}\text{ov}(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = -h_{ij}\sigma^2$, où $h_{ij} = \frac{1}{n} + (x_i - \bar{x})(x_j - \bar{x})/S_{xx}$.
- Si le postulat $\mathcal{H}_4$ est vérifié, $\hat{\varepsilon}_i \sim \mathcal{N}\{0, \sigma^2(1 - h_{ii})\}$.

Les *résidus studentisés*, r_i : Le $i$ème résidu studentisé est défini par

$$r_i = \frac{\hat{\varepsilon}_i}{s\sqrt{1 - h_{ii}}}.$$

Les résidus studentisés permettent d'éliminer la non-homogénéité de la variance des résidus. Ils sont distribués selon une loi normale centrée et réduite. Leur interprétation dans un graphique est donc beaucoup plus directe.

2.10.2 Linéarité

Pour vérifier la linéarité de la relation, on utilise habituellement trois types de graphiques :

- le diagramme en nuage de points de la variable endogène Y_i en fonction de la variable exogène x_i ;
- le diagramme en nuage de points des résidus $\hat{\varepsilon}_i$ (ou des résidus studentisés r_i) en fonction de la valeur prédite $\hat{Y}_i$;
- le diagramme en nuage de points des résidus $\hat{\varepsilon}_i$ en fonction de la variable exogène x_i.

La figure 2.1 présente six exemples de diagrammes de dispersion de la variable endogène en fonction de la variable exogène. L'allure de la relation entre les deux variables doit être linéaire afin que la relation soit appropriée. Dans ce cas, les diagrammes a) et b) présentent une relation linéaire, ce qui n'est pas le cas des quatre autres diagrammes.

La figure 2.2 illustre quatre diagrammes de dispersion des résidus en fonction de la variable endogène. Si le postulat de linéarité est raisonnable, ce graphique devrait montrer un nuage de points centré verticalement en 0. Le nuage de points devrait exhiber une allure complètement aléatoire, c'est-à-dire qu'il ne devrait exister aucune tendance discernable (par exemple résidus croissants ou décroissants en fonction de Y_i, graphique à l'allure quadratique, etc.). Dans cet exemple, seul un diagramme résulte d'une relation linéaire. Les vraies relations desquelles sont issues les données sont indiquées dans le titre des graphiques.

La figure 2.3 montre quatre diagramme de dispersion des résidus en fonction de la variable exogène. Comme dans la figure 2.2, si le postulat de linéarité est vérifié, ce graphique devrait montrer un nuage de points centré verticalement en 0 et à l'allure aléatoire. À nouveau seul un diagramme est issu d'une relation linéaire. En comparant ces diagrammes en nuages de points avec ceux de la figure 2.2, nous constatons que ces deux types de graphiques sont similaires et permettent de tirer les mêmes conclusions.

Il est très important d'examiner les graphiques 2.1 et 2.2 ou 2.1 et 2.3. Si la relation n'est pas linéaire, la régression linéaire simple ne constitue pas un modèle approprié. Il est alors possible de transformer les données pour obtenir une relation linéaire. Par exemple, dans le cas d'une relation du type de celles présentées à la figure 2.1 e) et f), on peut tenter un modèle du genre :

$$1/Y_i = \beta_0 + \beta_1(1/x_i) + \epsilon_i$$

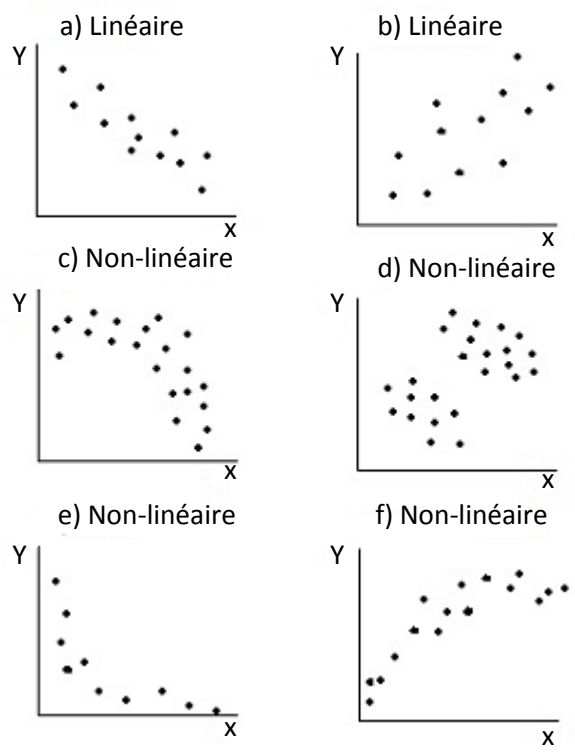

FIGURE 2.1: Diagramme de dispersion des la variable endogène en fonction de la variable exogène

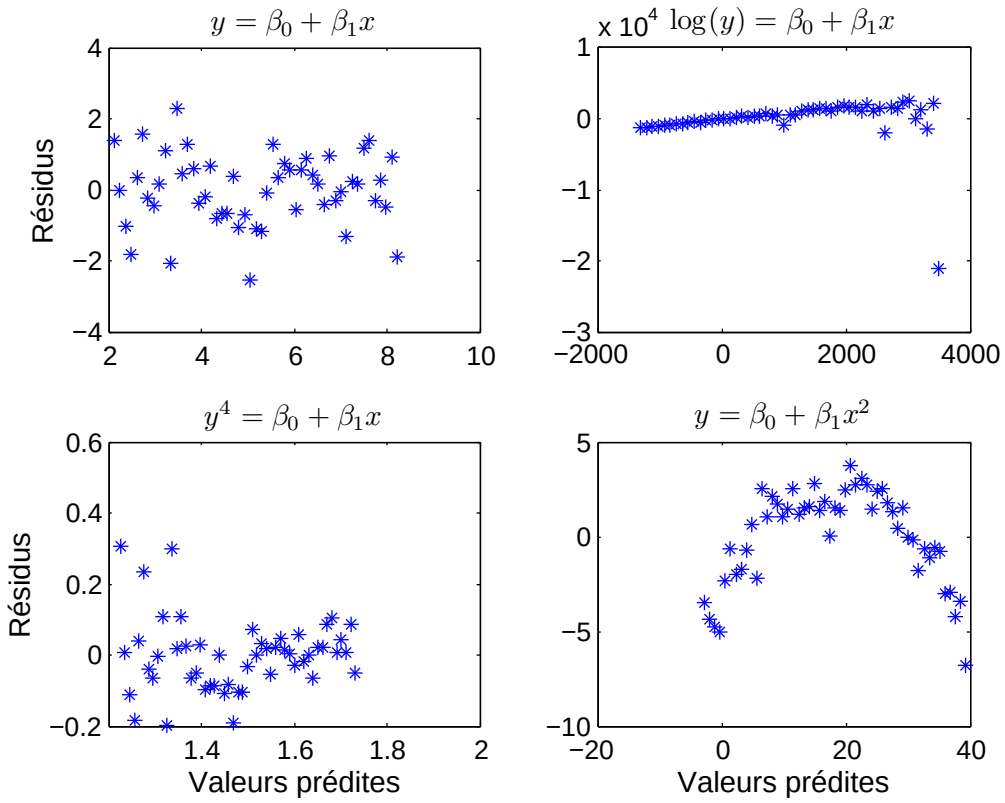

FIGURE 2.2: Exemples de quatre diagrammes de dispersion des résidus en fonction des valeurs prédites. Dans les quatre cas, le modèle de régression linéaire simple a été ajusté aux données. Le titre de chaque graphique indique le vrai modèle duquel les données sont issues.

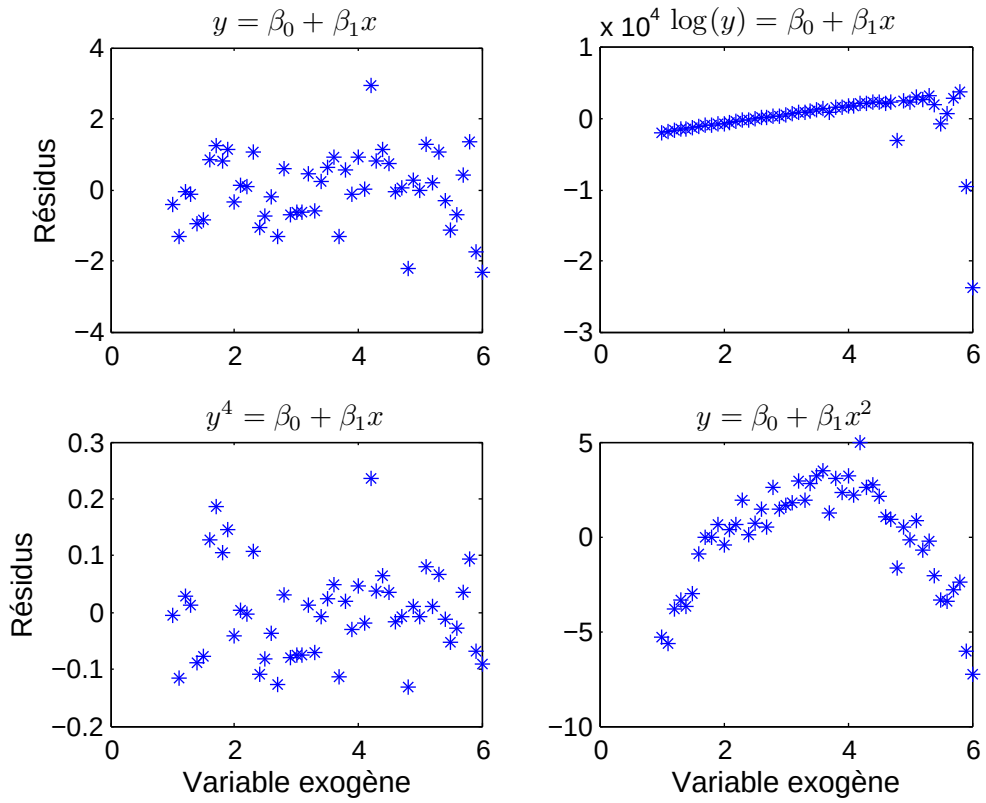

FIGURE 2.3: Exemples de quatre diagrammes de dispersion des résidus en fonction de la variable exogène. Dans les quatre cas, le modèle de régression linéaire simple a été ajusté aux données. Le titre de chaque graphique indique le vrai modèle duquel les données sont issues.

ou

$$\log(Y_i) = \beta_0 + \beta_1 \log(x_i) + \epsilon_i.$$

2.10.3 Homoscédasticité

On a vu que le graphique des résidus studentisés en fonction des valeurs prédites permet de vérifier la linéarité. Le même graphique permet de juger de l'homoscédasticité ($\mathbb{V}\text{ar}[\epsilon_i] = \sigma^2 \quad 1 \leq i \leq n$). L'hétéroscédasticité signifie que la variance des résidus n'est pas constante. Un graphique ayant une apparence d'entonnoir indique un tel problèmes. De plus, les résidus studentisés devraient en général se situer entre -3 et 3, car si le modèle de régression linéaire simple est approprié, ils sont issus d'une distribution normale centrée réduite. Si certains résidus possèdent des valeurs plus grandes que 3 en valeur absolue, ceci peut indiquer un manque de normalité ou la présence de données aberrantes. La figure 2.4 illustre de façon schématique divers problèmes d'hétéroscédasticité.

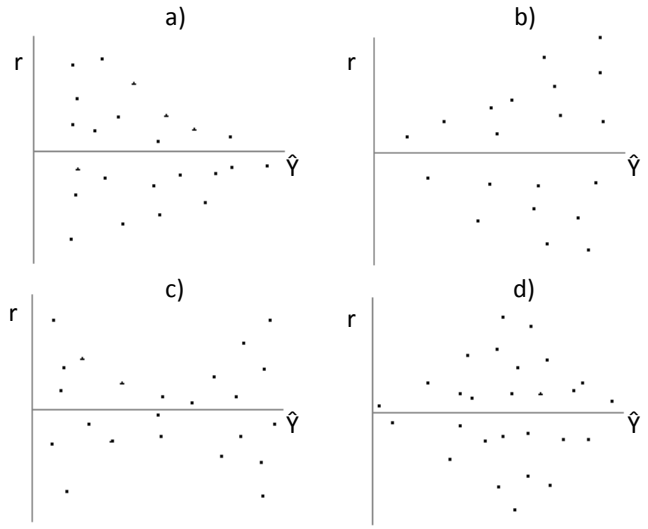

FIGURE 2.4: Exemples d'hétéroscédasticité illustrée avec le graphique des résidus studentisés.

Pour régler le problème d'hétéroscédasticité, on peut envisager une régression pondérée ou une transformation de la variable endogène.

2.10.4 Indépendance

L'indépendance des observations est difficile à tester. Si les données sont chronologiques, on peut tracer le graphique des résidus en fonction du numéro d'observation. Si aucun patron n'est détecté, le postu-

lat d'indépendance est valable. Sinon, il existe peut-être autocorrélation entre les données (dépendance temporelle). D'autres types de dépendance sont également possibles.

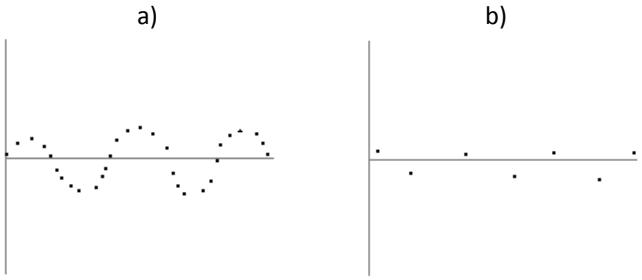

FIGURE 2.5: Exemple d'autocorrélation positive (a) et négative (b)

S'il y a présence d'autocorrélation, on peut l'éliminer avec un modèle du genre [1]

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 y_{i-1} + \epsilon_i.$$

2.10.5 Normalité

Ce postulat est moins critique que les autres. En effet, même si l'hypothèse de normalité n'est pas vérifiée, $\hat{\beta}_0$, $\hat{\beta}_1$ et $\hat{\sigma}^2$ possèdent un sens dans le cadre du modèle. En revanche, les intervalles de confiance, les tests et les seuils observés sont tous biaisés en l'absence de normalité. Il est possible de vérifier la normalité par des méthodes graphiques (histogramme, droite de Henry) ou par des tests formels (Shapiro-Wilk et Kolmogorov-Smirnov, ils ne seront pas étudiés dans le cours). Dans tous les cas, ces tests sont basés sur les résidus studentisés, car même si ces derniers sont dépendants, ils présentent l'avantage de posséder la même variance.

L'histogramme

L'histogramme permet de vérifier la symétrie et l'unimodalité. Une distribution normale est unimodale et symétrique, comme illustré à la figure 2.6a).

Le QQ-plot normal

Le diagramme de dispersion des résidus studentizés $r_{(i)}$ en fonction des quantiles de la loi normale centrée réduite $u_{(i)}$ permet de vérifier la normalité des résidus. Il est appelé le QQ-plot normal.

Si les résidus sont normaux, les points devraient être alignés. Lorsque les points forment une courbe concave ou convexe, la distribution des résidus est non symétrique. Un graphique en forme "d'intégrale

1. Pour plus d'information sur ce type de modèle, on se référera au cours sur les séries chronologiques.

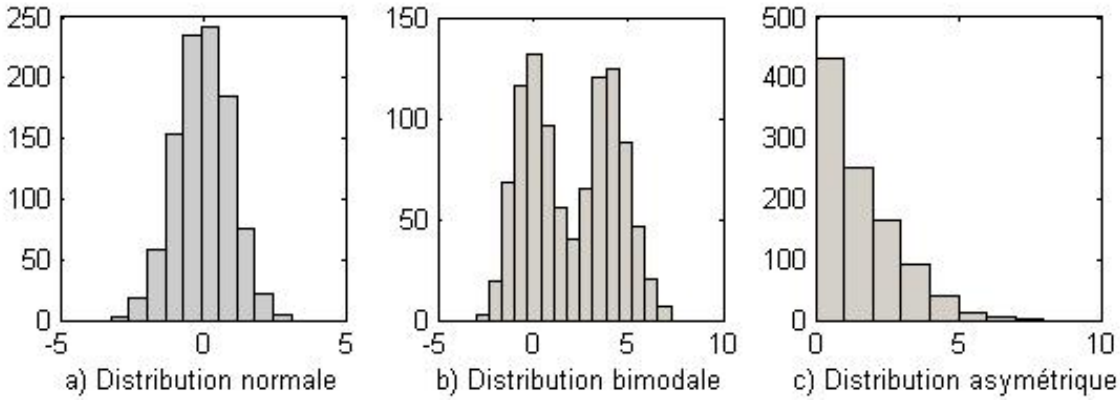

FIGURE 2.6: Histogramme présentant la distribution des résidus. Seule la distribution normale (a) est jugée acceptable. Dans les cas b) et c), le modèle apparaît comme n'étant pas adéquat.

inversée couchée" (voir par exemple la figure 2.9) indique que la loi normale a des queues plus épaisses que la distribution des résidus. Un graphique tel que la Figure 2.8) indique que la distribution normale a des queues plus légères que la distribution des résidus. C'est un problème fréquent et non négligeable, comme cela représente un risque supplémentaire pour l'assureur.

2.11 Transformation des données

Dans le cas où les postulats de la régression ne sont pas respectés, c'est-à-dire que le modèle $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ ne colle pas à la réalité, il est possible d'effectuer une transformation des variables qui permettra d'utiliser le modèle de régression linéaire de façon appropriée. Il existe souvent une fonction $g(\cdot)$ qui transforme la variable de façon à régler un problème de linéarité, d'hétéroscédasticité ou de normalité.

Dans cette section, on propose deux méthodes qui facilitent le choix de la transformation. La première méthode est particulièrement utile dans le cas où le postulat d'homoscédasticité n'est pas respecté. Elle consiste à trouver une approximation de la variance de la variable aléatoire transformée. La méthode de Box-Cox est quant à elle utilisée pour régler un problème de linéarité. Il peut aussi être intéressant d'essayer cette méthode dans le cas d'un problème de normalité.

Ces méthodes sont très utiles et il ne faut pas hésiter à les employer lorsque nécessaire. Cependant, il ne faut pas oublier de considérer aussi l'interprétation du modèle. Dans bien des cas, la nature même du phénomène peut nous éclairer sur le type de transformation à appliquer. Il faut aussi considérer qu'une variable à la puissance $1,8765$ est souvent plus difficile à concevoir et à interpréter qu'une variable à la puissance 2. Par conséquent, on préférera généralement arrondir les puissances. Un dernier point à noter avant d'amorcer la description des techniques de transformation de variable est qu'il vaut mieux appliquer la transformation à la variable exogène plutôt qu'à la variable endogène, l'interprétation des intervalles de

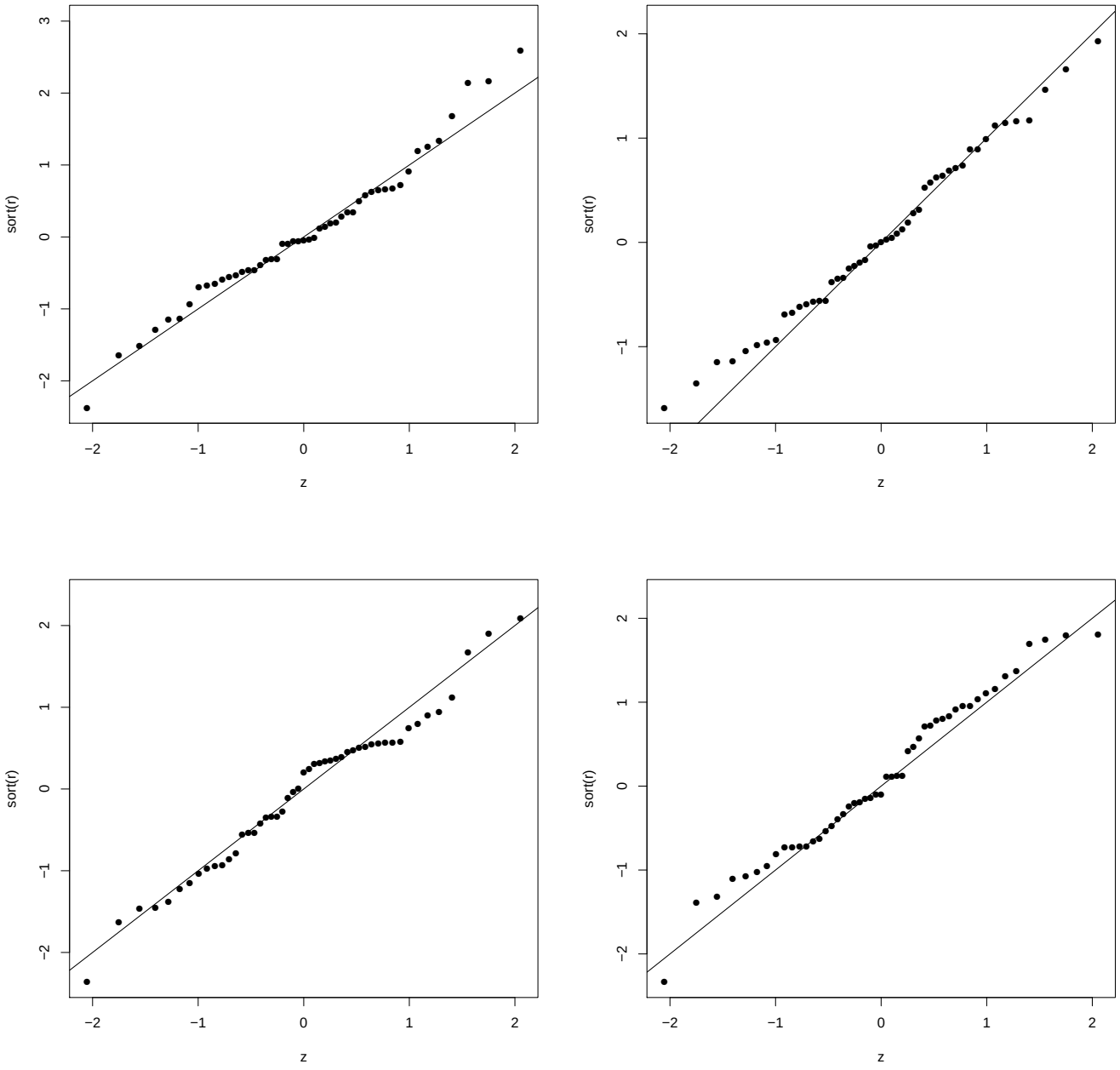

FIGURE 2.7: Exemples de QQ-plot normal où les résidus sont normaux

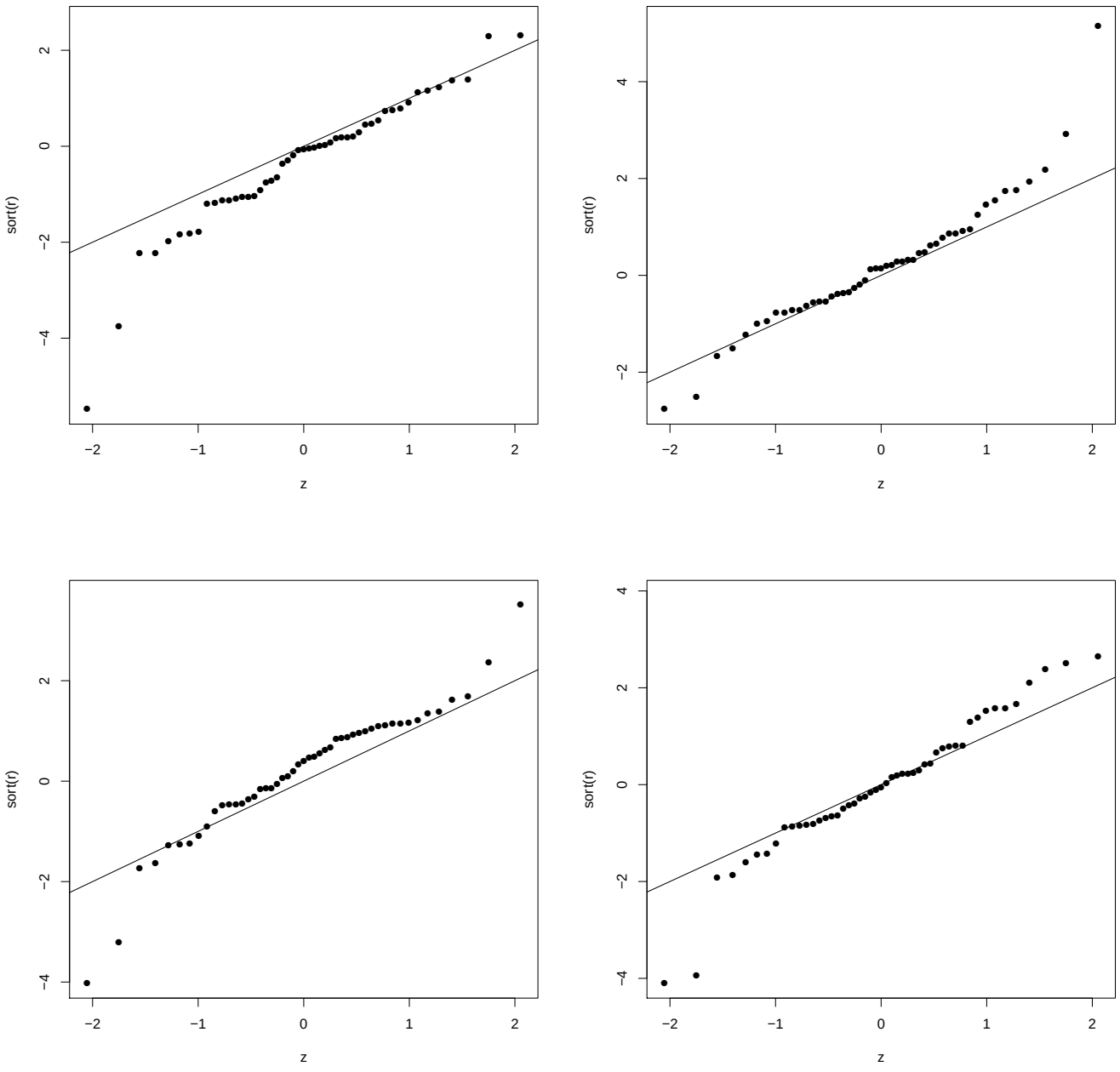

FIGURE 2.8: Exemples de QQ-plot normal où la distribution des résidus a des queues plus épaisses que la normale

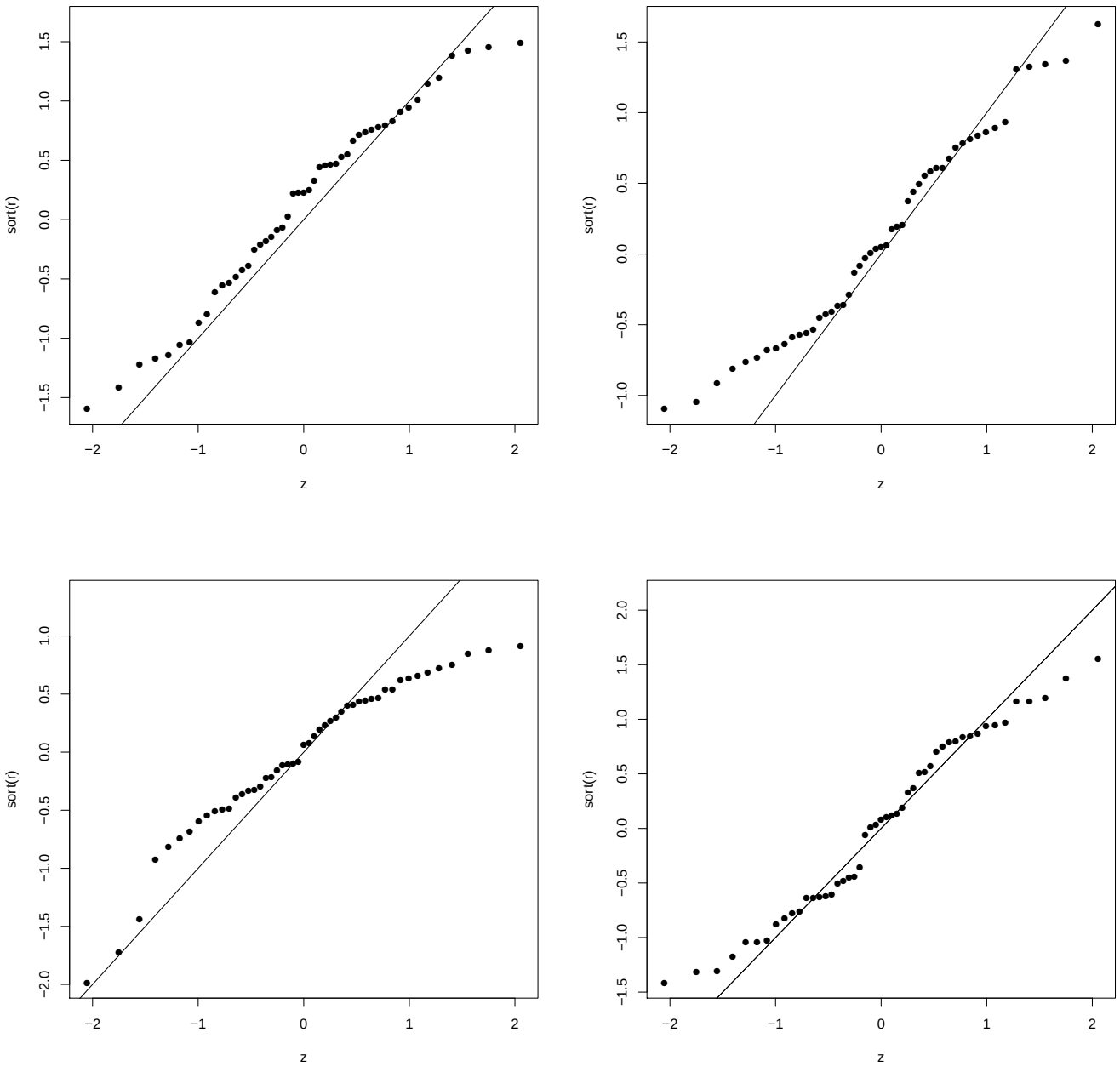

FIGURE 2.9: Exemples de QQ-plot normal où la distribution des résidus a des queues plus légères que la normale

confiance et de prévision sera plus facile.

2.11.1 Approximation de la variance d'une variable aléatoire transformée

Lorsque le diagramme de dispersion des résidus en fonction des valeurs ajustées montre un lien entre la variation des résidus et les valeurs ajustées, on se trouve dans un cas d'hétéroscédasticité. Le choix des transformations possibles pour stabiliser la variance se base sur la formule approximative pour la variance d'une variable transformée $g(Y)$. Le développement de Taylor du premier ordre conduit à

$$\begin{aligned} g(Y) &= g(\mathbb{E}[Y] + \{Y - \mathbb{E}[Y]\}) \\ &\approx g(\mathbb{E}[Y]) + g'(\mathbb{E}[Y])(Y - \mathbb{E}[Y]). \end{aligned}$$

La variance peut donc être approchée par

$$\mathbb{V}\mathrm{ar}[g(Y)] \approx \left\{g'(\mathbb{E}[Y])\right\}^2 \mathbb{V}\mathrm{ar}[Y].$$

Soit par exemple Y une variable aléatoire telle que son écart-type en fonction de l'espérance est $\sigma(\mu) = \mu^{1/2}$. La substitution dans la formule aproximative pour la variance nous montre que

$$\mathbb{V}\mathrm{ar}[g(Y)] \approx \{g'(\mu)\}^2\sigma^2 = \{g'(\mu)\}^2\mu.$$

La transformation $g(\cdot)$ telle que la variance de $g(Y)$ soit constante doit donc vérifier $g'(\mu)$ proportionnelle à $1/\sqrt{\mu}$, c'est-à-dire $g(Y) = \sqrt{Y}$ à une constante près. Lorsque Y prend des valeurs positives, les transformations les plus utiles sont résumées dans le tableau 2.2

TABLE 2.2: Transformation envisageable pour régler l'hétéroscédasticité

Transformation typique	Situation
$\sqrt{Y}$	$\mathbb{V}\mathrm{ar}[\epsilon_i] \propto \mathbb{E}[Y_i]$ (données de type Poisson), figure 2.4 b
$\log(Y)$	$\mathbb{V}\mathrm{ar}[\epsilon_i] \propto (\mathbb{E}[Y_i])^2$ (efficace si Y possède une très grande étendue)
$1/Y$	$\mathbb{V}\mathrm{ar}[\epsilon_i] \propto (\mathbb{E}[Y_i])^4$
$\arcsin(\sqrt{Y})$	$Y \in [0,1]$ et $\mathbb{V}\mathrm{ar}[\epsilon_i] \propto \mathbb{E}[Y_i](1 - \mathbb{E}[Y_i])$ (Y suit une loi Bernouilli)

2.11.2 La technique de Box-Cox

L'idée de Georges Box et David Cox, publiée en 1964, consiste à généraliser le modèle de régression en introduisant un paramètre supplémentaire λ. Le modèle devient alors :

$$g(Y_i, \lambda) = \beta_0 + \beta_1 x_i + \epsilon_i \tag{2.16}$$

où λ est un paramètre inconnu à estimer à partir des données et la fonction g est définie par

$$g(y; \lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0. \end{cases} \tag{2.17}$$

Il est à noter que $\lim_{\lambda \to 0} \dfrac{y^\lambda - 1}{\lambda} = \lim_{\lambda \to 0} \dfrac{y^\lambda \log(y)}{1} = \log(y)$. On suppose la normalité des termes d'erreurs et on maximise la vraisemblance par rapport à $\beta_0, \beta_1, \sigma^2$ et λ. Pour cela on pose $W_i(\lambda) = \beta_0 + \beta_1 x_i + \epsilon_i$. Comme on a supposé la normalité des termes d'erreur, on a que $W_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$. La vraisemblance s'exprime alors comme

$$L(\beta_0, \beta_1, \sigma, \lambda) = \prod_{i=1}^{n} f(W_i) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) \exp \left[-\frac{1}{2} \left\{ \frac{W_i - (\beta_0 + \beta_1 x_i)}{\sigma} \right\}^2 \right] J(\lambda),$$

où $J(\lambda)$ note le jacobien de la transformation. Le problème qui se pose à ce point est que l'on ne peut résoudre $\frac{\partial L}{\partial \beta_0} = \frac{\partial L}{\partial \beta_1} = \frac{\partial L}{\partial \sigma^2} = \frac{\partial L}{\partial \lambda} = 0$ explicitement et qu'il faut procéder numériquement. Certains logiciels tels que `SAS` (proc `transreg`) et `R` (fonction `boxcox` du package `car`), permettent d'utiliser la technique de Box-Cox et donnent un intervalle de confiance pour la valeur de λ.

La technique peut être contestée du fait qu'elle suppose la normalité des erreurs. En pratique, on utilise simplement la transformation Y_i^λ si $\lambda \neq 0$ plutôt que $\frac{y^\lambda - 1}{\lambda}$.

2.12 Prévision sous transformation

Il faut être prudent lorsqu'on applique une transformation à la variable endogène. En effet les prévisions sont alors obtenues en termes de la variable endogène transformée. Or, en général on a que,

$$\mathbb{E}[g(Y)] \neq g(\mathbb{E}[Y]).$$

Il est donc incorrect de simplement appliquer la transformation inverse afin d'obtenir une prévision en termes de la variable aléatoire endogène Y de base (non transformée). La détermination d'un intervalle de confiance est par contre direct, en utilisant le fait que lorsque $g(\cdot)$ est une fonction monotone, on a que

$$\mathbb{P}\{a \leq g(Y) \leq b\} = \mathbb{P}\{g^{-1}(a) \leq Y \leq g^{-1}(b)\},$$

où g^{-1} désigne la transformation inverse.

Exemple 2.4. *Soit le modèle de régression linéaire simple suivant :*

$$g(Y_i) = Y_i^* = \beta_0 + \beta_1 x_i + \epsilon_i,$$

où $g(Y) = \sqrt{Y}$. *Un intervalle de confiance à 90 % pour* $g(Y)$ *lorsque* $x = x_0$ *est donné par* $(0.2; 2.4)$. *On désire trouver un intervalle de prévision à 90 % lorsque* $x = x_0$. *Solution : On a que* $Y^* = \sqrt{Y} \Leftrightarrow Y = (Y^*)^2$. *L'intervalle de prévision en termes de la variable endogène de base* Y *sera donc* $(0.2^2; 2.4^2) = (0.04; 5.76)$.

2.13 Exemple

Cet exemple est tiré de Montgomery *et al.* (2012). On s'intéresse à modéliser le temps nécessaire à la livraison de boissons gazeuses (en minutes) en fonction du nombre de caisses livrées. Les données sont présentées dans le tableau 2.3.

Ces données sont placées dans un fichier texte (qui est disponible sur le site du cours). Pour accéder aux données, on peut enregistrer le fichier texte dans le dossier de travail de `R` et utiliser la commande :

```
dat <- read.table("TempsBoissons.txt",header=TRUE)
```

On trace d'abord un nuage de points avec les données :

```
plot(dat, pch=16)
```

et on voit le résultat dans le graphique 2.10. Il semble qu'une relation linéaire soit un modèle raisonnable. On considère donc le modèle

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

L'estimation des paramètres peut être effectuée simplement en `R` :

```
> modele <- lm(Temps~Caisses,data=dat)
> summary(modele)

Call:
lm(formula = Temps ~ Caisses, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5811 -1.8739 -0.3493  2.1807 10.6342

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.321      1.371   2.422   0.0237 *
```

Observation i	Temps y_i	Caisses x_i
1	16.68	7
2	11.5	3
3	12.03	3
4	14.88	4
5	13.75	6
6	18.11	7
7	8	2
8	17.83	7
9	79.24	30
10	21.5	5
11	40.33	16
12	21	10
13	13.5	4
14	19.75	6
15	24	9
16	29	10
17	15.35	6
18	19	7
19	9.5	3
20	35.1	17
21	17.9	10
22	52.32	26
23	18.75	9
24	19.83	8
25	10.75	4

TABLE 2.3: Données pour l'exemple 2.13

```
Caisses        2.176      0.124  17.546 8.22e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.181 on 23 degrees of freedom
Multiple R-squared: 0.9305,Adjusted R-squared: 0.9275
F-statistic: 307.8 on 1 and 23 DF,  p-value: 8.22e-15
```

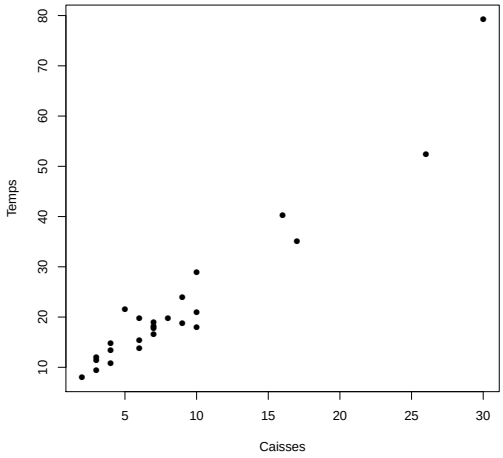

FIGURE 2.10: Nuage de points des données pour l'exemple 2.13

Cette sortie, obtenue avec la commande `summary`, montre que les estimations des paramètres selon la méthode des moindres carrés sont

$$\hat{\beta}_0 = 3.321 \text{ et } \hat{\beta}_1 = 2.176.$$

On peut interpréter ces valeurs de la manière suivante :

- Lorsque le nombre de caisses à livrer augmente de 1, alors en moyenne, le temps de livraison augmente de 2.176 minutes.
- Si on avait à faire une livraison avec aucune caisse, le temps de livraison serait de 3.321 minutes. On remarque que le paramètre β_0 n'a pas toujours une interprétation logique, en particulier lorsque $x = 0$ est impossible.

Les résultats présentés ci-haut contiennent également les écart-types des paramètres, soient

$$\sqrt{Var(\hat{\beta}_0)} = 1.371 \text{ et } \sqrt{Var(\hat{\beta}_1)} = 0.124.$$

On peut également lire les statistiques t pour tester si les paramètres sont égaux à 0. Pour tester

$$H_0 : \beta_0 = 0 \text{ vs } H_1 : \beta_0 \neq 0,$$

on trouve $t_{obs} = \frac{3.321}{1.371} = 2.422$. De plus, la valeur p de ce test est

$$2P[|T| > 2.422] = 0.0237,$$

si $T \sim \mathcal{T}(23)$. On rejette donc l'hypothèse nulle que $\beta_0 = 0$. De même, on trouve que le coefficient de régression lié à la variable `Caisses` est hautement significatif. En effet, la valeur p est $8.22e-15$, ce qui est très près de 0. La valeur estimée de σ^2 est

$$s^2 = 4.181.$$

On peut aussi lire que le coefficient de détermination, est 93.05%, ce qui signifie que plus de 93% de la variabilité dans le temps de livraison est expliqué par le nombre de caisses livrées. Ensuite, on peut ajouter la droite de régression sur le nuage de points avec la commande :

```
abline(modele).
```

On obtient le tableau ANOVA comme suit :

```
> anova(modele)
Analysis of Variance Table

Response: Temps
          Df Sum Sq Mean Sq F value   Pr(>F)
Caisses    1 5382.4  5382.4  307.85 8.22e-15 ***
Residuals 23  402.1    17.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Le ratio $F_0 = 307.85$ de Fisher est contenu dans ce tableau mais est également présenté dans les résultats sommaires du modèle. Selon ce test, on peut rejeter l'hypothèse nulle que le modèle n'est pas valide à un niveau de confiance de 1%, ou même à un niveau de confiance de 0.05%.

Pour calculer les intervalles de confiance sur la droite de régression, on utilise la fonction `predict` avec l'argument `interval="confidence"`. Il faut d'abord créer un objet de classe `data.frame` qui contient les nouvelles données pour les prédictions. Pour obtenir un intervalle de confiance pour les prédictions, on utilise l'argument `interval="prediction"`.

```
> newdat <- data.frame(Caisses=2*(1:5))
> predict(modele,newdata=newdat,type="response",interval="confidence")
        fit       lwr      upr
```

```
1  7.673113  5.223396 10.12283
2 12.025447  9.907812 14.14308
3 16.377780 14.508479 18.24708
4 20.730113 18.989183 22.47104
5 25.082447 23.323459 26.84143
> predict(modele,newdata=newdat,type="response",interval="prediction")
        fit       lwr      upr
1  7.673113 -1.316965 16.66319
2 12.025447  3.120124 20.93077
3 16.377780  7.528222 25.22734
4 20.730113 11.906779 29.55345
5 25.082447 16.255532 33.90936
```

On voit que les intervalles de confiance pour les prédictions sont beaucoup plus large que ceux pour la droite de régression, tel qu'attendu. Dans la figure 2.11, on voit les bornes de l'intervalle de confiance pour la droite de régression, en forme d'hyperbole, ce qui souligne que les prévisions sont plus précises lorsque x_0 est près de sa moyenne.

Pour vérifier le postulat de la linéarité, on a tracé le graphique de Y_i en fonction de x_i et on a vu que la relation était approximativement linéaire. On peut également tracer les graphiques des résidus en fonction de la valeur prédite :

```
plot(fitted(modele),residuals(modele),pch=16,xlab="Valeur Prédite", ylab="Residus")
```

Ou encore le graphique des résidus studentisés en fonction de la valeur prédite :

```
plot(fitted(modele),rstandard(modele),pch=16,xlab="Valeur Prédite", ylab="Residus studentises")
```

Et finalement le graphique des résidus en fonction de la variable exogène :

```
plot(dat[,1],residuals(modele),pch=16,xlab="Variable exogene", ylab="Residus").
```

Les trois graphiques sont présentés dans la Figure 2.12. L'hypothèse de linéarité semble adéquate, toutefois, la forme d'entonnoir caractérise de l'hétéroscédasticité. C'est-à-dire que la variance des résidus semble plus élevée pour les grandes valeurs de Y ou de x.

Pour vérifier le postulat de la normalité des résidus, on trace l'histogramme des résidus studentisés (voir Figure 2.13) :

```
hist(rstandard(modele),freq=FALSE,main="Histrogramme des residus studentises")
curve(dnorm(x),add=TRUE,col=2, lwd=2, lty=2)
```

La commande `curve(dnorm(x))` permet d'ajouter la courbe de la densité de la loi Normale(0,1) sur le graphique pour mieux comparer avec l'histogramme. On peut également tracer le QQ-Plot normal des résidus studentisés :

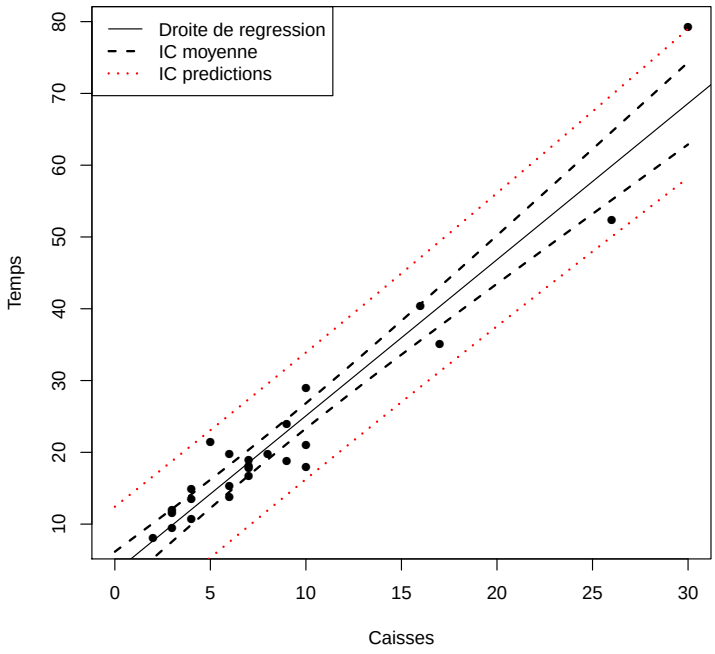

FIGURE 2.11: Droite de régression et intervalles de confiance pour l'exemple 2.13

```
qqnorm(rstandard(modele),pch=16)
qqline(rstandard(modele))
```

Il semble que la distribution normale ait des queues trop légère pour bien représenter les résidus. Les résultats des intervalles de confiance et des tests d'hypothèse pourraient en être affectés.

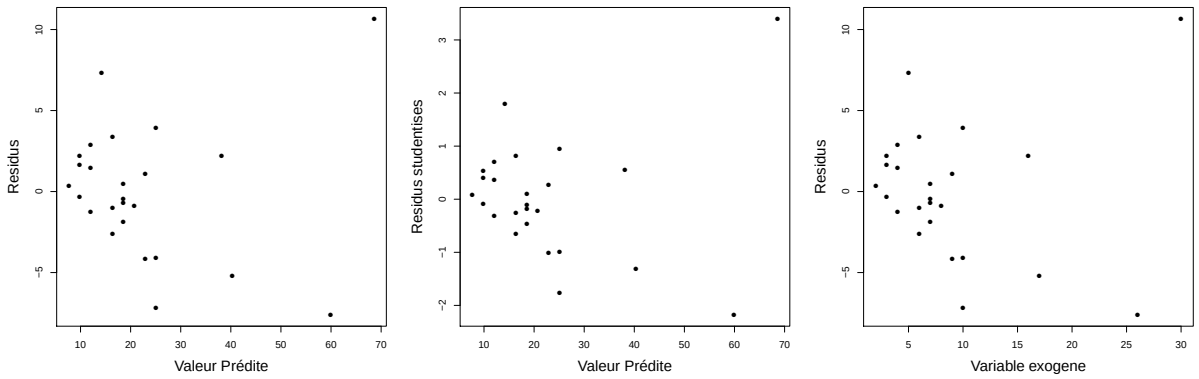

FIGURE 2.12: Vérification des postulats de linéarité et d'homoscédasticité

Si on utilise la méthode de Box-Cox sur ce modèle, on observe que la transformation racine-carrée est appropriée pour corriger le problème d'hétéroscédasticité. La fonction `boxcox` du package `car` trace la log-vraisemblance sur un intervalle donné.

```
library(car)
par(mfrow=c(1,2))
boxcox(Temps~Caisses,lambda=seq(-2,2,len=20),plotit=TRUE,data=dat)
boxcox(Temps~Caisses,lambda=seq(0,1,len=20),plotit=TRUE,data=dat)
```

Le résultat de la commande précédente est illustré dans la Figure 2.14. Ce nouveau modèle est donc :

```
> modele2 <- lm(I(Temps^0.5)~Caisses, data=dat)
> summary(modele2)

Call:
lm(formula = I(Temps^0.5) ~ Caisses, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59277 -0.25560  0.01804  0.22476  0.81350
```

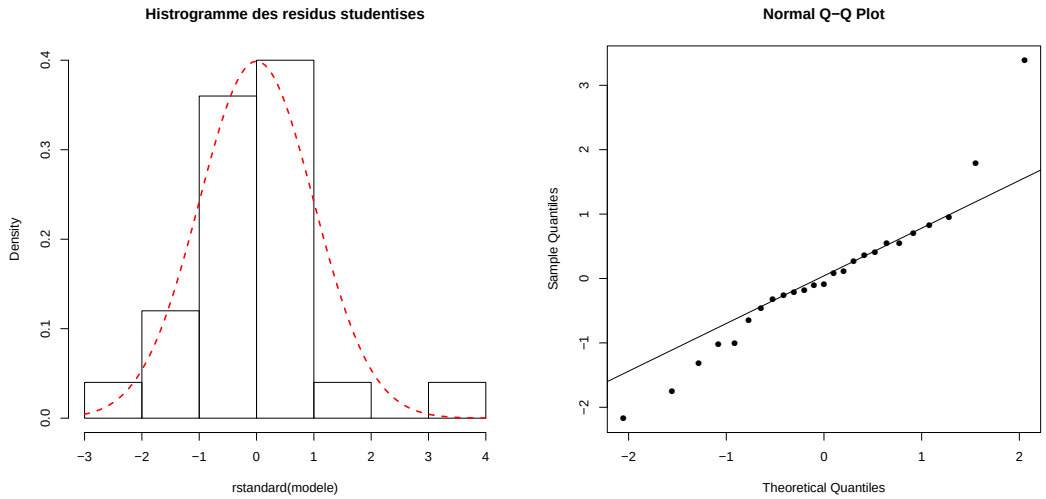

FIGURE 2.13: Vérification du postulat de normalité

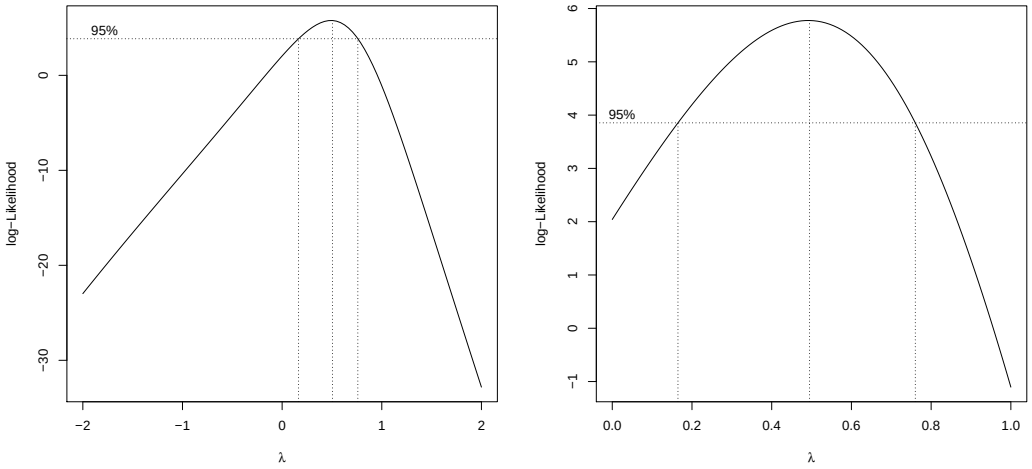

FIGURE 2.14: Intervalle de confiance à 95% pour la valeur de λ avec la technique de Box-Cox.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.87028    0.11886   24.15  < 2e-16 ***
Caisses      0.19061    0.01075   17.73 6.59e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3625 on 23 degrees of freedom
Multiple R-squared: 0.9318,Adjusted R-squared: 0.9288
F-statistic: 314.2 on 1 and 23 DF,  p-value: 6.594e-15
```

On peut réaliser l'analyse sur les paramètres encore une fois. Le modèle est valide, et les postulats de normalité et d'homoscédasticité sont maintenant vérifiés, comme montré dans la Figure 2.15.

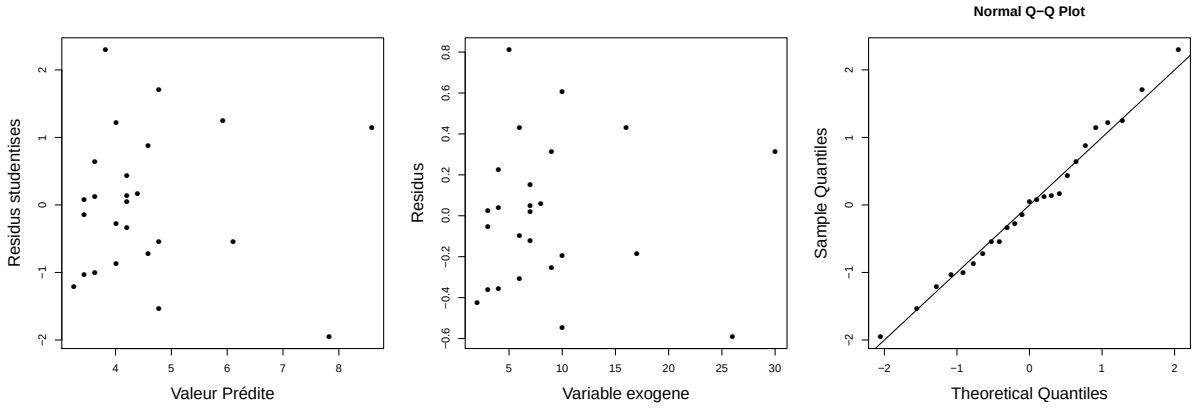

FIGURE 2.15: Vérification des postulats de normalité et d'homoscédasticité avec le modèle $\sqrt{Y} = \beta_0 + \beta_1 x$.

Il faut toutefois être prudents pour les prédictions avec ce nouveau modèle ; on ne peut pas directement effectuer une prédiction, mais on peut donner un intervalle de confiance en prenant le carré de l'intervalle de confiance obtenu. Les résultats sont présentés dans le graphique 2.16.

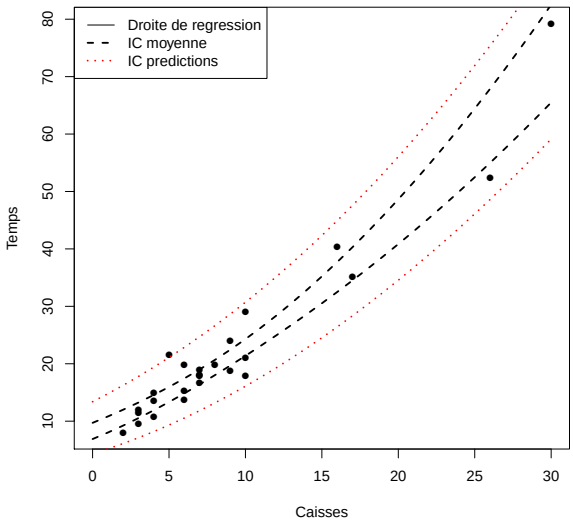

FIGURE 2.16: Intervalles de confiance pour la moyenne et pour la prédiction avec le modèle $\sqrt{Y} = \beta_0 + \beta_1 x$.

Bibliographie

Draper, N. R. et Smith, H. (1998). *Applied regression analysis.* Wiley Series in Probability and Statistics. Wiley, New Jersey, USA.

Faraway, J. (2002). *Practical Regression and Anova using R.* http ://cran.r-project.org/doc/contrib/Faraway-PRA.pdf.

Montgomery, D. C., Peck, E. A. et Vining, G. G. (2012). *Introduction to linear regression analysis. 5th Edition.* Wiley Series in Probability and Statstics. Wiley, New Jersey, USA.

Weisberg, S. (1985). *Applied linear regression. Second Edition.* Wiley Series in Probability and Mathematical Statistics. Wiley, New Jersey, USA.