PrivateTeacher
Maîtriser les Sciences Exactes

STATISTIQUES

Cours-Résumé de Psychologie Quantitative
Introduction aux modèles mathématiques
Régression linéaire simple

Julien RUPPEN

14 April, 2026

Abstract

La régression linéaire est une démarche qui consiste à superposer une droite à une série d'observations. Le but est d'utiliser la droite à la place des observations afin d'en simplifier la compréhension. Une droite en effet, est une relation explicite, facile à comprendre alors que les observations sont souvent complexes et difficiles à expliquer. Une droite de régression permet donc de simplifier les observations. On dit aussi qu'elle permet de s'en faire une représentation. On appel cette représentation **un modèle** mathématique. Un modèle s'écrit sous la forme d'une équation. Les modèles linéaires s'écrivent $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$, avec $\hat{\beta}_0$ et $\hat{\beta}_1$ les paramètre du modèle, l'ordonnée à l'origine et la pente respectivement. Le chapeau au dessus des paramètre indique qu'il s'agit d'estimation. Ce document montre l'imporance de **faire la difference entre le modèle et la réalité**. Cette distinction nous permettra de comprendre l'écart entre les valeurs prédites $\hat{Y}_i$ et les valeurs observées Y_i. Ce document démontre comment adopter une approche systématique dans l'interprétation des modèles mathématique dans le cadre de la psychologie quantitative.

Contents

1. Introduction

1.1. Modèles mathématiques

Les **modèles mathématiques** sont des instruments dont on se sert pour comprendre la réalité. Tous comme un tableau ou un dessin, les modèles sont des représentations simplifiées de la réalité et ne reproduisent donc pas exactement la réalité. Pour cette raison, il existe toujours une différence entre le modèle et la réalité. S'ils sont construit correctement cependant, les modèles peuvent servire à comprendre les phénomènes du monde réel, le comportement des individus en particulier.

1.2. Les équations

Un modèle mathématique s'écrit sous la forme d'une **égalité**. Une égalité est une relation entre deux grandeurs. Elle traduit l'idée que ces grandeurs sont liées. C'est ce qui permet de calculer l'une à partir de l'autre. C'est ainsi que l'on fait des prédictions. Ce document présente l'un des modèles les plus connu en psychologie quantitative: le **modèle linéaire**.

En psychologie quantitative, les modèles mathématiques servent à mettre en relation des variables du type scores, mesures comportementales ou indicateurs cliniques.

1.3. Le modèle linéaire

Deux grandeurs sont **proportionnelle** (on dit aussi "dans une relation de proportionnalité") si leur rapport reste constant. Considérons par exemple les paires de valeurs suivantes:

X	Y	Y/X
3	6	2
6	12	2
12	24	2

On le voit, le rapport Y/X est toujours égale à deux 2, quelle que soit la valeur de X et de Y. Pour cette raison, on dira que X et Y sont proportionnels.

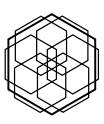
Le **modèle linéaire** formalise cette relation entre X et Y sous la forme d'une équation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Convention Le symbole ˆ (chapeau) signale que la valeur sous laquelle il se trouve n'est pas la valeur réelle, mais une estimation. $\hat{Y}$ désigne donc une estimation de la variable Y. Il s'agit de la valeur prédite par le modèle, non pas de la valeur observée.

1.3.1 Variable dépendante et Indépendante

Il est d'usage d'appeler la variable X une **variable indépendante**, car celle-ci ne dépend d'aucune autre variable et on la choisi librement. Une variable indépendante constitu



donc un point de départ pour expliquer l'autre variable, Y. X sert donc à expliquer Y pour cette raison on l'appel aussi une **variable explicative**. Y quant à elle est la variable que l'on cherche à expliquer. Elle dépend de la variable X on l'appel donc **variable dépendante**

1.3.2 Coefficient de proportionnalité

Le coefficient $\hat{\beta}_1$ quant à lui est le **coefficient de proportionnalité** entre X et Y: il exprime de combien $\hat{Y}$ varie pour chaque unité supplémentaire de X. Dans notre exemple, ce coefficient vaut 2 $\hat{\beta}_1 = 2$ Ce coefficient est aussi la pente de la droite. C'est lui qui porte toute l'information sur la relation entre les deux variables, c'est donc ce paramètre dont on va se servir lors de l'interprétation des résultats. Le coefficient $\hat{\beta}_0$ est l'ordonnée à l'origine. Il s'agit de la valeur prédite de Y lorsque $X = 0$.

1.4. Les résidus : mesure d'imprécision

Le modèle étant une représentation de la réalité, il existe toujours un écart entre la valeur prédite $\hat{Y}_i$ et la valeur observée Y_i. Cette différence s'appelle un **résidu**. On la note e_i:

$$e_i = Y_i - \hat{Y}_i$$

Lorsque l'on parle de résidu on sous-entend la plupart du temps non pas un résidu particulier e_i mais l'ensemble de tous les résidu e. S'il n'y a qu'un seul modèle en effet, il y a par contre plusieurs observations. Pour chaque observation, il existe donc une différence par rapport au modèle. Il y a donc toujours **plusieurs résidus**.

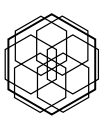
Un résidu unique n'est pas vraiment utile en soit. C'est plutôt l'ensemble qu'il est possible d'interpréter. Pour cette raison, on s'intéresse souvent à **la distribution des résidus**

1.4.1 Normalité des résidus

Lorsque le modèle s'ajuste correctement, les résidus ne présentent ni biais ni asymétrie. Les variation qui reste (résidu) de part et d'autre de la droite est supposé n'être que du bruit c'est à dire des valeurs aléatoires, sans structure particulière. Or **la signature statistique du bruit est la distribution gaussienne**. Pour cette raison, on cherche si les résidu se distribue normalement.

Pour évaluer si un modèle a correctement capturé la tendance des données, il est utile d'examiner les deux caractéristique suivante:

1. **Les résidus sont-ils centrés en zéro ?** Si le modèle surestime systématiquement ou sous-estime systématiquement, la moyenne des résidus sera différente de zéro. Cela indique un biais
2. **Les résidus sont-ils distribués de façon symétrique ?** Une asymétrie dans la distribution signale que les erreurs du modèle ne sont pas également réparties dans les deux sens. Cela peut indiquer une relation non linéaire que le modèle linéaire ne capture pas.



1.5. Représentation Graphique

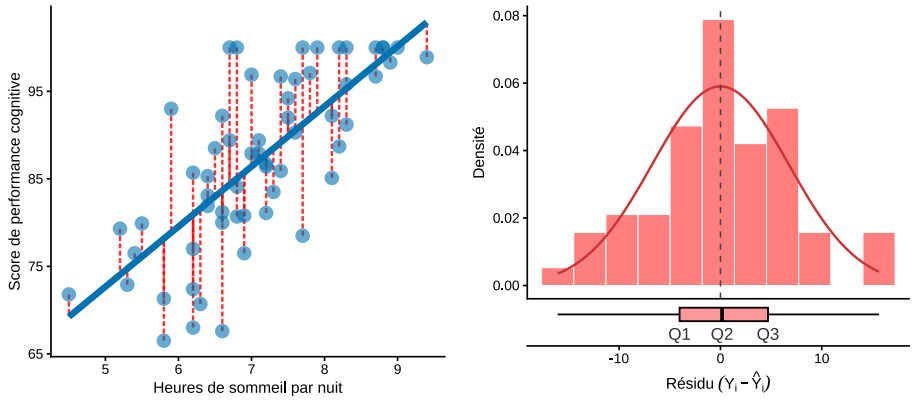


Figure 1: Gauche : relation entre les heures de sommeil (X) et le score de performance cognitive (Y). La droite est le modèle ajusté ; les segments rouges sont les résidus $e_i = Y_i - \hat{Y}_i$. Droite : distribution des résidus (histogramme rouge), courbe gaussienne ajustée (rouge épais) et boxplot horizontal.

Le panneau gauche montre la droite de régression au centre du nuage. Les segments rouges rendent visible ce que le modèle ne capture pas : pour chaque observation, la distance verticale entre le point et la droite. Le panneau droit montre leur distribution collective — c'est elle qui renseigne sur la qualité du modèle.

1.6. Fiabilité du modèle

Ajuster un modèle ne suffit pas. Encore faut-il pouvoir quantifier dans quelle mesure ce modèle est fiable. Trois statistiques permettent de le faire. Elles sont présentée ici et on les retrouvera dans chaque section de ce document.

1.6.1 Statistique F

La statistique F teste si le modèle **dans son ensemble** explique les observation Y mieux que le hasard. Si F est grand et la p-valeur associée inférieure à .05, on peut conclure que la relation entre X et Y est statistiquement significative. Si tel est le cas, alors on peut utiliser les coefficients du modèle pour faire des prédictions.

1.6.2 R2

Le R^2 mesure la **proportion de variance de** Y expliquée par le modèle. Il varie entre 0 (le modèle n'explique rien) et 1 (le modèle prédit Y parfaitement, sans résidu). Un $R^2 = 0.53$ signifie que le modèle capture 53 % de la variabilité du score, ce qui reste, 47 %, correspond à des sources de variation que X n'explique pas.

1.6.3 Normalité des résidus

La normalité des résidus est la vérification la plus directe de la qualité du modèle. Si le modèle a correctement capturé la tendance des données, ce qui reste — les résidus — ne doit être que du **bruit aléatoire**. Or la signature statistique du bruit aléatoire es

la **distribution gaussienne**. Or cette distribution est caractérisée par deux propriétés: centre en zéro et symmétri par rapport à zéro

Centrage en zéro. La médiane des résidus doit être proche de zéro. Un écart marqué signale que le modèle surestime ou sous-estime systématiquement les prédiction, ce qui représente un biais de prédiction. Dans l'exemple, la médiane vaut 0.16: aucun biais détectable.

Symétrie. La distribution doit être approximativement symétrique de part et d'autre de zéro. Dans la sortie R, on compare la queue gauche (Min → Q1 = 12 points) à la queue droite (Q3 → Max = 11 points). Ces deux valeurs sont proches : la distribution est donc symétrique.

1.6.3.1 QQ-plot et LOWESS Le **Q-Q plot** compare les quantiles des résidus à ceux d'une loi normale théorique. Si les points s'alignent sur la droite en pointillé, cela signifie que les quantile des résidus s'aligne sur les quantile d'une loi normale autrement dit, cela signifie que les résidus sont normalement distribués. Il s'agit la d'un condition nécessaire pour que l'inférences du modèle soient valides.

La **courbe LOWESS** détecte un biais systématique non capturé par le modèle. La ligne rouge doit longer la ligne zéro : un tracé plat confirme que les résidus sont bien centrés et qu'aucune structure ne subsiste. La Figure 2 présente les deux graphiques diagnostiques standards.

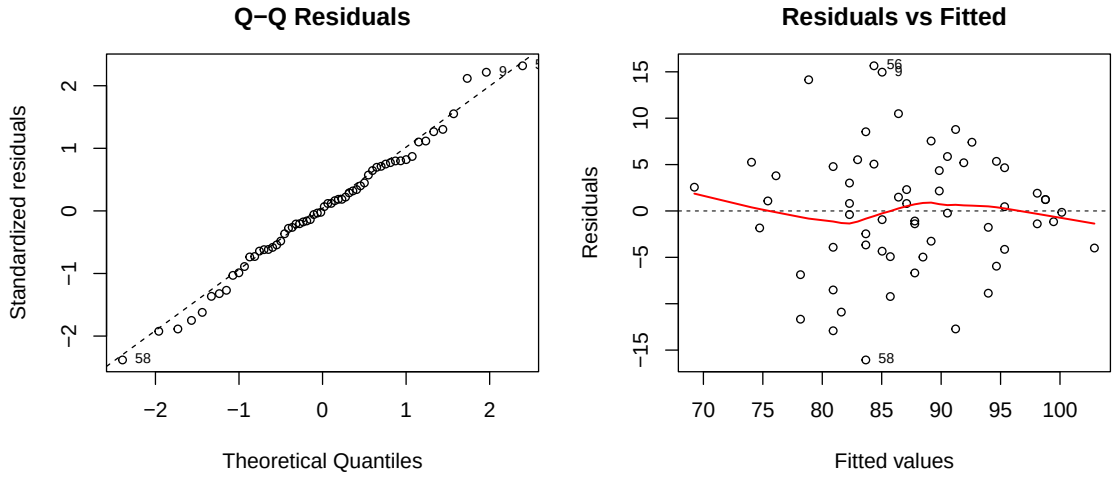


Figure 2: Gauche : Q-Q plot des résidus standardisés. Les points s'alignent sur la droite théorique : la normalité des résidus est confirmée par le test de Shapiro-Wilk (W = 0.989, p = 0.846). Droite : résidus en fonction des valeurs ajustées. La courbe LOWESS longe la ligne zéro : les résidus sont bien centrés autour de zéro (médiane = 0.16).

2. La régression linéaire simple

2.1. Une variable explicative continue

La variable explicative X est ici une variable **continue** : elle peut prendre n'importe quelle valeur numérique dans un intervalle. Dans le cadre de l'équation linéaire, une variable continue entre directement comme grandeur numérique dans le modèle. La relation entre X et Y s'écrit :

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Le coefficient $\hat{\beta}_1$ est le **coefficient de proportionnalité** entre X et Y : il indique de combien $\hat{Y}$ augmente pour chaque unité supplémentaire de X. Il s'agit de la pente de la droite.

2.2. Un exemple concret

Un chercheur en psychologie de l'éducation recrute 60 étudiants. Pour chacun, il mesure le nombre d'heures d'étude hebdomadaires (variable explicative X) et le score obtenu à un examen de fin de semestre sur 100 (variable critère Y). Les heures d'étude s'étendent de 2.1 à 19.8 heures par semaine. La situation est représentée sur la Figure 3.

2.2.1 Représentation Graphique

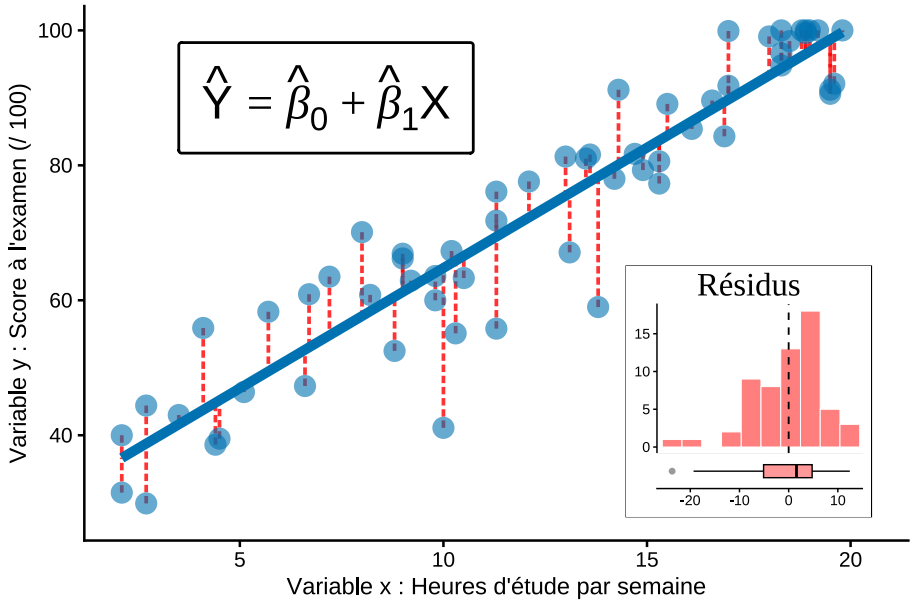


Figure 3: Relation entre les heures d'étude hebdomadaires et le score à l'examen. La droite représente le modèle linéaire ajusté par régression. Les segments rouges représentent les résidus $e_i = Y_i - \hat{Y}_i$: la distance verticale entre chaque observation et le modèle.

Le nuage de points révèle une relation positive entre X et Y : plus le nombre d'heures est élevé, plus le score est élevé. Les segments rouges représentent les résidus et rendent visible l'écart entre le modèle et la réalité.

2.3. Question de recherche

Si un étudiant consacre 25 heures par semaine à l'étude, quel score le modèle prédit-il pour cet étudiant ?

2.3.1 En d'autre termes

La question consiste à estimer le score à l'examen pour un étudiant dont le nombre d'heures d'étude ($X = 25$) se situe au-delà du domaine observé dans les données. Pour y répondre, il convient d'estimer les coefficients $\hat{\beta}_0$ et $\hat{\beta}_1$ par régression linéaire simple, puis d'évaluer l'équation $\widehat{\text{score}} = \hat{\beta}_0 + \hat{\beta}_1 \times 25$.

2.4. Méthode de résolution

1. **Identifier la variable que l'on cherche à expliquer.** C'est notre variable dépendante : le score à l'examen (Y).
2. **Identifier la ou les variables qui servent à expliquer.** C'est notre variable indépendante : les heures d'étude hebdomadaires (X).
3. **La relation entre ces variables est-elle statistiquement significative ?** On teste $H_0 : \beta_1 = 0$. Si $p < .05$, la relation est significative et $\hat{\beta}_1$ peut être interprété.
4. **Interpréter dans le cadre de la question.** On substitue $X = 25$ dans $\widehat{\text{score}} = \hat{\beta}_0 + \hat{\beta}_1 \times 25$ pour obtenir le score prédit.

2.5. Commande R

```
lm_sect1 <- lm(score ~ heures, data = data_sect1)
summary(lm_sect1)
```

La syntaxe `score ~ heures` se lit : *score en fonction de heures*. La variable critère Y figure à gauche du `~` ; la variable indépendante X à droite.

2.5.1 Lire la sortie logicielle

```
Call:
lm(formula = score ~ heures, data = data_sect1)

Residuals:
    Min      1Q  Median      3Q     Max
-23.691  -5.100   1.590   4.769  12.440

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   29.138      2.328   12.52   <2e-16 ***
 heures         3.565      0.174   20.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.224 on 58 degrees of freedom
Multiple R-squared:  0.8786,    Adjusted R-squared:  0.8765
F-statistic: 419.9 on 1 and 58 DF,  p-value: < 2.2e-16
```

2.5.2 Interpréter les résultats

La ligne `(Intercept)` donne $\hat{\beta}_0 = 29.14$: le score prédit par le modèle lorsque le nombre d'heures d'étude vaut zéro. Cette valeur est une extrapolation hors du domaine observé. (les données commencent à 2.1 heures). Il donne la valeur du score lorsqu'un étudiant ne travail pas (zéro heures) mais peut ne pas avoir de sense dans d'autre situation. Lorsque c'est le cas, on ne l'interprètera pas comme nous l'avons fait ici, mais on se contentera de traiter comme un simple **paramètre d'ajustement** du modèle sans d'autre utilité que de définir la droite de régression.

La ligne surlignée en bleu `heures` est la pente $\hat{\beta}_1 = 3.57$: c'est le **coefficient de proportionnalité** entre la variable indépendante et la variable critère. Pour une heure d'étude supplémentaire par semaine, le score augmente de 3.57 points.

2.5.3 Faire une prédiction

La régression a estimé $\hat{\beta}_0$ et $\hat{\beta}_1$. L'équation du modèle ajusté — **notre représentation de la réalité** — s'écrit :

$$\widehat{\text{score}} = 3.57 \times \text{heures} + 29.14$$

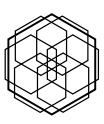
Cette équation représente notre modèle. Pour un étudiant qui consacre 25 heures par semaine à l'étude, on substitue $X = 25$:

$$\widehat{\text{score}} = 3.57 \times 25 + 29.14 = 118.3$$

Le modèle prédit $\widehat{\text{score}} = 118.3$ points. Cette prédiction est une extrapolation : $X = 25$ dépasse le maximum observé de 19.8 heures. La valeur prédite bien entendu, n'est pas exacte car il existe toujours une incertitude quantifiée par l'écart-type des résidus. (On appel cette valeur: RSE pour "Residual Standard Error")

2.6. Formuler une Réponse

Pour un étudiant qui étudie 25 heures par semaine, le modèle prédit $\widehat{\text{score}} = 118.3$ points ($\hat{\beta}_1 = 3.57$, $\hat{\beta}_0 = 29.14$, $R^2 = 0.88$, $p < .001$). Cette valeur est une extrapolation au-delà du domaine observé (2.1–19.8 heures).



2.7. Evaluer la qualité du modèle

2.7.1 Statistique F

La valeur $F = 419.9$ avec une p-valeur $< .001$ indique que le modèle dans son ensemble est statistiquement significatif : les heures d'étude apportent une information réelle sur le score.

2.7.2 R2

$R^2 = 0.879$: les heures d'étude expliquent 87.9 % de la variance du score. C'est la proportion de variabilité de Y que le modèle capture. Le reste — 12.1 % — correspond à des sources de variation que X n'explique pas.

2.7.3 Normalité des résidus

Centrage en zéro. La médiane des résidus (zone surlignée en rouge dans la sortie) vaut 1.59 : proche de zéro, aucun biais systématique détectable. Le modèle ne surestime ni ne sous-estime de façon systématique.

Symétrie. La queue gauche (Min → Q1) s'étend sur 18.6 points ; la queue droite (Q3 → Max) sur 7.7 points. Ces deux valeurs sont comparables : la distribution est approximativement symétrique.

2.7.3.1 QQ-plot et LOWESS

La Figure 4 confirme ces deux points visuellement.

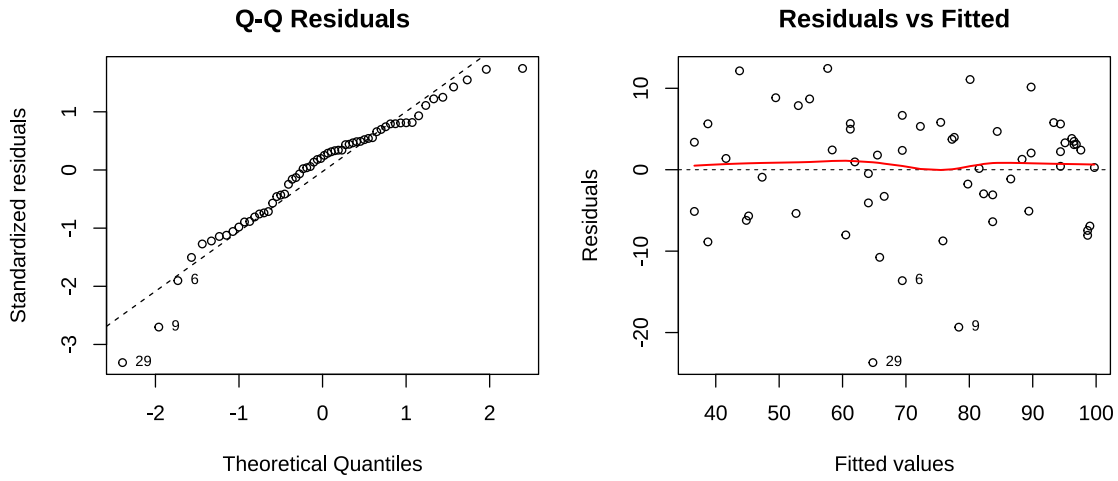


Figure 4: Gauche : Q-Q plot des résidus standardisés. Les points s'écartent de la droite théorique : la normalité des résidus est rejetée (W = 0.95, p = 0.016). Droite : résidus en fonction des valeurs ajustées. La courbe LOWESS longe la ligne zéro : les résidus sont bien centrés autour de zéro (médiane = 1.59).

3. Variable indépendante ordinale

3.1. Une variable explicative ordinale

Dans la section précédente, la variable explicative X était continue : elle pouvait prendre n'importe quelle valeur dans un intervalle. Ici, la variable explicative est **ordinale** : elle ne prend qu'un nombre limité de valeurs entières ordonnées. Dans ce cas, on peut encore utiliser le même modèle $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ à condition d'accepter l'**hypothèse d'équidistance**.

l'**hypothèse d'équidistance:** Une variable ordinale encode un ordre entre les modalités, mais pas nécessairement des distances égales entre elles. Le modèle linéaire suppose que le passage de 1 à 2 représente la même différence que le passage de 4 à 5. Cette hypothèse est rarement vérifiable mais elle est acceptée en pratique pour les échelles avec suffisamment de modalités (5 ou plus).

3.2. Un exemple concret

Un chercheur en psychologie quantitative recrute 75 participants en consultation ambulatoire. Pour chacun, il mesure le niveau de stress auto-rapporté sur une échelle ordinale de 1 à 5 (variable indépendante X) et le score de bien-être psychologique de 0 à 100 (variable critère Y, adapté du WHO-5). La situation est représentée sur la Figure 5.

3.2.1 Représentation Graphique

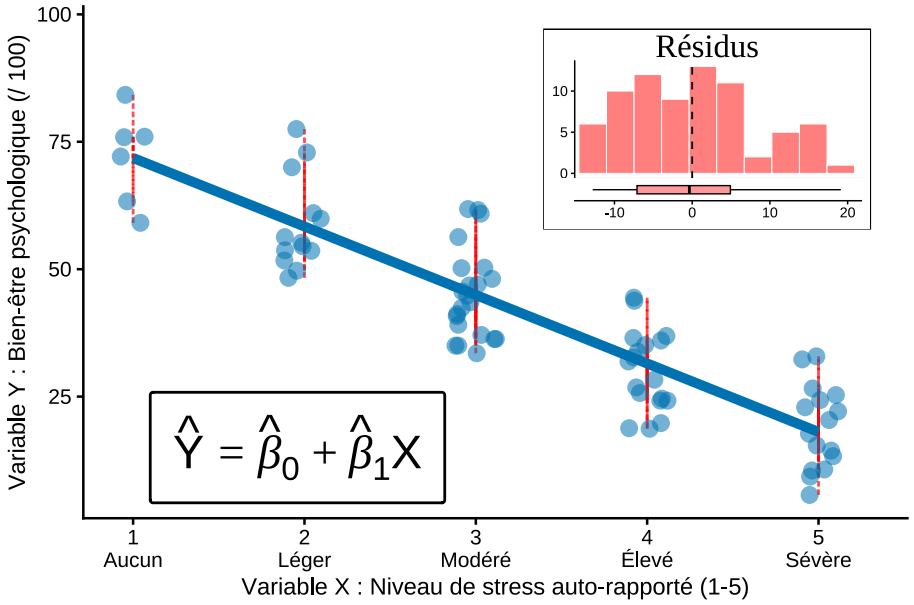


Figure 5: Relation entre le niveau de stress auto-rapporté (variable indépendante ordinale, 1–5) et le score de bien-être psychologique. Les points sont légèrement décalés horizontalement pour visualiser la densité à chaque modalité. La droite représente le modèle linéaire ajusté. Les segments rouges représentent les résidus $e_i = Y_i - \hat{Y}_i$.

La nature ordinale de la variable indépendante est immédiatement visible : les points se regroupent en cinq colonnes verticales correspondant aux cinq modalités, à la différenc

du nuage continu observé à la section précédente. La droite de régression traverse ce nuage avec une pente négative — plus le stress est élevé, plus le bien-être est faible.

3.3. Question de recherche

Quel score de bien-être le modèle prédit-il pour un patient dont le niveau de stress est de 3.5 ?

3.3.1 En d'autre termes

La question consiste à estimer le score de bien-être pour un patient dont le niveau de stress ($X = 3.5$) correspond à une valeur intermédiaire entre deux modalités entières de l'échelle ordinale. Pour y répondre, il convient d'appliquer l'équation du modèle $\widehat{\text{bien-être}} = \hat{\beta}_0 + \hat{\beta}_1 \times 3.5$, en acceptant l'hypothèse d'équidistance qui autorise le traitement numérique de la variable indépendante ordinale.

3.4. Méthode de résolution

1. **Identifier la variable que l'on cherche à expliquer.** C'est notre variable dépendante : le score de bien-être (Y).
2. **Identifier la ou les variables qui servent à expliquer.** C'est notre variable indépendante : le niveau de stress (X)
3. **La relation entre ces variables est-elle statistiquement significative ?** On teste $H_0 : \beta_1 = 0$. Si $p < .05$, la relation est significative et $\hat{\beta}_1$ peut être interprété.
4. **Interpréter dans le cadre de la question.** On substitue $X = 3.5$ dans $\widehat{\text{bien-être}} = \hat{\beta}_0 + \hat{\beta}_1 \times 3.5$ une interpolation entre deux modalités entières.

3.5. Commande R

```
lm_sect2 <- lm(bienetre ~ stress, data = data_sect2)
summary(lm_sect2)
```

La syntaxe est identique à celle de la section précédente. R traite la variable `stress` comme numérique, ce qui est précisément l'hypothèse d'équidistance qui est implicitement posée ici.

3.5.1 Lire la sortie logicielle

```
Call:
lm(formula = bienetre ~ stress, data = data_sect2)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-12.8010  -7.1010  -0.3624   4.9183  19.1219

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  85.2552     2.8415   30.00   <2e-16 ***
stress      -13.4386     0.8009  -16.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.425 on 73 degrees of freedom
Multiple R-squared:  0.7941,    Adjusted R-squared:  0.7913
F-statistic: 281.5 on 1 and 73 DF,  p-value: < 2.2e-16
```

3.5.2 Interpréter les résultats

La ligne `(Intercept)` donne $\hat{\beta}_0 = 85.26$: le score de bien-être prédit lorsque le stress vaut zéro. Cette valeur n'a pas d'interprétation réelle, mais reste nécessaire pour définir la droite.

La ligne surlignée en bleu `stress` est la pente $\hat{\beta}_1 = -13.44$: c'est le **coefficient de proportionnalité** entre la variable indépendante et la variable critère. Pour une unité supplémentaire sur l'échelle de stress, le bien-être diminue de 13.44 points.

3.5.3 Faire une prédiction

La régression a estimé $\hat{\beta}_0$ et $\hat{\beta}_1$. L'équation du modèle ajusté s'écrit :

$$\widehat{\text{bien-être}} = -13.44 \times \text{stress} + 85.26$$

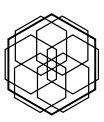
Cette équation représente notre modèle. Pour un patient dont le niveau de stress hypothétique est 3.5, on substitue $X = 3.5$:

$$\widehat{\text{bien-être}} = -13.44 \times 3.5 + 85.26 = 38.2$$

Le modèle prédit $\widehat{\text{bien-être}} = 38.2$ points. Cette valeur se situe entre la prédiction pour stress=3 (44.9 points) et celle pour stress=4 (31.5 points), ce qui illustre le traitement continu d'une variable indépendante pourtant discrète. Le modèle impose une pente constante de -13.44 points par unité, alors que les différences inter-niveaux observées varient entre -15 et -11.1 points.

3.6. Formuler une Réponse

Pour un patient dont le niveau de stress hypothétique est 3.5, le modèle prédit $\widehat{\text{bien-être}} =$ 38.2 points ($\hat{\beta}_1 = -13.44$, $\hat{\beta}_0 = 85.26$, $R^2 = 0.79$, $p < .001$). Cette prédiction repose su



l'hypothèse d'équidistance. La valeur 3.5 est une interpolation.

3.7. Evaluer la qualité du modèle

3.7.1 Statistique F

La valeur $F = 281.5$ avec une p-valeur $< .001$ indique que le modèle dans son ensemble est statistiquement significatif : le niveau de stress apporte une information réelle sur le bien-être.

3.7.2 R2

$R^2 = 0.794$: le niveau de stress explique 79.4 % de la variance du bien-être. Le reste, 20.6 %, ne correspond à des sources de variation que X n'explique pas.

3.7.3 Normalité

Centrage en zéro. La médiane des résidus (zone surlignée en rouge dans la sortie) vaut -0.36 : proche de zéro, aucun biais systématique détectable. Le modèle ne surestime ni ne sous-estime de façon systématique.

Symétrie. La queue gauche (Min → Q1) s'étend sur 5.7 points ; la queue droite (Q3 → Max) sur 14.2 points. Ces deux valeurs sont comparables : la distribution est approximativement symétrique.

3.7.3.1 QQ-plot et LOWESS

La Figure 6 confirme ces deux points visuellement.

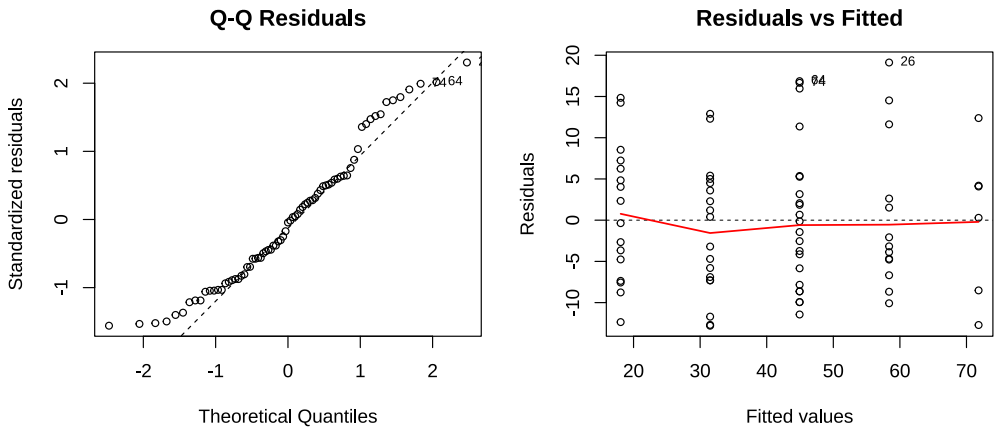


Figure 6: Gauche : Q-Q plot des résidus standardisés. Les points s'écartent de la droite théorique : la normalité des résidus est rejetée (W = 0.956, p = 0.01). Droite : résidus en fonction des valeurs ajustées. La courbe LOWESS longe la ligne zéro : les résidus sont bien centrés autour de zéro (médiane = -0.36).

4. Variable continue et variable dummy

4.1. Deux VIs, dont une dummy

Les sections précédentes utilisaient une seule variable indépendante X. Le modèle linéaire se généralise naturellement à plusieurs variables indépendantes. Avec deux variables X_1 et X_2, le modèle s'écrit :

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

Chaque coefficient $\hat{\beta}_k$ est un **coefficient partiel** : il exprime l'effet de X_k sur $\hat{Y}$ à valeur constante de toutes les autres variables indépendantes. C'est la notion de **contrôle statistique**.

Dans cette section, X_1 est une variable continue et X_2 est une **variable dummy** : elle ne prend que deux valeurs, 0 ou 1, pour coder une appartenance catégorielle. L'effet de la dummy est simple : elle déplace l'ordonnée à l'origine de $\hat{\beta}_2$ points pour le groupe codé 1, sans modifier la pente $\hat{\beta}_1$. Le résultat graphique est deux droites **parallèles** — même coefficient de proportionnalité entre X_1 et $\hat{Y}$, mais niveaux différents selon le groupe.

4.2. Un exemple concret

Un chercheur en psychologie quantitative recrute 90 patients répartis en deux groupes : 45 hommes ($X_2 = 0$) et 45 femmes ($X_2 = 1$). Pour chaque patient, il mesure le nombre de séances de psychothérapie reçues (variable indépendante continue X_1, de 1 à 19.8 séances) et le score de qualité de vie de 0 à 100 (variable critère Y). La situation est représentée sur la Figure 7.

4.2.1 Représentation Graphique

Le graphique montre deux nuages de points distincts et deux droites parallèles. La pente est identique dans les deux groupes — c'est $\hat{\beta}_1$, le coefficient de proportionnalité entre séances et qualité de vie. Le décalage vertical entre les deux droites est $\hat{\beta}_2$, l'effet du genre **à nombre de séances constant**. C'est ce que signifie contrôler pour X_1 : comparer les deux groupes à un même niveau de la variable indépendante continue.

4.3. Question de recherche

Quel score de qualité de vie le modèle prédit-il pour une patiente (femme) ayant reçu 25 séances de psychothérapie ?

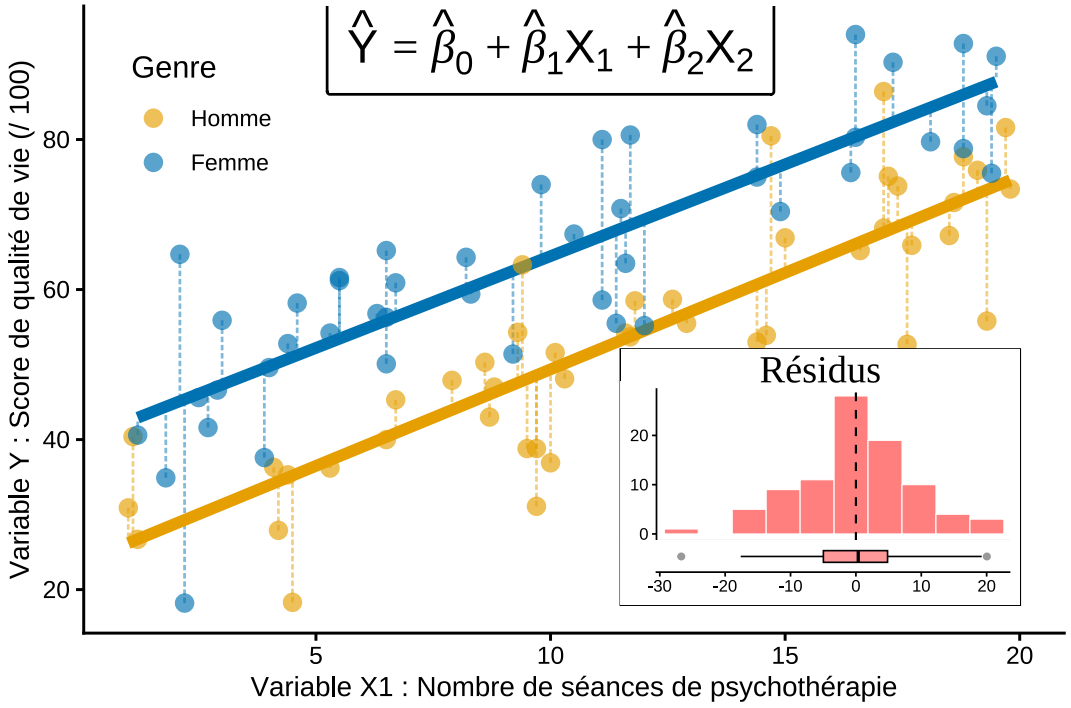


Figure 7: Relation entre le nombre de séances et le score de qualité de vie, selon le genre. Les deux droites parallèles illustrent l'effet additif de la variable dummy : même pente $\hat{\beta}_1$, ordonnées à l'origine décalées de $\hat{\beta}_2$ points. Les segments matérialisent les résidus $e_i = Y_i - \hat{Y}_i$.

4.3.1 En d'autre termes

La question consiste à estimer le score de qualité de vie d'une patiente ($X_2 = 1$) ayant reçu un nombre de séances ($X_1 = 25$) supérieur au domaine observé. Pour y répondre, il convient d'estimer les trois coefficients du modèle ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$) par régression multiple, puis d'évaluer $\widehat{\text{qualité de vie}} = \hat{\beta}_0 + \hat{\beta}_1 \times 25 + \hat{\beta}_2 \times 1$.

4.4. Méthode de résolution

1. **Identifier la variable que l'on cherche à expliquer.** C'est notre variable dépendante : le score de qualité de vie (Y).
2. **Identifier la ou les variables qui servent à expliquer.** Ce sont nos variables indépendantes : séances (X_1, continue) et genre (X_2, dummy 0/1) — chaque $\hat{\beta}_k$ est interprété toutes choses égales par ailleurs.
3. **La relation entre ces variables est-elle statistiquement significative ?** On teste $H_0 : \beta_1 = 0$ et $H_0 : \beta_2 = 0$ séparément ; le test F évalue leur contribution conjointe.
4. **Interpréter dans le cadre de la question.** On substitue $X_1 = 25$ et $X_2 = 1$ dans l'équation du modèle — extrapolation hors du domaine observé.

4.5. Commande R

```
lm_sect3 <- lm(qv ~ seances + genre, data = data_sect3)
summary(lm_sect3)
```

La syntaxe `qv ~ seances + genre` se lit : *qualité de vie en fonction de séances et de genre*. R traite `genre` (0/1) comme numérique — c'est le codage dummy. Le `+` indique un modèle additif sans interaction : les deux droites sont contraintes à être parallèles.

4.5.1 Lire la sortie logicielle

```
Call:
lm(formula = qv ~ seances + genre, data = data_sect3)

Residuals:
     Min       1Q   Median       3Q      Max
-26.7120  -4.9741   0.3561   4.8414  20.0388

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.3887     2.3136  10.541  < 2e-16 ***
seances       2.5082     0.1639  15.299  < 2e-16 ***
genre        15.0053     1.8714   8.018 4.54e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.74 on 87 degrees of freedom
Multiple R-squared:  0.7518,    Adjusted R-squared:  0.746
F-statistic: 131.7 on 2 and 87 DF,  p-value: < 2.2e-16
```

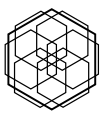
4.5.2 Interpréter les résultats

La sortie comporte maintenant **trois lignes de coefficients** au lieu de deux — c'est la nouveauté de la régression multiple.

`(Intercept)` : $\hat{\beta}_0 = 24.39$ — score de qualité de vie prédit pour un homme ($X_2 = 0$) avec zéro séance. Valeur hors du domaine observé, sans interprétation directe.

`seances` (surlignée) : $\hat{\beta}_1 = 2.51$ — c'est le **coefficient de proportionnalité** entre le nombre de séances et la qualité de vie, **à genre constant**. Pour une séance supplémentaire, le score de qualité de vie augmente de 2.51 points, indépendamment du genre du patient.

`genre` (surlignée) : $\hat{\beta}_2 = 15.01$ — c'est l'effet du genre, **à nombre de séances constant**. À nombre de séances égal, les femmes présentent en moyenne 15.01 points de qualité de vie de plus que les hommes. C'est le contrôle statistique en action : on compare les deux groupes à X_1 fixé.



4.5.3 Faire une prédiction

La régression a estimé $\hat{\beta}_0$, $\hat{\beta}_1$ et $\hat{\beta}_2$.

$$\widehat{\text{qualité de vie}} = 2.51 \times \text{séances} + 15.01 \times \text{genre} + 24.39$$

Pour une patiente ($X_2 = 1$) ayant reçu 25 séances ($X_1 = 25$) :

$$\widehat{\text{qualité de vie}} = 2.51 \times 25 + 15.01 \times 1 + 24.39 = 102.1$$

Le modèle prédit 102.1 points. Cette valeur dépasse 100 — le maximum de l'échelle. C'est une conséquence directe de l'extrapolation : en dehors du domaine observé (1–19.8 séances), le modèle linéaire n'est plus contraint par les données et peut produire des valeurs hors échelle.

4.6. Formuler une Réponse

Pour une patiente ayant reçu 25 séances, le modèle prédit $\widehat{\text{qualité de vie}} = 102.1$ points ($\hat{\beta}_1 = 2.51$, $\hat{\beta}_2 = 15.01$, $R^2 = 0.75$, $p < .001$). Cette valeur dépasse le maximum de l'échelle — elle est une extrapolation au-delà du domaine observé (1–19.8 séances).

4.7. Evaluer la qualité du modèle

4.7.1 Statistique F

La valeur $F = 131.7$ avec une p-valeur $< .001$ indique que le modèle dans son ensemble est statistiquement significatif : les deux variables indépendantes apportent ensemble une information réelle sur la qualité de vie.

4.7.2 R2

$R^2 = 0.752$: les séances et le genre expliquent ensemble 75.2 % de la variance de la qualité de vie. Le R^2 ajusté vaut 0.746 — légèrement inférieur, il pénalise l'ajout de variables indépendantes supplémentaires et constitue une mesure plus conservative de l'ajustement.

4.7.3 Normalité

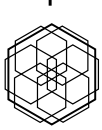
Deux points à examiner systématiquement.

Centrage en zéro. La médiane des résidus (zone surlignée en rouge dans la sortie) vaut 0.36 : proche de zéro, aucun biais systématique détectable. Le modèle ne surestime ni ne sous-estime de façon systématique.

Symétrie. La queue gauche (Min → Q1) s'étend sur 21.7 points ; la queue droite (Q3 → Max) sur 15.2 points. Ces deux valeurs sont comparables : la distribution est approximativement symétrique.

4.7.3.1 QQ-plot et LOWESS

La Figure 8 confirme ces deux points visuellement.



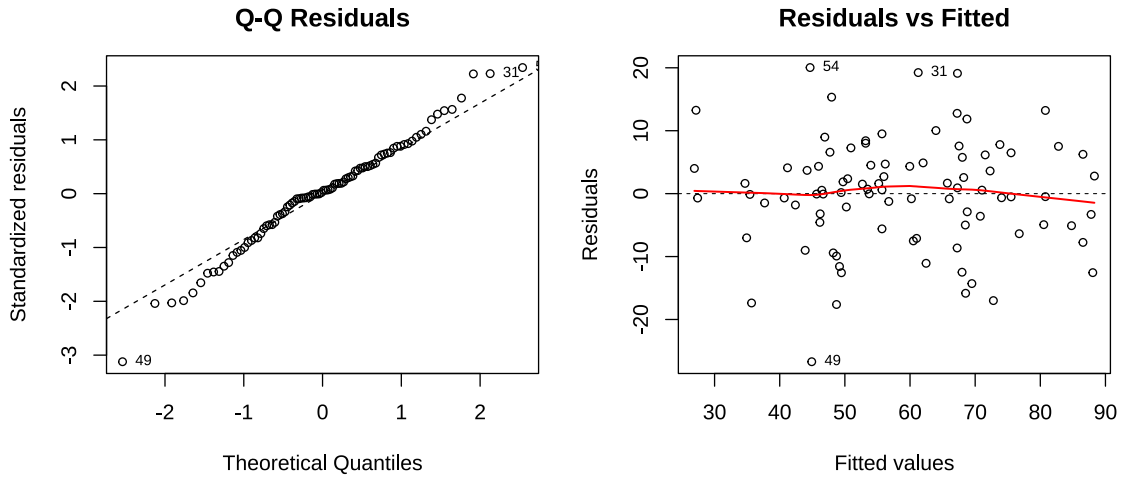


Figure 8: Gauche : Q-Q plot des résidus standardisés. Les points s'alignent sur la droite théorique : la normalité des résidus est confirmée (W = 0.985, p = 0.394). Droite : résidus en fonction des valeurs ajustées. La courbe LOWESS longe la ligne zéro : les résidus sont bien centrés autour de zéro (médiane = 0.36).

5. Conclusion

5.1. Valeurs observées et valeurs estimées

Ce document démontre comment interpréter les paramètre d'une droite de régression à l'aide des instrument statistique qui permette de quantifier l'incertitude sur les prédiction. Dans ce document, nous avons appris à bien faire la différence entre le modèle et la réalité, autrement dit, entre les valeur prédites et les valeur observée. Nous avons adopté un démarche systématique qui s'applique dans tous les cas de figure ou l'on utilise un modèle mathématique pour représenter une série d'observation.

5.2. La méthode en quatre étapes

Ce document démontre également comment aborder l'interprétation des paramètre d'un modèle de manière systématique. Cette méthode se résume en quatre étapes:

1. **Quelle est la variable que l'on cherche à expliquer ?** C'est la variable dépendante Y — le score à l'examen, le bien-être, la qualité de vie.
2. **Quelles sont les variables qui servent à expliquer ?** Ce sont les variables indépendantes X_k — continues, ordinales ou dummy. Chaque $\hat{\beta}_k$ est leur coefficient de proportionnalité avec Y.
3. **La relation est-elle statistiquement significative ?** On teste $H_0 : \beta_k = 0$. Si $p < .05$, la relation est significative et le coefficient peut être interprété.
4. **Quelle est la réponse à la question posée ?** On substitue les valeurs cibles dans l'équation du modèle pour obtenir une prédiction, en signalant toute extrapolation hors du domaine observé.

5.3. Qualité du modèle et normalité des résidus

Trois statistiques permettent d'évaluer la qualité du modèle:

Statistique F. Elle teste si le modèle dans son ensemble explique Y mieux que le hasard. Une p-valeur $< .05$ est le seuil minimal pour que les coefficients soient interprétables.

R^2. Il exprime la proportion de variance de Y expliquée par le modèle. Un R^2 modeste n'invalide pas le modèle — en psychologie quantitative, expliquer 20 à 30 % de la variance d'un comportement avec une seule variable explicative est souvent substantiel.

Normalité des résidus. C'est le premier bloc à lire dans la sortie logicielle, avant les coefficients. Deux vérifications rapides :

- *Centrage en zéro* : la médiane des résidus doit être proche de zéro. Un écart marqué signale un biais systématique de prédiction.
- *Symétrie* : comparer l'étendue de la queue gauche (Min → Q1) à celle de la queue droite (Q3 → Max). Une forte asymétrie suggère que la structure des données n'est pas bien capturée par le modèle linéaire.

Ces trois points de repère permettent de porter un regard informé sur la qualité du modèle et nous permette d'évaluer sa fiabilité avant de se livrer à une interprétation des paramètres.