Impact des valeurs aberrantes sur une régression linéaire

Université XYZ

February 15, 2025

1 Introduction

Les valeurs aberrantes sont des observations qui s'écartent fortement des autres données. Elles peuvent fausser les estimations des paramètres et conduire à des erreurs de prédiction importantes. Cette étude examine leur impact sur une régression linéaire expliquant le *Salaire* en fonction de l'*Expérience* et de l'*Éducation*.

2 Présentation des données

Nous utilisons un jeu de données hypothétique contenant 500 observations avec les variables suivantes :

- **Salaire** : revenu annuel en milliers d'euros.
- **Expérience** : nombre d'années d'expérience.
- **Éducation** : nombre d'années d'études après le secondaire.

Exemple de données :

Salaire (k€)	Expérience (années)	Éducation (années)
25	2	5
40	5	6
120	15	10
300	**4**	**3**
35	3	5

3 Détection des valeurs aberrantes

3.1 Méthodes graphiques

- **Boxplot** : permet d'identifier visuellement les valeurs extrêmes.
- **Diagramme de dispersion** : met en évidence les observations éloignées.

3.2 Méthodes statistiques

- **Distance interquartile (IQR)** :

$$\text{Valeurs aberrantes} \in [Q1 - 1.5IQR, Q3 + 1.5IQR] \tag{1}$$

- **Z-score** : une observation est aberrante si son Z-score est supérieur à 3.

4 Impact sur la régression

Nous estimons le modèle suivant :

$$Salaire_i = \beta_0 + \beta_1 Experience_i + \beta_2 Education_i + \varepsilon_i \tag{2}$$

4.1 Comparaison des méthodes de traitement

Méthode	Effet sur les coefficients	Avantages	
Suppression des outliers	Coefficients plus stables	Facile à interpréter	
Transformation logarithmique	Réduction des valeurs extrêmes	Utile pour asymétrie	Moins
Régression robuste	Estimations plus fiables	Gère bien les outliers	Plu

5 Application en R

Le code suivant permet de générer les données et d'appliquer différentes méthodes de détection et de correction des valeurs aberrantes.

5.1 Chargement des bibliothèques et génération des données

```r
install.packages(c("ggplot2", "dplyr", "MASS", "car"))
library(ggplot2)
library(dplyr)
library(MASS)
library(car)

set.seed(123)
n <- 500
data <- data.frame(
  Experience = runif(n, 1, 20),
  Education = sample(3:10, n, replace = TRUE),
  Salaire = 25 + 3 * runif(n, 1, 20) + 2 * sample(3:10, n,
     replace = TRUE) + rnorm(n, 0, 5)
)
data$Salaire[c(10, 50, 100)] <- c(150, 300, 500)
head(data)
```

Listing 1: Génération des données

5.2 Détection des valeurs aberrantes

```r
Q1 <- quantile(data$Salaire, 0.25)
Q3 <- quantile(data$Salaire, 0.75)
IQR <- Q3 - Q1
seuil_inf <- Q1 - 1.5 * IQR
seuil_sup <- Q3 + 1.5 * IQR
data$outlier_iqr <- ifelse(data$Salaire < seuil_inf | data$
   Salaire > seuil_sup, TRUE, FALSE)
```

Listing 2: Détection des valeurs aberrantes

5.3 Régression linéaire et robuste

```r
model1 <- lm(Salaire ~ Experience + Education, data = data)
summary(model1)
data_clean <- data[!data$outlier_iqr, ]
model2 <- lm(Salaire ~ Experience + Education, data = data_
   clean)
summary(model2)
model3 <- rlm(Salaire ~ Experience + Education, data = data)
summary(model3)
```

Listing 3: Régression et impact des outliers

6 Conclusion

Les valeurs aberrantes faussent les estimations et augmentent les erreurs de prédiction. La suppression ou l'utilisation de méthodes robustes permet d'améliorer la fiabilité du modèle.