S8_Bonus_residus

Table of contents

1 Contexte général : ce que fait un modèle

On observe, pour chaque individu i :

- une (ou plusieurs) variable(s) explicative(s) X_i
- une variable réponse Y_i

Exemple :

- $X_i = $ âge du patient (en années)
- $Y_i = $ pression artérielle systolique (en mmHg)

On veut résumer la relation entre X et Y par un modèle linéaire :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

avec :

- β_0 : intercept (valeur moyenne de Y quand $X = 0$)
- β_1 : pente (variation moyenne de Y pour une unité de X)
- ε_i : terme d'erreur théorique, la partie de Y_i non expliquée par le modèle

En pratique, on ne connaît pas β_0, β_1, on les estime à partir des données.

□

2 Observé, prédit, résidu : bien distinguer

Une fois le modèle ajusté, on obtient des estimations $\hat{\beta}_0$ et $\hat{\beta}_1$.

On peut alors calculer, pour chaque individu i :

Valeur observée :

$$y_i$$

Valeur prédite par le modèle :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Résidu :

$$e_i = y_i - \hat{y}_i$$

Interprétation :

- si $e_i > 0$: l'observation est au-dessus de ce que le modèle prédit
- si $e_i < 0$: l'observation est en dessous de ce que le modèle prédit
- si $e_i = 0$: le modèle tombe pile sur la valeur observée

Les résidus :

- sont dans la même unité que Y (mmHg, g/L, kg, etc.)
- mesurent l'erreur de prédiction du modèle pour chaque observation

□

3 Terme d'erreur ε_i vs résidu e_i

3.A Terme d'erreur ε_i (théorique)

Dans le modèle :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

on suppose que ε_i vérifie (dans le modèle linéaire gaussien) :

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

avec :

- $E(\varepsilon_i) = 0$
- Var$(\varepsilon_i) = \sigma^2$
- indépendance entre individus

ε_i est un objet théorique : on ne l'observe jamais directement.

3.B Résidu e_i (observé)

Le résidu se définit par :

$$e_i = y_i - \hat{y}_i$$

- il se calcule à partir des données et du modèle ajusté
- il dépend des estimations $\hat{\beta}_0$, $\hat{\beta}_1$
- c'est une estimation du terme d'erreur ε_i

Résumé :

- ε_i = erreur idéale dans le modèle théorique
- e_i = erreur observée (ce qu'on voit réellement dans les données)

□

4 Rôle des résidus dans la régression linéaire

4.A Méthode des moindres carrés

Dans la régression linéaire classique, les coefficients $\hat{\beta}_0, \hat{\beta}_1$ sont choisis pour minimiser la somme des carrés des résidus :

$$\text{SSR} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

L'algorithme cherche donc à rendre les résidus globalement aussi petits que possible (au sens des carrés).

4.B Hypothèses du modèle = hypothèses sur les résidus

Dans le modèle linéaire gaussien, on fait des hypothèses sur les ε_i.
En pratique, on vérifie ces hypothèses sur les résidus

$$e_i$$

:

4.B.1 Moyenne nulle

Théorique :

$$E(\varepsilon_i) = 0$$

En pratique : les résidus doivent osciller autour de 0.

4.B.2 Variance constante (homoscédasticité)

Théorique :

$$\text{Var}(\varepsilon_i) = \sigma^2 \quad \text{(indépendante de } X_i\text{)}$$

En pratique : la dispersion des résidus ne doit pas augmenter ou diminuer systématiquement avec les valeurs prédites.

4.B.3 Indépendance

Théorique : les ε_i sont indépendants.
En pratique : les résidus ne doivent pas montrer de structure dans le temps, par centre, par patient, etc.

4.B.4 Normalité (dans le cadre gaussien)

Théorique :

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

En pratique : l'histogramme et le QQ-plot des résidus doivent être compatibles avec une loi normale.

Toute l'analyse de diagnostic (graphiques de résidus, tests) repose sur le comportement des résidus.

□

5 Exemple concret en R avec un jeu de données fictif

On crée un jeu de données fictif simple :

- X = âge du patient, entre 20 et 80 ans
- Y = pression artérielle systolique (PAS, en mmHg)

On suppose que le « vrai » modèle (utilisé pour simuler les données) est :

$$Y_i = 90 + 0{,}6X_i + \varepsilon_i$$

avec :

$$\varepsilon_i \sim \mathcal{N}(0, 10^2)$$

```r
# Création d'un jeu de données fictif

n <- 100                              # nombre de patients
age <- runif(n, min = 20, max = 80)   # âges entre 20 et 80 ans

# Paramètres "vrais" du modèle de simulation
beta_0 <- 90      # intercept
beta_1 <- 0.6     # pente
sigma  <- 10      # écart-type de l'erreur

# Terme d'erreur théorique simulé
epsilon <- rnorm(n, mean = 0, sd = sigma)

# Valeur observée de PAS (simulée)
PAS <- beta_0 + beta_1 * age + epsilon

# Data frame final
```

```r
df <- data.frame(
    age = age,
    PAS = PAS
)

head(df)
```

```
       age      PAS
1 37.25465 114.8860
2 67.29831 130.0935
3 44.53862 116.2945
4 72.98104 147.4746
5 76.42804 133.5991
6 22.73339 118.8047
```

5.A Ajuster un modèle linéaire et récupérer les résidus

On ajuste le modèle linéaire :

$$\text{PAS}_i = \beta_0 + \beta_1 \cdot \text{age}_i + \varepsilon_i$$

avec `lm()`.

```r
# Ajustement du modèle linéaire
mod <- lm(PAS ~ age, data = df)

# Résumé du modèle
summary(mod)
```

```
Call:
lm(formula = PAS ~ age, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-22.3797  -6.1323  -0.1973   5.9633  22.1723

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 90.20984    3.00425   30.03   <2e-16 ***
age          0.58503    0.05697   10.27   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.693 on 98 degrees of freedom
Multiple R-squared:  0.5183,    Adjusted R-squared:  0.5134
F-statistic: 105.5 on 1 and 98 DF,  p-value: < 2.2e-16
```

On extrait ensuite :

- les valeurs prédites $\widehat{y}_i$
- les résidus $e_i = y_i - \widehat{y}_i$

```
# Valeurs prédites et résidus
df$y_hat  <- fitted(mod)   # valeurs prédites
df$residu <- resid(mod)    # résidus

head(df)
```

```
       age      PAS    y_hat      residu
1 37.25465 114.8860 112.0049  2.88111844
2 67.29831 130.0935 129.5812  0.51227617
3 44.53862 116.2945 116.2662  0.02828346
4 72.98104 147.4746 132.9058 14.56884788
5 76.42804 133.5991 134.9224 -1.32327681
6 22.73339 118.8047 103.5095 15.29522857
```

Dans ce tableau :

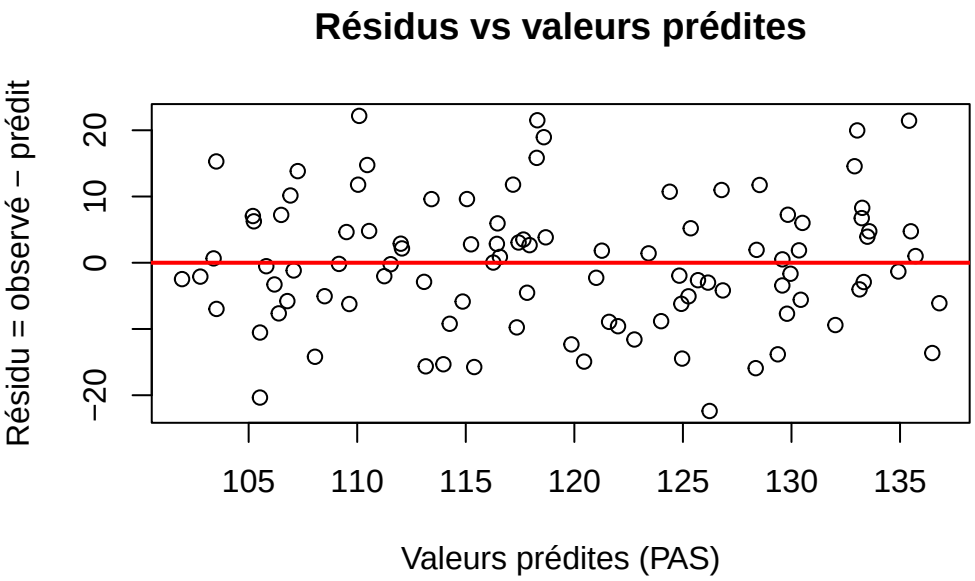
- PAS = y_i (valeur observée)
- y_hat = $\widehat{y}_i$ (valeur prédite par le modèle)
- residu = $e_i = y_i - \widehat{y}_i$

5.A.1 Graphe des résidus vs valeurs prédites

On trace les résidus en fonction des valeurs prédites :

- on souhaite les voir centrés autour de 0
- sans forme particulière (pas de « V », pas de structure claire)

```
plot(
    x = df$y_hat,
    y = df$residu,
    xlab = "Valeurs prédites (PAS)",
    ylab = "Résidu = observé - prédit",
    main = "Résidus vs valeurs prédites"
)
abline(h = 0, col = "red", lwd = 2)
```



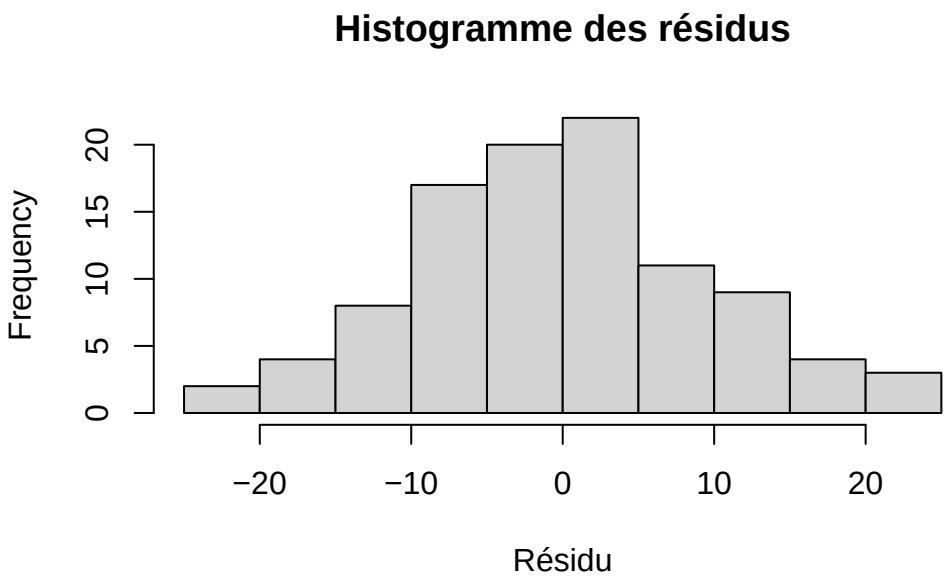
Interprétation :

- points au-dessus de la ligne rouge : PAS observée plus élevée que prévu
- points en dessous : PAS observée plus basse que prévu
- si le modèle est correct, le nuage doit être « diffus » autour de la ligne à 0, sans structure évidente

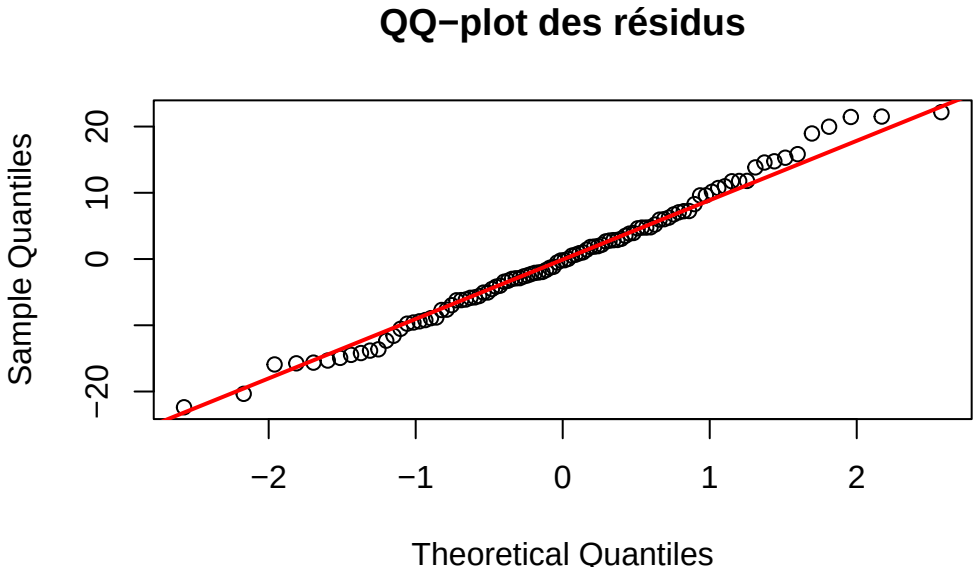
5.B Histogramme et QQ-plot des résidus

On regarde la distribution des résidus.

```
hist(
    df$residu,
    breaks = 15,
    main = "Histogramme des résidus",
    xlab  = "Résidu"
)
```




```
#| label: residuals-qqplot
#| echo: true
qqnorm(df$residu, main = "QQ-plot des résidus")
qqline(df$residu, col = "red", lwd = 2)
```



Idéalement :

- l'histogramme est approximativement symétrique autour de 0
- le QQ-plot montre les points proches de la droite : compatible avec une loi normale

□

6 Lien avec les modèles plus complexes

Dans des modèles plus complexes (GLM, modèles linéaires mixtes, etc.), on garde la même idée de base :

$$\text{résidu} = \text{valeur observée} - \text{valeur prédite (ou espérée) par le modèle}$$

- Dans un GLM (logistique, Poisson), la définition est adaptée pour tenir compte du fait que la variance dépend de la moyenne (résidus de Pearson, de déviance, etc.).
- Dans un modèle linéaire mixte, on définit les résidus après avoir pris en compte :
  - les effets fixes (âge, sexe, traitement…)
  - les effets aléatoires (patient, centre, etc.)

Mais le sens reste le même : les résidus mesurent ce qui n'est pas expliqué par le modèle pour chaque observation.

□

7 Résumé

Pour chaque observation i :

$$e_i = y_i - \hat{y}_i$$

- un résidu, c'est la différence entre la valeur observée et la valeur prédite
- les résidus sont au cœur :
  - de l'estimation (moindres carrés minimisent la somme des carrés des résidus)
  - du diagnostic du modèle (vérifier les hypothèses sur les erreurs)

Le terme d'erreur ε_i est théorique,
le résidu e_i est observable et c'est lui qu'on regarde dans les sorties et les graphiques.