Auto-évaluation 12 – Régression linéaire simple et corrélation

Solutions

1. Importez le fichier `Pression_sanguine.txt` qui présente les données relatives à la pression sanguine en mm de Hg chez des patients de différents âges.

Réponse :

Nous importons le jeu de données :

```
> sang <- read.table(file = "Pression_sanguine.txt", header = TRUE)
> head(sang)
  Age Pression
1  30      108
2  30      110
3  30      106
4  40      125
5  40      120
6  40      118
> str(sang)
'data.frame':        20 obs. of  2 variables:
 $ Age     : int  30 30 30 40 40 40 40 50 50 50 ...
 $ Pression: int  108 110 106 125 120 118 119 132 137 134 ...
```

a) Effectuez une régression linéaire simple afin de déterminer l'effet de l'âge sur la pression sanguine.

Réponse :

```
> m.sang <- lm(Pression ~ Age, data = sang)
> summary(m.sang)
Call:
lm(formula = Pression ~ Age, data = sang)

Residuals:
   Min     1Q Median     3Q    Max
-4.005 -1.919 -0.442  2.026  4.089

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  68.7849     2.2161      31   <2e-16 ***
Age           1.3031     0.0408      32   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.57 on 18 degrees of freedom
Multiple R-squared:  0.983,        Adjusted R-squared:  0.982
F-statistic: 1.02e+03 on 1 and 18 DF,  p-value: <2e-16
```

b. Vérifiez les suppositions nécessaires à la régression linéaire simple. Apportez des transformations si nécessaire. La régression linéaire est-elle justifiée ici ?

Réponse :

Les suppositions de linéarité, d'homogénéité des variances et de normalité des résidus sont respectées (fig. 1). La régression est donc appropriée avec ces données.

```
> par(mfrow = c(2, 2))
> ##linéarité
> plot(sang$Pression ~ sang$Age,
      main = "Linéarité")
> ##homogénéité des variances
> plot(rstudent(m.sang) ~ fitted(m.sang),
      main = "Homoscédasticité")
> ##normalité des résidus
> qqnorm(rstudent(m.sang),
        main = "Normalité des résidus")
> qqline(rstudent(m.sang))
```

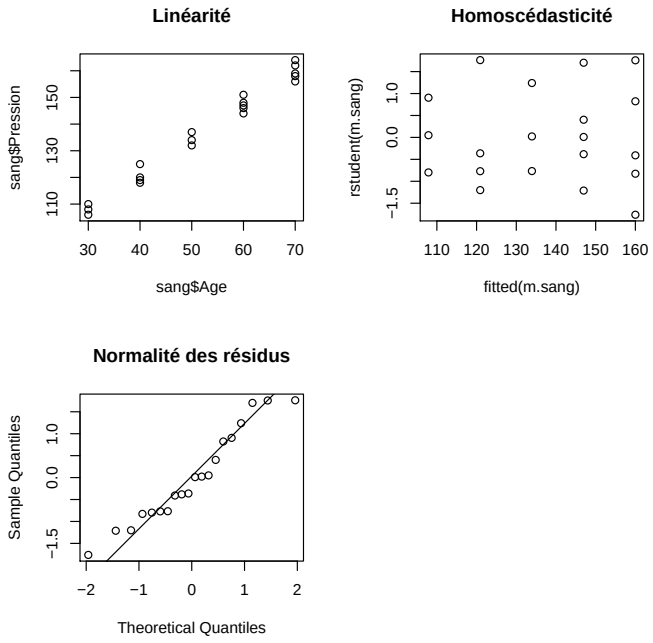

FIGURE 1 – Diagnostics de la régression linéaire effectuée sur les données de pression sanguine en fonction de l'âge de patients.

c. Interprétez les résultats et commentez la valeur des coefficients ainsi que le pouvoir prédictif de la régression.

Réponse :

```
> summary(m.sang)

Call:
lm(formula = Pression ~ Age, data = sang)


Residuals:
   Min     1Q Median     3Q    Max
-4.005 -1.919 -0.442  2.026  4.089


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  68.7849     2.2161      31   <2e-16 ***
Age           1.3031     0.0408      32   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 2.57 on 18 degrees of freedom
Multiple R-squared:  0.983,        Adjusted R-squared:  0.982
F-statistic: 1.02e+03 on 1 and 18 DF,  p-value: <2e-16
```

Nous remarquons que l'effet de l'âge est très positif avec une pente de 1.3 et une excellente précision (erreur-type très faible relative à la valeur du coefficient de la pente : 0.04). De plus, le R^2 indique que l'âge explique 98.3 % de la variabilité de la pression sanguine. La régression décrit très bien les données.

d. Utilisez l'équation de régression pour prédire la pression sanguine d'un patient de 55 ans. Construisez un intervalle de confiance autour de la prédiction.

Réponse :

L'équation de régression est :

$$\begin{aligned} \hat{y}_i &= 68.78 + 1.3 \cdot Age_i \\ \hat{y}_i &= 68.78 + 1.3 \cdot 55 \\ \hat{y}_i &= 140.46 \end{aligned}$$

Nous pouvons calculer la valeur prédite et un intervalle de confiance autour de cette valeur :

```
> ##jeu de données à partir duquel on fait des prédictions
> jeu.pred <- data.frame(Age = 55)
> ##on effectue la prédiction avec SE
> pred <- predict(m.sang, newdata = jeu.pred, se.fit = TRUE)
> ##ajout à jeu.pred
> jeu.pred$fit <- pred$fit
> jeu.pred$se.fit <- pred$se.fit
> ##on calcule IC à 95%
> jeu.pred$inf95 <- jeu.pred$fit +
    qt(p = 0.025, df = m.sang$df.residual) * jeu.pred$se.fit
> jeu.pred$sup95 <- jeu.pred$fit -
    qt(p = 0.025, df = m.sang$df.residual) * jeu.pred$se.fit
> jeu.pred
  Age    fit  se.fit  inf95  sup95
1  55 140.46 0.58369 139.23 141.68
```

Nous concluons qu'un patient âgé de 55 ans aura une pression de 140.5 mm de Hg avec un intervalle de confiance à 95 % : (139.2, 141.7).

e. Présentez la droite de régression sous forme graphique. Ajoutez les limites de confiance autour de la droite.

Réponse :

Calculons les valeurs prédites et leurs intervalles de confiance respectifs pour chacune des valeurs observées d'âge :

```
> ##jeu de données à partir duquel on fait des prédictions
> jeu.pred <- data.frame(Age = seq(from = min(sang$Age),
                                   to = max(sang$Age),
                                   by = 1))
> ##on effectue la prédiction avec SE
> pred <- predict(m.sang, newdata = jeu.pred, se.fit = TRUE)
> ##ajout à jeu.pred
> jeu.pred$fit <- pred$fit
> jeu.pred$se.fit <- pred$se.fit
> ##on calcule IC à 95%
> jeu.pred$inf95 <- jeu.pred$fit +
    qt(p = 0.025, df = m.sang$df.residual) * jeu.pred$se.fit
> jeu.pred$sup95 <- jeu.pred$fit -
    qt(p = 0.025, df = m.sang$df.residual) * jeu.pred$se.fit
```

La figure 2 illustre la droite de prédiction ainsi que les intervalles de confiance autour des valeurs prédites. Nous remarquons l'excellente précision dans les prédictions à partir de la régression, tel qu'indiqué par les intervalles de confiance très étroits autour des valeurs prédites.

```
> ##graphique vide
> par(cex = 1.2)
> plot(jeu.pred$fit ~ jeu.pred$Age, type = "n",
       ylab = "Pression sanguine (mm de Hg)",
       xlab = "Âge (années)",
       ylim = c(min(jeu.pred$inf95), max(jeu.pred$sup95)))
> ##ajoute droite
> lines(y = jeu.pred$fit, x = jeu.pred$Age)
> ##ajoute limites de confiance
> lines(y = jeu.pred$inf95, x = jeu.pred$Age, lty = "dotted")
> lines(y = jeu.pred$sup95, x = jeu.pred$Age, lty = "dotted")
```

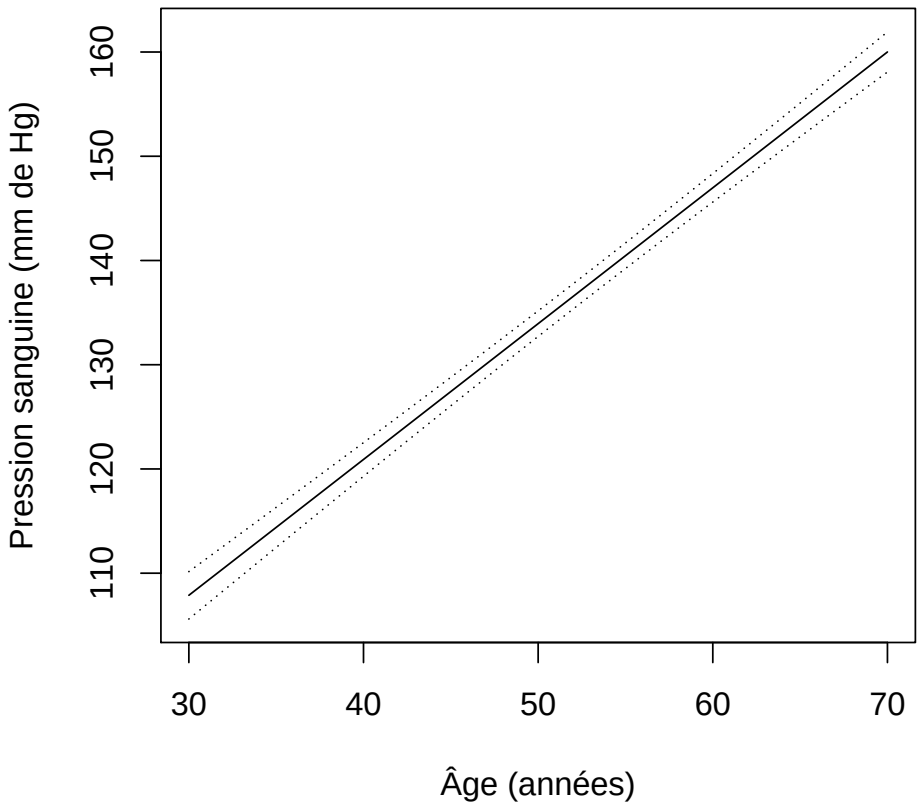

FIGURE 2 – Pression sanguine en fonction de l'âge des patients. Les pointillés représentent les limites de confiance à 95 % autour des valeurs prédites obtenues à partir de la régression linéaire.

2. Dans une étude d'observation, on s'intéresse au nombre de visites effectuées par des insectes pollinisateurs sur des rosiers sauvages (*Rosa blanda*). Pour ce faire, nous sélectionnons aléatoirement 50 rosiers dans notre aire d'étude et effectuons des observations pendant 3 heures sur chaque rosier en notant le nombre de visites par des insectes pollinisateurs. Lors de cette étude d'observation, nous avons également établi un quadrat de 1 m x 1 m centré sur chaque rosier et nous avons mesuré la surface du quadrat couverte par des champignons (exprimée en %).

a) Quelle analyse sera la plus approriée pour décrire la relation entre le nombre de visites par des insectes pollinisateurs sur des rosiers et le pourcentage de recouvrement de champignons sous ces rosiers ? Justifiez votre réponse.

Réponse :

Une corrélation serait plus appropriée pour décrire la relation entre les deux variables. On ne s'attend pas à une relation de cause à effet entre le nombre de visites par des insectes pollinateurs sur des rosiers et le recouvrement de champignon sous ces rosiers. Il est peu probable que la surface couverte par des champignons influence directement la fréquence de visites d'insectes pollinisateurs à des rosiers. Une plus grande humidité pourrait améliorer les conditions de croissance pour les champignons. Toutefois, une trop grande humidité pourrait potentiellement nuire à la croissance des rosiers, entraînant une production plus faible de fleurs. Ceci pourrait ensuite se traduire **indirectement** par une diminution du nombre de visites par les insectes pollinisateurs.