Regression - Guide R

Irene Gannaz

26 février 2020

Ajustement

```
x1 <- rnorm(20)
x2 <- rnorm(20,2)
y <- -1+2*x1+0.01*x2 + rnorm(20,0,0.05)

modele <- lm(y~x1+x2) #ajuste le modele
summary(modele) # realise tous les tests utiles sur le modele
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.057062 -0.035054 -0.004125  0.026651  0.134748
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.05869    0.02633 -40.210   <2e-16 ***
## x1           1.99800    0.01359 147.042   <2e-16 ***
## x2           0.03117    0.01129   2.761   0.0134 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05002 on 17 degrees of freedom
## Multiple R-squared:  0.9992, Adjusted R-squared:  0.9991
## F-statistic: 1.109e+04 on 2 and 17 DF,  p-value: < 2.2e-16
```

- **Intercept** = constante
- **Estimate** = valeurs des coefficients correspondant à chaque variable
- **Pr(>|t|)** = p-valeur du test (H0) coefficient=0 contre (H1) coefficient $\neq 0$. 't' car c'est un test de Student. Les * dans la marge servent à repérer les valeurs significatives
- **Residual standard error** = le $\hat{\sigma}$ du cours
- **Multiple R-squared** = le R^2 du cours en dimension 2, En plus grande dimension, $R^2 = \frac{\|\hat{y}_{\cdot} - \overline{y}_{\cdot}\|}{\|y_{\cdot} - \overline{y}_{\cdot}\|}$ est le ratio entre la variance expliquée par le modèle et la variance des données : si le ratio est proche de 1 cela signifie que les observations s'éloignent peu du modèle.
- **F-statistic** = test de Fisher de pertinence du modèle. Un modèle pertinent est un modèle tel que R^2 est significativement supérieur à 0. $F = \frac{R^2}{1-R^2}\frac{N-K}{K-1}$ avec K le nombre de variables dans le modèle (hors constante). La p-valeur donne la probabilité de se tromper si on affirme que le modèle n'est pas pertinent.

S'il n'y a qu'une seule variable dans le modèle, tester (H0) coefficient=0 contre (H1) coefficient $\neq 0$ ou faire le test de pertinence global de Fisher sont parfaitement équivalents.

Etude des résidus

Il faut vérifier les hypothèses sur les résidus. Il y en a 4 :

- loi normale
- espérance nulle
- variance constante
- indépendance

Si l'une de ces hypothèse est remise en cause, alors le modèle n'est plus valable (aucun des tests ci-dessus n'est valable et l'ajustement pas les moindres carrés est également discutable).

L'étude de ces hypothèses se fait par

- une étude graphique
- des tests

Des tests pouvant être utilisés pour les hypothèses sont :

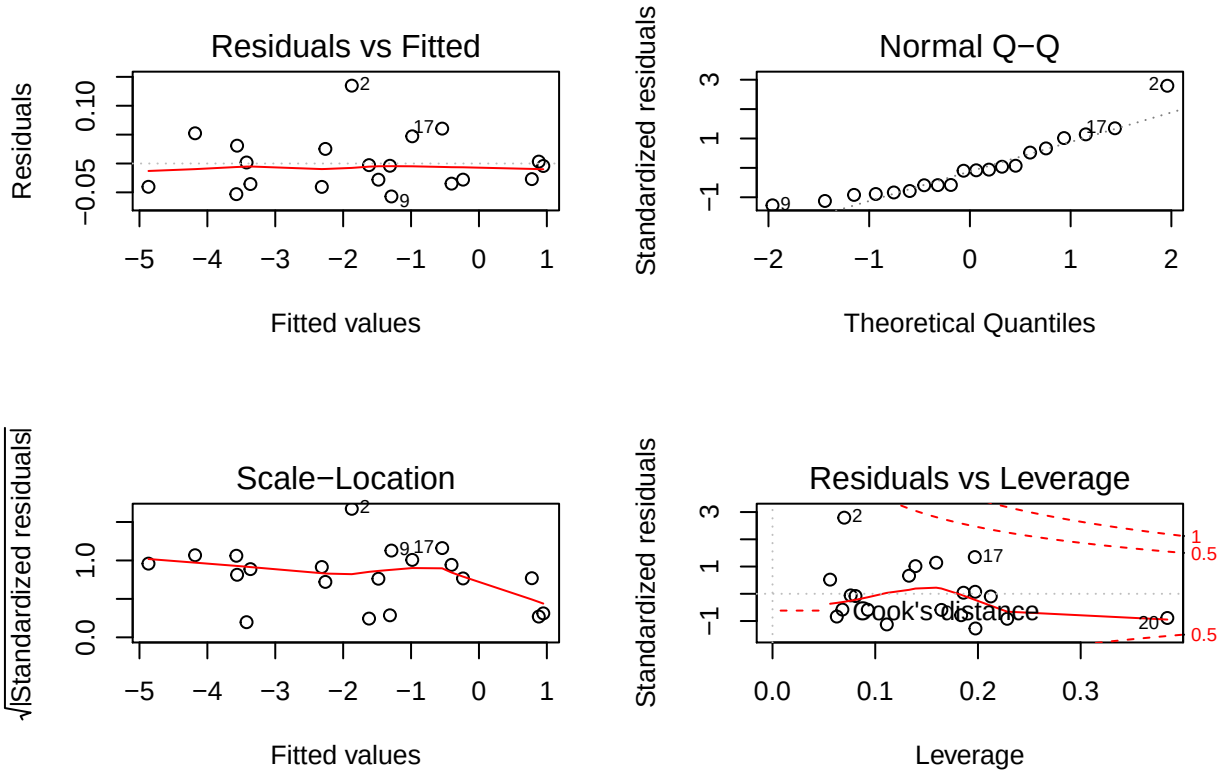
- loi normale : Shapiro-Wilk, `shapiro.test`
- espérance nulle : test Rainbow `raintest` dans le package `lmtest`
- variance constante : test de Breush-Pagan `bptest` dans le package `lmtest`
- indépendance : test de Durbin-Watson `dwtest` dans le package `lmtest`

Dans chaque cas les tests doivent être appliqués sur les résidus du modèle, et une p-valeur petite signifie un rejet de l'hypothèse, donc du modèle de régression.

Reprenons l'exemple. Les résidus sont données par

Les principaux graphiques sur les résidus peuvent être obtenus à l'aide de

```
par(mfrow=c(2,2))
plot(modele)
```



- **Residuals vs Fitted :**
  - Si on observe une tendance trop marquée des points sur le graphique, cela signifie que l'espérance des résidus n'est pas nulle, mais qu'elle est positive sur certaines sections et négatives sur d'autres. Ce problème peut souvent être corrigé avec un changement de variable. On reste assez "tolérant" sur les tendances et il faut qu'elles soient marquées pour rejeter le modèle.
  - Si on observe que le nuage de point s'écarte (forme de trompette) la variance des résidus n'est pas constante. On dit que les résidus sont hétéroscédastiques.
- **Normal Q-Q :** Compare la distribution des résidus à une loi normale. En abscisse, les quantiles empiriques des résidus et en ordonnée les quantiles de la loi normale, avec estimation des paramètres sur les résidus. Si les distribution sont identiques ou presque alors l'ensemble des points sont sur la diagonale. Sinon on observera la plupart du temps des deviation aux extremité ce qui sous-entend que les queues de distribution sont différentes.
- **Scale location :** Idem que Residuals vs Fitted mais avec des résidus normalisés.
- **Residuals vs Leverage :** Montre l'influence des echantillons (plus un point est à droite et plus il en a). Si un point est un outliers il apparaitra trés éloigné des autres et en dehors des bornes par rapport à la distance de Cook. Ces bornes sont représentées par des lignes rouge en pointillé. Il faut reprendre le modèle en enlevant les points concernés s'il y en a pour vérifier qu'ils ne déterminent pas le modèle à eux tout seuls...