Modèles Linéaires

Devoir Maison - Correction
Licence 3 MIASHS (2023 - 2024)

Guillaume Metzler

Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Résumé

Le présent travail est à rendre, sur feuille, pour le jeudi 7 mars 2024.
Il vous permettra de vous préparer pour votre premier contrôle du 14 mars 2024.

y	1	4	2	4.5	2.2	4.2
x_1	-2	3	-1	1	-4	3
x_2	0	2	0	0	0	-2

TABLE 1 – Tableau des données : y représente les valeurs de la variable dépendante, x_1 et x_2 les valeurs des covariables indépendantes.

Etude d'un problème de régression

On considère le jeu de données présenté en Table 1.

On considère le modèle de régression multiple suivant

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

1. Rappeler les hypothèses du modèle linéaire gaussien.

   Les hypothèses du modèle linéaire gaussien sont les suivantes :

   - La variable Y est supposée gaussien de moyenne $\mathbf{X}\boldsymbol{\beta}$ et de variance inconnue σ^2. D'ailleurs seule Y sera aléatoire et X sera déterministe par la suite,
   - les observations $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ sont supposées $i.i.d.$,
   - les erreurs du modèles seront supposées normalement distribuées et indépendantes. Plus précisément $\boldsymbol{\varepsilon} \underset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, où σ^2 est inconnue.

2. Dans le cas présent, définir les objets $\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}$ et $\boldsymbol{\varepsilon}$ ainsi que leurs dimensions.

   Le vecteur $\mathbf{y}$ est un vecteur de $\mathbb{R}^6$, la matrice design $\mathbf{X}$ est une matrice de $\mathscr{M}_{6,3}$, $\boldsymbol{\varepsilon} \in \mathbb{R}^6$ et $\boldsymbol{\beta} \in \mathbb{R}^3$.

3. On considère le problème d'optimisation

   $$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2.$$

   (a) Donner l'expression littérale de $\hat{\boldsymbol{\beta}}$, l'estimateur de $\boldsymbol{\beta}$.

   L'estimateur $\hat{\boldsymbol{\beta}}$ des moindres carrés ordinaires est donné par

   $$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}.$$

(b) Calculer sa valeur numérique, à la main. On donnera un résultat précis au centième près, soit deux chiffres après la virgule.

Nous effectuerons les calculs sous R pour cette correction

```
n = 6
p = 2
# Création de la matrice de design

x0 = rep(1,6)
x1 = c(-2,3,-1,1,-4,3)
x2 = c(0,2,0,0,0,-2)

X = cbind(x0,x1,x2)

# Création du vecteur des valeurs réponses

y = matrix(c(1,4,2,4.5,2.2,4.2), ncol =1)

# Valeur de l'estimateur beta_hat

beta_hat = solve(t(X)%*%X)%*%t(X)%*%y
beta_hat
```

```
##          [,1]
## x0  2.983333
## x1  0.407500
## x2 -0.050000
```

4. A l'aide du logiciel R.

(a) Tester la significativité des paramètres du modèle. On prendra le soin de définir le test employé, la définition et la valeur de la statistique de test ainsi que le distribution associée à ce test.

Pour tester la significativité des paramètres β_j, nous formulons les hypothèses suivantes :

$$H_0 : \beta_j = 0 \quad \text{vs} \quad \beta_j \neq 0.$$

C'est un test statistique biltéral qui repose sur la loi de Student ! Plus préciséement, une loi de Student à $n - p - 1$ degrés de liberté soit, 3 degrés de

liberté dans le cas présent, où p représente le nombre de variables dans le modèle.

$$t_{test} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{Var[\hat{\beta}_j]}}.$$

La variance des estimateurs est donnée par

$$Var[\hat{\beta}] = \sigma^2 (X^\top X)^{-1},$$

où σ^2 est inconnue et dont une estimation est donnée par

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

Il ne nous reste plus qu'à effectuer les applications numériques

```
data = data.frame(X[,c(2,3)],y)
data
```

```
##   x1 x2   y
## 1 -2  0 1.0
## 2  3  2 4.0
## 3 -1  0 2.0
## 4  1  0 4.5
## 5 -4  0 2.2
## 6  3 -2 4.2
```

```
mymodel = lm(y~., data= data)
beta_hat = mymodel$coefficients

# Extraire les résidus
res = mymodel$residuals

# Estimation de la variance des résidus
sigma2_hat = (1/(n-p-1))*sum(res^2)

# Matrice de variance de l'estimateur
var_hat = sigma2_hat * solve(t(X)%*%X)

# On peut extraire l'écart_type des différents coefficients
s_beta_hat = sqrt(diag(var_hat))


# Calcul des valeurs de la statistique de test
```

```r
t_test = beta_hat / s_beta_hat

# Test de student
p_val = 2*(1-pt(abs(t_test),n-p-1))
p_val
```

```
## (Intercept)          x1          x2
## 0.007048412 0.102017396 0.906297903
```

(b) Comment tester si le modèle est globalement significatif ? Définir le test, calculer sa valeur et conclure quant à la significativité du modèle.

Pour tester si le modèle est globalement significatif, on procède au test suivant

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_p = 0 \quad \text{vs} \quad H_1 : \exists j \in [[1, p]] \text{ t.q. } \beta_j \neq 0.$$

C'est un test qui repose sur la loi de Fisher. Plus précisément sur la loi de Fisher à p et $n - p - 1$ degrés de libertés.
On rappelle la formule de décomposition de la variance

$$SCT = SCE + SCR,$$

où :

i) SCT : *somme des carrés totaux* qui est égale à $\sum_{i=1}^{n}(y_i - \bar{y})^2$.
ii) SCE : *somme des carrés expliqués* qui est égale à $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$.
iii) SCR : *somme des carrés résiduels* qui est égale à $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

On peut calculer des variances qui sont débiaisées. Pour la variance résiduelle : $MSR = \dfrac{SCR}{n-p-1}$. Pour la variance expliquée par le modèle : $MSE = \dfrac{SCE}{p}$.
La statistique de test de Fisher est donnée par

$$F_{test} = \frac{\frac{SCE}{p}}{\frac{SCR}{n-p-1}} = \frac{MSE}{MSR},$$

qui va suivre une distribution de Student à p et $n - p - 1$ degrés de liberté.

Calculons toutes ces quantités

```r
# Test de Fisher

SCR = sum(res^2)
SCT = sum((data$y - mean(data$y))^2)
SCE = SCT - SCR

F_test = (SCE/p) / (SCR/(n-p-1))

# Calcul de la p-value
p_val = 1-pf(F_test, p, n-p-1)
p_val
```

```
## [1] 0.2114746
```

5. On s'intéresse maintenant à la qualité du modèle.

   (a) Calculer le coefficient de détermination R^2 du modèle de régression construit.

   On rappelle que l'on a la relation suivante

$$R^2 = 1 - \frac{\frac{SCR}{n}}{\frac{SCT}{n}} = 1 - \frac{SCR}{SCT} = \frac{SCE}{SCT}.$$

   On prend ici en compte les estimateurs biaisés des variances.

```r
# Calcul du R^2

R_squared = 1 - SCT/SCT
R_squared
```

```
## [1] 0
```

   (b) En déduire la valeur du coefficient de détermination ajusté R^2_{aj}.

```r
# Calcul du R^2 ajusté

R_adj_squared = 1 - (SCR/SCT) * (n-1)/(n-p-1)
R_adj_squared
```

```
## [1] 0.40841
```

(c) Calculer le BIC du modèle.

On rappelle que le BIC d'un modèle de régression est donnée par la relation

$$BIC = n(\ln(2\pi) + 1) + n \ln\left(\frac{SCR}{n}\right) + (p+2)\ln(n).$$

```
# Calcul du BIC du modèle

BIC(mymodel)

## [1] 21.23849

n*log(2*pi)+n + n*log(SCR/n) + (p+2)*log(n)

## [1] 21.23849
```

6. Retour sur un modèle linéaire simple. On considère le modèle suivant, qui n'utilise que la covariable $\mathbf{x}_1$, *i.e.*,

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \boldsymbol{\varepsilon}.$$

(a) Effectuer la régression linéaire simple, à la main, et déterminer le coefficient de détermination ajusté de modèle.

On se contentera de faire les calculs à l'aide de R.

```
# Modèle linéaire simple

my_simple_model = lm(y ~ x1,data = data)
summary(my_simple_model)

##
## Call:
## lm(formula = y ~ x1, data = data)
##
## Residuals:
##         1         2         3         4         5         6
## -1.168333 -0.205833 -0.575833  1.109167  0.846667 -0.005833
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.9833     0.3919   7.612   0.0016 **
## x1            0.4075     0.1518   2.685   0.0550 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.96 on 4 degrees of freedom
## Multiple R-squared:  0.6431,Adjusted R-squared:  0.5539
## F-statistic: 7.208 on 1 and 4 DF,  p-value: 0.05496
```

```
# On peut extraire directement le coefficient
# de détermination ajusté via la commande suivante

summary(my_simple_model)$adj.r.squared
```

```
## [1] 0.553887
```

(b) Comparer le résultat obtenu au modèle multiple et dire qu'elle est le meilleur modèle.

```
        # Modèle linéaire simple

        summary(my_simple_model)$adj.r.squared
```

```
## [1] 0.553887
```

```
        # Modèle linéaire multiple

        summary(mymodel)$adj.r.squared
```

```
## [1] 0.40841
```

Le modèle linéaire multiple présente un R^2_{aj} plus faible que le modèle linéaire simple. Le modèle multiple est donc moins intéressant que le modèle simple.