



## Modèles Linéaires

Contrôle Continu - Première Epreuve - 1h30

Correction

Licence 3 MIASHS (2023-2024)

Guillaume Metzler, Francesco Amato & Alejandro Rivera

Université de Lyon, Université Lumière Lyon 2

Laboratoire ERIC UR 3083, Lyon, France

[guillaume.metzler@univ-lyon2.fr](mailto:guillaume.metzler@univ-lyon2.fr); [francesco.amato@univ-lyon2.fr](mailto:francesco.amato@univ-lyon2.fr);

[alejandro.rivera@univ-lyon2.fr](mailto:alejandro.rivera@univ-lyon2.fr)

L'usage de matériel électronique (téléphone, ordinateur), des notes de cours ou des notes personnelles n'est pas autorisé pendant la durée de cette épreuve

En revanche, l'usage de la calculatrice est autorisé.

### Résumé

Les exercices de cet examen sont indépendants. Comme dans tout examen, une grande importance sera accordée à la qualité de la rédaction des réponses. Une réponse avec un résultat sans justification ne pourra prétendre à l'obtention de la totalité des points.

## Exercice 1

Les questions de ce premier exercice sont des questions de cours et des questions d'interprétations des résultats sur le modèle de régression linéaire simple et multiple.

1. On considère le modèle de régression linéaire simple

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- (a) Rappeler les hypothèses du modèle linéaire gaussien.

Le modèle linéaire gaussien suppose que les données  $\{(y_i, x_i)\}_{i=1}^n$  sont toutes indépendante et identiquement distribuées. On suppose également que la variable aléatoire  $Y$  suit une loi gaussienne de moyenne  $\beta_0 + \beta_1 X$  et de variance inconnue  $\sigma^2$ . Enfin, les erreurs du modèles sont supposées indépendantes et identiquement distribuées de loi normale de moyenne 0 et de variance inconnue  $\sigma^2$ .

- (b) Supposons que l'on dispose d'un échantillon  $S = \{(y_i, x_i)\}_{i=1}^m$ . Enoncer le problème d'optimisation que l'on cherche à résoudre pour estimer les paramètres  $\beta_0$  et  $\beta_1$  de la régression ?

Dans le modèle de régression linéaire simple, la quantité que l'on cherche à minimiser peut s'écrire sous la forme

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

De façon équivalente, sous forme matricielle, on a

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

où  $\boldsymbol{\beta}$  est le vecteur des paramètres,  $\mathbf{y}$  le vecteur des valeurs de la variable dépendante  $Y$  et  $\mathbf{X} \in \mathbb{R}^{n \times 2}$  est notre matrice de design.

- (c) Donner les expressions littérales de  $\beta_0$  et  $\beta_1$  en fonction de l'espérance, de la variance de  $X$  et de la covariance des variables aléatoires  $X$  et  $Y$ .

Les solutions de notre problème d'optimisation de la question précédente sont données par

$$\hat{\beta}_1 = \frac{\text{Cov}[X, Y]}{\text{Var}[X]} \quad \text{et} \quad \hat{\beta}_0 = \mathbb{E}[Y] - \hat{\beta}_1 \mathbb{E}[X].$$

- (d) A quelle autre quantité statistique est liée la pente de la droite de régression. Pourquoi ?

La pente de la droite de régression est également reliée au coefficient de corrélation entre les variables  $X$  et  $Y$ , noté  $\rho_{X,Y}$  par la relation

$$\rho_{X,Y} = \hat{\beta}_1 \sqrt{\frac{\text{Var}[X]}{\text{Var}[Y]}}.$$

Dans le cas de la régression linéaire simple, la pente permet également de mesurer le lien entre les variables  $X$  et  $Y$ . Attention cependant, une pente élevée ne signifie pas que la corrélation sera importante car cela dépend également des variances des variables aléatoires  $X$  et  $Y$ .

2. On considère maintenant le modèle de régression multiple à  $p$  variables :

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

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

Dans le cas où notre matrice de design  $\mathbf{X}$  est de rang plein, égal à  $p + 1$ , nous avons

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

- (b) Il est aussi nécessaire d'estimer la variance  $\sigma^2$  des erreurs de notre modèle. Donner un estimateur sans biais de cette variance en fonction de la taille  $n$  et de l'échantillon et du nombre de variables indépendantes  $p$ .

Un estimateur sans biais  $\hat{\sigma}^2$  de la variance  $\sigma^2$  de notre modèle est donné par

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

où, pour tout  $i \in \llbracket 1, n \rrbracket$ ,  $\hat{\varepsilon}_i$  désigne le résidu associé au  $i$ -ème exemple.

- (c) On souhaite tester la significativité des paramètres du modèle, quel test doit-on effectuer ? On précisera les hypothèses, la définition de la statistique de test ainsi que sa distribution.

Pour tester la significativité d'un paramètre  $\beta_j$  du modèle, on procède au test suivant

$$H_0 : \beta_j = 0 \quad \text{v.s.} \quad H_1 : \beta_j \neq 0.$$

Il s'agit donc d'un test bilatéral, qui, dans le cas présent, repose sur la loi de Student.

La statistique de test employée est définie par

$$t_{\text{test}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\text{Var}[\hat{\beta}_j]}} \underset{\text{sous } H_0}{\sim} T_{n-p-1}.$$

La statistique de test suit donc une loi de student à  $n - p - 1$  degrés de libertés.

3. On considère un jeu de données de régression et on utilise le logiciel  pour effectuer notre régression linéaire. Les sorties du logiciel  sont les suivantes :

```
##
## Call:
## lm(formula = medv ~ ., data = BostonHousing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00  7.144 3.28e-12 ***
## crim        -1.080e-01  3.286e-02 -3.287 0.001087 **
## zn           4.642e-02  1.373e-02  3.382 0.000778 ***
## indus        2.056e-02  6.150e-02  0.334 0.738288
## chas1        2.687e+00  8.616e-01  3.118 0.001925 **
## nox         -1.777e+01  3.820e+00 -4.651 4.25e-06 ***
## rm           3.810e+00  4.179e-01  9.116 < 2e-16 ***
## age          6.922e-04  1.321e-02  0.052 0.958229
## dis         -1.476e+00  1.995e-01 -7.398 6.01e-13 ***
## rad          3.060e-01  6.635e-02  4.613 5.07e-06 ***
## tax         -1.233e-02  3.760e-03 -3.280 0.001112 **
## ptratio     -9.527e-01  1.308e-01 -7.283 1.31e-12 ***
## b            9.312e-03  2.686e-03  3.467 0.000573 ***
## lstat       -5.248e-01  5.072e-02 -10.347 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

Analyser les sorties du logiciel et

- (a) préciser les variables non significatives dans le modèle

L'analyse fait ressortir deux variables non significatives : les variables *indus* et *age* comme le montrent les *p*-valeurs supérieures à un seuil significatif de 0.05.

- (b) dire si le modèle est globalement significatif ou non.  
Sur quel test repose la significativité global du modèle ?

Le modèle est globalement significatif. En effet, le test de Fisher permettant de vérifier cela renvoie une *p*-valeur de  $2.2 \times 10^{-16}$ .

- (c) donner le coefficient de détermination du modèle

Le coefficient de détermination correspond à la valeur du  $R^2$  qui est égale à 0.7406. Nous pouvons dire que nous avons un modèle qui est globalement satisfaisant.

4. Lorsque l'on enlève les variables non significatives du modèles et que l'on effectue à nouveau la régression, on trouve les informations suivantes : (i) l'estimation de l'erreur standard du modèle est égale à 4.736 ; (ii) le coefficient de détermination vaut 0.7406 ; (iii) le coefficient de détermination ajusté vaut 0.7348 ; (iv) la *p*-value associée à la significativité global du modèle est de  $2.2 \times 10^{-16}$ .

- (a) Peut-on dire que ce nouveau modèle est plus intéressant que le précédent ?  
Pourquoi ? Sur quel critère se base votre conclusion ?

Pour comparer deux modèles qui reposent sur un nombre différents de variables, on doit se baser sur le coefficient de détermination ajusté  $R_{aj}^2$ . Dans le cas présent, le  $R_{aj}^2$  du second modèle est plus élevé que le  $R_{aj}^2$  du premier modèle (complet). Le second modèle est donc plus intéressant.

Pour voir si cette différence est singificative, nous pourrions également procéder à un test statistique (mais ce dernier n'a pas été étudié).

- (b) Citer un autre critère, vu en cours qui permet d'évaluer la qualité d'un modèle et qui utilise une des quantités susmentionnée.

Nous pourrions également évaluer la qualité d'un modèle à l'aide du **BIC** qui met en jeu la somme des carrés des résidus et donc  $\hat{\sigma}^2$ .

## Exercice 2

Dans cet exercice, on considère le jeu de données qui figure dans la table ci-dessous

| $y$   | 2  | 3 | 4 | 5  | 6 | 7 | 8  | 9  |
|-------|----|---|---|----|---|---|----|----|
| $x_1$ | -1 | 0 | 1 | -2 | 0 | 4 | -2 | 0  |
| $x_2$ | 0  | 0 | 0 | 1  | 2 | 0 | -1 | -2 |

On va également considérer le modèle multiple suivant

$$Y = \beta \mathbf{X} + \varepsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$

où les notations ont le même sens que celles employées en cours.

1. Donner les valeurs de l'estimateur  $\hat{\beta}$ , on arrondira le résultat au centième près.

Les valeurs exactes du paramètre  $\beta$  sont données par :

```
n = 8
p = 2

# Création de la matrice de design
x0 = rep(1,n)
x1 = c(-1,0,1,-2,0,4,-2,0)
x2 = c(0,0,0,1,2,0,-1,-2)

# Valeurs de la variable réponse/dépendante
y = c(2:9)

X = cbind(x0, x1, x2)

# Estimateur des paramètres du modèle
beta_hat = solve(t(X)%*%X)%*%t(X)%*%y
beta_hat
```

```
##          [,1]
## x0  5.5000000
## x1  0.1538462
## x2 -0.9000000
```

2. L'estimateur de la variance  $\hat{\sigma}^2$  associé à ce modèle a une valeur égale à 6.66.

- (a) En déduire l'écart-type associé à chaque paramètre du modèle. On arrondira le résultat à deux chiffres significatifs.

On rappelle que la variance de l'estimateur  $\hat{\beta}$  est donnée par

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

La variance des différents paramètres se lit ensuite sur les différents éléments diagonaux de cette matrice.

Ce qui nous donne :

```
sigma2 = 6.66
var_beta_hat = diag(sigma2*solve(t(X)%*%X))
sqrt(var_beta_hat)

##          x0          x1          x2
## 0.9124144 0.5061164 0.8160882
```

- (b) A l'aide d'un test approprié, on indiquera si les paramètres du modèle sont significatifs ou non, au risque d'erreur  $\alpha = 0.05$ .

On pourra, pour cela, utiliser une des valeurs des quantiles suivants :

$$t_{6,0.95} = 1.943, t_{5,0.95} = 2.015, t_{5,0.975} = 2.571 \quad \text{ou} \quad t_{6,0.975} = 2.447$$

On effectue le test décrit dans l'exercice précédent et qui repose sur la loi de Student.

```
# On calcule les différentes valeurs de statistique de test.

t_test = beta_hat/sqrt(var_beta_hat)
t_test

##          [,1]
## x0  6.0279629
## x1  0.3039738
## x2 -1.1028219
```

3. On cherche maintenant à étudier la performance global du modèle.

- (a) Etant donnée la valeur de  $\hat{\sigma}^2$ , en déduire la valeur de **SCR** (la somme des carrés des résidus).

On a directement

```
# Calcul de SCR
SCR = (n-p-1)*sigma2
SCR
## [1] 33.3
```

- (b) Calculer la valeur de **SCT** de votre jeu de données (la somme des carrés totaux).

Il s'agit simplement de déterminer, à un facteur multiplicatif prêt, la variance de  $Y$ .

```
# Caclul de SCT
SCT = sum((y-mean(y))^2)
SCT
## [1] 42
```

- (c) En déduire la valeur du coefficient de détermination  $R^2$  et commenter.

On rappelle que le coefficient de détermination  $R^2$  est défini par

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

L'application numérique nous donne

```
R_square = 1-SCR/SCT
R_square
## [1] 0.2071429
```

- (d) En déduire la valeur du coefficient de détermination ajusté  $R_{aj}^2$ .

On rappelle que le coefficient de détermination ajusté  $R_{aj}^2$  est défini par

$$R_{aj}^2 = 1 - \frac{SCR}{SCT} \times \frac{n-1}{n-p-1}.$$

L'application numérique nous donne

```
R_square = 1-(SCR/SCT)*(n-1)/(n-p-1)
R_square
## [1] -0.11
```

*Remarque : la valeur de ce dernier peut être négative.*

4. On cherche maintenant à tester la significativité globale du modèle
- (a) A l'aide des questions précédentes, en déduire la valeur de **SCE**, la somme des carrés expliqués par le modèle.

Nous avons la relation

$$SCT = SCE + SCR,$$

où *SCE* est la somme des carrés expliquée par le modèle.

```
SCE = SCT - SCR
SCE
## [1] 8.7
```

- (b) Donner la valeur de la statistique de test permettant de tester la significativité globale du modèle.

La statistique de test est définie comme le rapport entre la variance expliquée par le modèle ( $SCE/p$ ) et la variance résiduelle ( $SCR/(n-p-1)$ ), *i.e.*,

$$F_{\text{test}} = \frac{\frac{SCE}{p}}{\frac{SCR}{n-p-1}} = \frac{SCE}{SCR} \times \frac{n-p-1}{p}.$$

```
F_test = (SCE/SCR)*(n-p-1)/p
F_test
## [1] 0.6531532
```