compte-rendu CTD10

2025-11-21

Contents

Semaine 10

Objectifs

L'objectif de cette séance est de se familiariser avec la **régression linéaire multiple** comportant **deux variables explicatives**. On étudiera le cas où les prédicteurs sont indépendants et celui où ils sont corrélés.

Cette séance introduira également la notion de **facteur de confusion**, essentielle pour une interprétation correcte des coefficients estimés dans un modèle de régression.

Exercice 2. CTD9. Formules des coefficients de régression ($d = 2$)

Contexte

```r
df = read.csv2("./data/co2_per_capita.csv")
View(df)
```

On considère le modèle

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \epsilon\,,$$

où (X_1, X_2) sont des variables aléatoires indépendantes de ϵ, mais pas nécessairement indépendantes entre elles. On note $K_{\mathbf{x}}$ la matrice de covariance du vecteur (X_1, X_2), que l'on suppose de déterminant non nul.

$$K_{\mathbf{x}} = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}(X_2) \end{pmatrix}.$$

L'objectif de l'exercice est d'obtenir des formules mathématiques pour les coefficients de régression, similaires à celles obtenues dans le cas de la régression linéaire simple.

Question 1

Dans le cas d'une régression linéaire simple, rappeler la formule permettant de calculer la pente de la droite de régression.

Solution. Se souvenir que si le prédicteur est standardisé (écart-type égal à 1), alors la pente de la régression est égale à la covariance entre les deux variables.

Réponse:

$$b_1 = \frac{Cov(X,Y)}{\mathbb{V}(X)}$$

Question 2

Montrer que

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = K_{\mathbf{x}}^{-1} \begin{pmatrix} \mathrm{Cov}(Y, X_1) \\ \mathrm{Cov}(Y, X_2) \end{pmatrix}$$

et que

$$b_0 = \mathbb{E}[Y] - b_1\mathbb{E}[X_1] - b_2\mathbb{E}[X_2]\,.$$

Solution. Pour la première équation, développer les expressions des covariances $\mathrm{Cov}(X_1, Y)$ et $\mathrm{Cov}(X_2, Y)$. On obtient alors un système linéaire à deux inconnues que l'on peut inverser facilement.

Réponse:

On sait que,

$$Y = b_0 + b_1.X_1 + b_2.X_2 + \epsilon$$

On note

$$K_X$$

la matrice de covariance du couple (X_1, X_2).

Par bilinéarité de la covariance, et en sachant que $Cov(X,X) = \mathbb{V}(X)$ on a :

$$\begin{cases} Cov(Y, X_1) = b_1.\mathbb{V}(X_1) + b_2.Cov(X_2, X_1) \\ Cov(Y, X_1) = b_1.Cov(X_1, X_2) + b_2.\mathbb{V}(X_2) \end{cases}$$

Ainsi,

$$\begin{pmatrix} Cov(Y, X_1) \\ Cov(Y, X_2) \end{pmatrix} = K_X . \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

Donc,

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = K_X^{-1} . \begin{pmatrix} Cov(Y, X_1) \\ Cov(Y, X_2) \end{pmatrix}$$

Par linéarité de l'espérance on trouve directement,

$$b_0 = \mathbb{E}[Y] - b_1\mathbb{E}[X_1] - b_2\mathbb{E}[X_2]\,.$$

$\mathbb{E}(\epsilon) = 0$ car $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Question 3

O

À partir du modèle suivant on reprend les données du fichier `co2_per_capita.csv`. L'objectif est d'expliquer l'indice de développement humain (HDI/IDH) d'un pays à partir de deux variables explicatives : le revenu national brut par habitant (`gni.per.capita`, log10), et l'espérance de vie à la naissance (`life.expectancy`).:

```
mod_hdi <-  lm(hdi ~ gni.per.capita  + life.expectancy, data = df)
```

Estimez les coefficients b_1 et b_2 à l'aide des valeurs empiriques correspondant à la question 2.

Solution. Pour estimer les tailles d'effet b_1 et b_2, les fonctions utiles sont `cov`, `solve` et le produit martriciel.

```
## definition des variables

## Échantillon pour y qui correspond à la colonne HDI de df
y <- df[,"hdi"]

## Échantillon pour x_1 qui correspond à la colonne "gni.per.capita" de df
## de même pour x_2 avec "life.expectancy".
x_1 <- df[,"gni.per.capita"]
x_2 <-  df[,"life.expectancy"]

## Matrice des échantillons x_1 et x_2.

x <- cbind(x_1, x_2)
```

Calculer la formule obtenue

```
## definition de K_x

K_x <- cov(x)

## completer en verifiant les dimensions des matrices
## produit de 2 matrices : %*%
## inversion matricielle : solve()
Mc <- c(cov(y,x_1),cov(y,x_2))
b <-  solve(K_x) %*% matrix(Mc,nrow = 2)

b <- t(b)
colnames(b) <- c("gni.per.capita","life.expectancy")
b
```

```
##      gni.per.capita life.expectancy
## [1,]      0.2032053     0.006674646
```

Question 4

Vérifier les résultats obtenus en comparant les calculs avec les coefficients de régression estimés par la fonction `lm`.

Solution. Inclure le code R correspondant. Utiliser la commande `coefficients(mod_hdi)` ou consulter le résumé du modèle avec `summary(mod_hdi)`.

```
summary(mod_hdi)
```

```
##
## Call:
## lm(formula = hdi ~ gni.per.capita + life.expectancy, data = df)
```

```
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.099424 -0.013816  0.004781  0.019540  0.070776
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.578747   0.021785  -26.57   <2e-16 ***
## gni.per.capita   0.203205   0.008930   22.75   <2e-16 ***
## life.expectancy  0.006675   0.000595   11.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03025 on 166 degrees of freedom
## Multiple R-squared:  0.9617, Adjusted R-squared:  0.9612
## F-statistic:  2083 on 2 and 166 DF,  p-value: < 2.2e-16
```

On retrouve bien les mêmes valeurs pour b_1, b_2 .

Question 5

À partir des colonnes `Estimate` et `Std. Error` du résumé du modèle, calculer les valeurs des statistiques t associées aux coefficients des prédicteurs `gni.per.capita` et `life.expectancy`.

Solution. On peut extraire les coefficients de régression et leurs écarts-types à partir du résumé du modèle, puis calculer les t-scores demandés (hypothèses $b_1 = 0$ et $b_2 = 0$).

```r
# remplacer eval = FALSE par eval = TRUE
summary_tab <-  coef(summary(mod_hdi))[, c("Estimate", "Std. Error")]
summary_tab
```

```
##                     Estimate   Std. Error
## (Intercept)     -0.578747485 0.0217851085
## gni.per.capita   0.203205301 0.0089305095
## life.expectancy  0.006674646 0.0005949774
```

```r
t1 <- summary_tab[2]/summary_tab[5]
cat("P-valeur pour b1",2*(pt(abs(t1),df = 166,lower = FALSE)),"\n")
```

```
## P-valeur pour b1 6.659563e-53
```

```r
t2 <- summary_tab[3]/summary_tab[6]
cat("P-valeur pour b2",2*(pt(abs(t2),df = 166,lower = FALSE)),"\n")
```

```
## P-valeur pour b2 4.277345e-22
```

```r
summary(mod_hdi)
```

```
##
## Call:
## lm(formula = hdi ~ gni.per.capita + life.expectancy, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.099424 -0.013816  0.004781  0.019540  0.070776
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)     -0.578747   0.021785  -26.57   <2e-16 ***
## gni.per.capita   0.203205   0.008930   22.75   <2e-16 ***
## life.expectancy  0.006675   0.000595   11.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03025 on 166 degrees of freedom
## Multiple R-squared:  0.9617, Adjusted R-squared:  0.9612
## F-statistic:  2083 on 2 and 166 DF,  p-value: < 2.2e-16
```

```
t1
```

```
## [1] 22.75405
```

```
t2
```

```
## [1] 11.21832
```

Vérification : On retrouvera cette valeur dans la colonne “t value” du résumé.

Question 6

Calculer la p-valeur du test de l'hypothèse nulle “$H_0 : b_1 = 0$” en utilisant la commande `pt`. Vérifier le résultat à l'aide du résumé du modèle.

Solution : La p-valeur du test bilatéral correspond à

$$p = \mathbb{P}(|T_{n-3}| > |t|) = 2\mathbb{P}(T_{n-3} > |t|)$$

où T_{n-3} suit la loi de Student à $n - 3$ degrés de libertés.

Réponse:

```
## On calcule la p-valeur à partir de la statistique t1 précédente :

2*pt(abs(t1),df = 166, lower = FALSE)
```

```
## [1] 6.659563e-53
```

La p-valeur est très faible, on peut donc rejetter l'hypothèse “$b_1 = 0$” , ce qui signifie que le GNI a un effet significatif sur le HDI.

Question 7. Test de pertinence de la régression

Calculer la p-valeur du test de l'hypothèse nulle “$H_0 : b_1 = b_2 = 0$” en utilisant la commande `pf`. Vérifier le résultat à l'aide du résumé du modèle.

Solution : D'après le cours, la p-valeur du test correspond à

$$p = \mathbb{P}(F_{2,n-3} > f_{2,n-3})$$

où $F_{2,n-3}$ est de loi de Fisher $F(2, n - 3)$ et

$$f_{2,n-3} = \frac{n-3}{2}\frac{R^2}{1-R^2} = \frac{n-3}{2}\frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{RSS}}.$$

```r
## On calcule tout d'abord f_{2,n-3}

f = (166/2)*(summary(mod_hdi)$r.squared)/(1-summary(mod_hdi)$r.squared)

## La p-valeur est alors:

pf(f,df1 = 2,df2 = 166, lower.tail = FALSE)
```

```
## [1] 2.64025e-118
```

La p-valeur est très faible, donc la régression linéaire est très pertinente.

Question 8. Interprétation du coefficient de détermination.

On considère à nouveau le modèle probabiliste de la régression multiple

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \epsilon \,,$$

où $X = (X_1, X_2)$ forme un couple de variables aléatoires indépendantes de ϵ. La loi de ϵ est d'espérance nulle et de variance σ^2.

- Déterminer l'espérance conditionnelle

$$Z = \mathbb{E}[Y|X] \,.$$

Vérifier que Z est constante si et seulement si $b_1 = b_2 = 0$.

- Montrer que la part de variance de Y expliquée par X vérifie

$$\rho^2 = \frac{\mathrm{Var}(\mathbb{E}[Y|X])}{\mathrm{Var}(Y)} = 1 - \frac{\sigma^2}{\mathrm{Var}(Y)} \,.$$

Vérifier que $\rho^2 = 0$ si et seulement si $b_1 = b_2 = 0$.

Solution : Rappeler l'interprétation de ρ^2 lorsque $d = 1$. Remarquer que les calculs probabilistes sont très simples (la variance de Y expliquée par X est égale à la variance de Y expliquée par Z).

Réponse:

- Quand $d = 1$, ρ^2 est le carré du coefficient de corrélation linéaire entre X et Y.
- On a directement que,

$$Z = b_0 + b_1.X_1 + b_2.X_2$$

par linéarité et car $\mathbb{E}(\epsilon|X) = 0$ car ϵ est d'espérance nulle.

Il est clair que Z est constant si et seulement si $b_1 = b_2 = 0$.

- On a par la formule de la variance totale,

$$\mathbb{V}(Y) = \mathbb{V}(\mathbb{E}(Y|X)) + \mathbb{E}(\mathbb{V}(Y|X))$$

Soit,

$$\mathbb{V}(Y) = \mathbb{V}(\mathbb{E}(Y|X)) + \sigma^2$$

Soit,

$$\rho^2 = 1 - \frac{\sigma^2}{\text{Var}(Y)}$$

Question 9. Interprétation du coefficient de détermination.

- Expliquer pourquoi, comme dans la régression linéaire simple, la part de variance de Y expliquée par X peut être estimée par le coefficient de détermination

$$\mathrm{R}^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}\,.$$

- Calculer la part de la variation de la variable `hdi` expliquée par les prédicteurs `gni.per.capita` et `life.expectancy`

Solution : Reprendre le résumé du modèle `mod_hdi`

- La corrélation ρ mesure la force de la relation linéaire entre X et Y.

or, R^2 mesure "combien" la linéarité entre X et Y explique Y ;
Nous avons que R^2 est une approximation de ρ^2, pour obtenir R^2 il suffit donc de remplacer σ^2 par $\frac{RSS}{n}$ et $Var(Y)$ par $\frac{TSS}{n}$ car ce sont leurs approximations respectives.

```
mod = summary(mod_hdi)
mod$r.squared
```

```
## [1] 0.9616828
```

```
y <- df$hdi
ychap <- mod_hdi$fitted.values
# Calcul du TSS et RSS
TSS <- sum( (y - mean(y))^2 )
RSS <- sum( (y - ychap)^2 )
# Calcul du R-squared
R2 <- 1 - RSS/TSS
R2
```

```
## [1] 0.9616828
```

Nous obtenons une part de variance expliquée de 0.96, en cherchant r.squared dans les summary de mod_hdi nous retrouvons la même valeur.

Exercice 1 CTD10

On considère ici une variable aléatoire Y définie par le modèle de régression linéaire multiple suivant :

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \epsilon,$$

où les prédicteurs X_1 et X_2 sont des variables aléatoires indépendantes du terme d'erreur ϵ et de variance égale à 1 :

$$\text{Var}(X_1) = \text{Var}(X_2) = 1.$$

On note $\rho_X = \text{Cor}(X_1, X_2)$ leur corrélation linéaire, supposée strictement inférieure à 1 en valeur absolue.theta <- 0.9

Question 1 Sous les hypothèses du modèle (notamment $\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = 1$ et $\mathrm{Cov}(X_1, \epsilon) = \mathrm{Cov}(X_2, \epsilon) = 0$, montrer que

$$b_1 = \mathrm{Cov}(X_1, Y) - b_2\, \rho_X .$$

Solution.

Développons $\mathrm{Cov}(X_1, Y)$.

$\mathrm{Cov}(X_1, Y) = \mathrm{Cov}(X_1, b_0 + b_1 X_1 + b_2 X_2 + \epsilon)$

$\mathrm{Cov}(X_1, Y) = b_1 \mathrm{Cov}(X_1, X_1) + b_2 \mathrm{Cov}(X_1, X_2) + \mathrm{Cov}(X_1, \epsilon)$ par bilinéarité de la covariance, soit,

$\mathrm{Cov}(X_1, Y) = b_1 \mathrm{var}(X_1) + b_2 \mathrm{Cov}(X_1, X_2) + \mathrm{Cov}(X_1, \epsilon)$

Or d'après les hypothèses $\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = 1$ on a $\mathrm{Cov}(X_1, X_2) = \rho_x$ et X_1 et ϵ donc $\mathrm{Cov}(X_1, \epsilon) = 0$

Finalement nous obtenons le résultat : $b_1 = \mathrm{Cov}(X_1, Y) - b_2\, \rho_X$.

Question 2

On ajuste un **modèle de régression linéaire incomplet** de la forme

$$Y = a_0 + a_1 X_1 + \epsilon',$$

à partir de n observations issues du **modèle complet** présenté dans l'introduction.

On appelle **biais sur la taille d'effet du prédicteur** X_1 la différence entre la valeur théorique a_1 que l'on pourrait estimer à partir du modèle incomplet et la vraie valeur b_1 du modèle complet :

$$\text{biais} = a_1 - b_1 .$$

1. Sous les hypothèses du modèle de régression linéaire simple, montrer que $a_1 = \mathrm{Cov}(X_1, Y)$.
2. Montrer que le biais sur la taille d'effet de X_1 s'écrit

$$\text{biais} = b_2\, \rho_X .$$

3. Donner une interprétation succincte de ce biais : expliquer pourquoi la formule obtenue est intuitive compte tenu du rôle de la corrélation entre X_1 et X_2.

Solution.

1. Montrons que $a_1 = \mathrm{Cov}(X_1, Y)$

   On rappelle que le modèle incomplet et celui ci : $Y = a_0 + a_1 X_1 + \epsilon'$, donc, puisqu'on a un modèle de régression linéaire simple, on a alors : $a_1 = \frac{\mathrm{Cov}(X_1, Y)}{Var(X_1)}$ or, ici $Var(X_1) = 1$ donc, $a_1 = \mathrm{Cov}(X_1, Y)$.

2. Montrons que biais $= b_2\, \rho_X$

   Puisqu'on vient de montrer que $a_1 = \mathrm{Cov}(X_1, Y)$ soit biais $= \mathrm{Cov}(X_1, Y) - b_1$. : et que dans la question 1, on a $\mathrm{Cov}(X_1, Y) = b_1 + b_2\, \rho_X$, On en déduit alors que : biais $= b_2\, \rho_X$.

3. Donnons une interprétation

   Le biais relie donc les deux variables aléatoires par leur corrélation linéaire qui indique la relation enre les 2 variables mais aussi la force de cette relation. Cela est cohérent car le biais représente la différence d'évolution entre les 2 modèles. Plus l'effet de X_2 est fort ou plus X_1 et X_2 sont corrélés, plus le biais est fort. A l'inverse, si X_1 et X_2 ne sont pas corrélés, le biais est nul.

Question 3

On considère un échantillon simulé de la manière suivante ($n = 1000$).

```
# Graine du générateur aléatoire
set.seed(1962)

# Taille d'echantillon et paramètre de simulation
n <- 1000
theta <-  0.9

# Simulation des variables X1 et X2 corrélées
x_1 <- rnorm(n, m = 0, sd = 1)
x_2 <- theta *  x_1 + rnorm(n, m = 0, sd = sqrt(1-theta^2))

# Simulation de la réponse Y
y = 1.2 + 0.5*x_1 + 3*x_2 + rnorm(n, m = 0, sd = .5)
```

1. Montrer que les variables X_1 et X_2 sont de variance 1.
2. Montrer que la valeur `theta` choisie dans la simulation est la corrélation théorique entre les variables aléatoires X_1 et X_2.
3. Calculez la corrélation empirique entre les vecteurs simulés `x_1` et `x_2` et comparez-la à la valeur théorique `theta`.

Solution.

1. X_1 suit une loi normale de paramètre 0,1 donc par définition, la variance de X_1 est 1.

   En notant X' la variable aléatoire de loi $\mathcal{N}(0, 1 - \theta$š$)$, on a $X_2 = \theta X_1 + X'$. Alors,

   $Var(X_2) = Var(\theta X_1 + X') = Var(\theta X_1) + Var(X') + 2Cov(\theta X_1, X_2)$

   $Var(X_2) = \theta$š $+ [\sqrt{(1 - \theta$š$)}]$š $= 1$

```
#On vérifie que les moyennes empiriques sont proches de zéro
cat("La moyenne de X_1 est : ", mean(x_1), "\n")
```

```
## La moyenne de X_1 est :  0.02068673
```

```
cat("La moyenne de X_2 est : ", mean(x_2), "\n")
```

```
## La moyenne de X_2 est :  0.01886366
```

```
#On vérifie les valeurs par approche expérimentale

cat("La variance de X_1 est : ", var(x_1), "\n")
```

```
## La variance de X_1 est :  0.9611809
```

```
cat("La variance de X_2 est : ", var(x_2))
```

```
## La variance de X_2 est :  0.9226189
```

2.

On a dans un premier temps :

$\text{Cov}(X_1, X_2) = \text{Cov}(X_1, \theta X_1 + X') = \theta \text{Cov}(X_1, X_1) + \text{Cov}(X_1, X') = \theta$ car $\mathbb{V}(X_1) = \mathbb{V}(X_2) = 1$ et aussi grâce au fait que X_1 et X' sont indépendants.

Ainsi, on peut calculer :

$\rho(X_1, X_2) = \frac{\text{Cov}(X_1,X_2)}{\mathbb{V}(X_1)\mathbb{V}(X_2)} = \text{Cov}(X_1, X_2) = \theta$

3- Corrélation empirique

```
# calcul de theta empirique
theta_emp = cor(x_1, x_2)
cat("valeur empirique de théta : ",theta_emp,"\n")
```

```
## valeur empirique de théta :  0.8838507
```

```
#comparaison avec la valeur théorique de theta
cat("valeur théorique de théta : ",theta)
```

```
## valeur théorique de théta :  0.9
```

Avec la fonction $cor(x_1, x_2)$ on obtient un théta empirique de 0,88 qui est donc très proche de la valeur théorique 0,9 de théta.

Question 4

Ajuster un modèle de regression linéaire incomplet à la variable `y` en utilisant uniquement la variable x_1.

1. Quelle est la valeur (théorique) du biais sur la taille d'effet b_1 ?
2. Quelle est l'estimation de ce biais (calculer $\hat{a}_1 - \hat{b}_1$ et $\hat{b}_2 \times \hat{\theta}$) ?
3. Commenter les résultats et conclure.

Solution.

On récupère le modèle incomplet précédent :

$$Y = a_0 + a_1 X_1 + \epsilon',$$

1. Calculons le biais :

   $biais = a_1 - b_1 = b_2 * Cor(X_1, X_2) = 3 * \theta = 2,7$

2. Retrouvons ce biais par estimation :

   ```
   # Estimation de a1, b1 et b2
   a1_hat <- cov(x_1,y) / var(x_1)

   den <- var(x_1) * var(x_2) - cov(x_1,x_2)^2
   b1_hat <- (cov(x_1,y) * var(x_2) - cov(x_1,x_2) * cov(x_2,y)) / den
   b2_hat <- (var(x_1) * cov(x_2,y) - cov(x_1,x_2) * cov(x_1,y)) / den

   # calcul du biais par a_1 - b_1
   a1_hat -b1_hat
   ```

   ```
   ## [1] 2.619018
   ```

   ```
   # calcul du biais par b2*theta
   b2_hat*theta_emp
   ```

   ```
   ## [1] 2.67319
   ```

3. Le biais est positif, nous voyons donc que X_2 a un effet sur Y et que ce n'est pas uniquement X_1 qui agit sur Y.