Contrôle Continu

Durée 2h00. Les documents, les téléphones portables, tablettes, ordinateurs ne sont pas autorisés. La calculatrice est autorisée. Les exercices sont indépendants. La qualité de la rédaction sera prise en compte. Tous les résultats numériques seront donnés avec une précision de deux chiffres après la virgule.

Exercice 1. QCM. On se place dans le cadre du modèle de régression linéaire multiple gaussien : $Y = X\beta + \epsilon$, avec Y un vecteur de taille n, β un vecteur de taille p, X une matrice de rang p, et $\epsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$. Répondez aux questions suivantes. Une seule réponse est acceptée par question.

1. La matrice X est de taille :
   (a) $n \times p$
   (b) $p \times n$

   $n \times p$

2. Dans ce modèle, les Y_i, $1 \leq i \leq n$, sont aléatoires, indépendants et identiquement distribués (i.i.d.).
   (a) Vrai
   (b) Faux

   Faux : non identiquement distribués (l'espérance change).

3. Dans ce modèle, les ϵ_i, $1 \leq i \leq n$, sont aléatoires i.i.d.
   (a) Vrai
   (b) Faux

   Vrai

4. Dans ce modèle, les β_k, $1 \leq k \leq p$, sont aléatoires i.i.d.
   (a) Vrai
   (b) Faux

   Faux : non aléatoire.

5. L'estimateur des moindres carrés $\hat{\beta}$ est la projection de Y sur l'espace engendré par les colonnes de X.
   (a) Vrai
   (b) Faux

   Faux : c'est $\hat{Y} = X\hat{\beta}$ qui est la projection de Y sur l'espace engendré par les colonnes de X.

6. L'estimateur des moindres carrés $\hat{\beta}$ est l'estimateur de β qui a la plus petite variance (au sens matriciel).
   (a) Vrai
   (b) Faux

   Faux : $\hat{\beta}$ est l'estimateur non biaisé linéaire de β qui a la plus petite variance parmi les estimateurs non biaisé linéaire.

7. L'estimateur du maximum de vraisemblance de β est non biaisé.
   (a) Vrai
   (b) Faux

   Vrai, il est égal à l'estimateur des moindres carrés.

8. L'estimateur du maximum de vraisemblance de σ^2 est non biaisé.
   (a) Vrai
   (b) Faux

   Faux, son espérance vaut $\frac{n-p}{n}$.

9. Les estimateurs des moindres carrés $\hat{\beta}$ et $\hat{\sigma}^2$ sont indépendants.
   (a) Vrai
   (b) Faux

   Vrai (théorème de Cochran)

10. Les estimateurs des moindres carrés $\hat{\beta}_k$, $1 \leq k \leq p$, sont indépendants.
    (a) Vrai
    (b) Faux

    Faux.

Exercice 2. On souhaite faire la régression simple d'une variable Y en fonction d'une variable explicative X (avec intercept). Pour cela, on dispose de n observations (x_i, y_i), $1 \leq i \leq n$, et des statistiques résumées suivantes :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 1 \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 2.07$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0.35 \qquad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 15.93$$

$$s_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 2.22$$

$$\bar{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 1.34 \qquad s_{x^2}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - \bar{x^2})^2 = 1.56$$

$$s_{x^2y}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - \bar{x^2})(y_i - \bar{y}) = 4.85$$

1. Posez un modèle de régression en précisant les hypothèses, et donnez l'expression des estimateurs des coefficients.

   Voir le cours.

2. Calculez les coefficients en utilisant les données fournies.

   On estime les paramètre du modèle $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Les estimateurs des moindres carrés sont:

   $$\hat{\beta}_1 = \frac{s_{xy}^2}{s_x^2} = \frac{2.22}{0.35} = 6.34$$

   et

   $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -4.27$$

   La droite des MC est $y_i = -4.27 + 6.34 \times x_i$.

3. Rappelez la définition du R^2 du modèle et son interprétation.

   Voir le cours.

4. On rappelle que, dans le cas d'une régression simple, on a $R^2 = \rho_{x,y}^2$, avec $\rho_{x,y}$ le coefficient de corrélation entre x et y. Calculez le R^2 du modèle.

   Le coefficient de corrélation entre x et y est donné par:

   $$\begin{aligned}\rho_{x,y} &= \frac{s_{xy}^2}{\sqrt{s_x^2 s_y^2}} \\ &= \frac{2.22}{\sqrt{0.35 \times 15.93}} = 0.94\end{aligned}$$

   D'où $R^2 = \rho_{x,y}^2 = 0.88$.

5. On souhaite maintenant faire la régression de Y contre X^2. Posez le modèle associé, calculez les estimateurs de ses coefficients et le R^2 de ce modèle.

   On fait la régression $y_i = \beta_0' + \beta_1' x_i^2 + \epsilon_i$.

Les estimateurs des moindres carrés sont:

$$\hat{\beta}_1' = \frac{s^2_{x^2y}}{s^2_{x^2}} = \frac{4.85}{1.56} = 3.11$$

et

$$\hat{\beta}_0' = \bar{y} - \hat{\beta}_1' \bar{x^2} = -2.09$$

Le coefficient de corrélation entre x^2 et y est donné par:

$$\begin{aligned}\rho_{x^2,y} &= \frac{s^2_{x^2y}}{\sqrt{s^2_{x^2} s^2_y}} \\ &= \frac{4.85}{\sqrt{1.56 \times 15.93}} = 0.97\end{aligned}$$

D'où $R^2 = \rho^2_{x^2,y} = 0.95$.

6. D'après ces résultats, pouvez-vous préférer un modèle plutôt qu'un autre ?

   Le second modèle a un meilleur R^2, et le même nombre de paramètres que le premier, on préfère donc la régression sur X^2.

Exercice 3. On examine l'évolution d'une variable Y en fonction de deux variables x et z. On dispose de $n = 50$ observations de ces variables. On note $X = \begin{pmatrix}\mathbb{1} & x & z\end{pmatrix}$ où $\mathbb{1}$ est le vecteur constant et x, z sont les vecteurs des variables explicatives. On suppose que l'on a calculé :

$$X^TX = \begin{pmatrix} ? & 0 & 0 \\ ? & 4.28 & 5.41 \\ ? & ? & 103.86 \end{pmatrix} \quad , \quad Y^TY = \|Y\|^2 = 147.24 \quad , \quad \hat{\beta} = \begin{pmatrix} 0.1 \\ -2 \\ 1 \end{pmatrix}.$$

1. Donnez les valeurs manquantes dans la matrice.

   $$n = \mathbb{1}^T\mathbb{1} = [X^TX]_{11} = 50$$

   Puis, par symétrie, on a:

   $$X^TX = \begin{pmatrix} 50 & 0 & 0 \\ 0 & 4.28 & 5.41 \\ 0 & 5.41 & 103.86 \end{pmatrix}$$

2. Donnez l'expression de $\hat{\beta}$ et calculez X^TY.

   On a $X^TX\hat{\beta} = X^TY$, d'où

   $$X^TY = \begin{pmatrix} 50 & 0 & 0 \\ 0 & 4.28 & 5.41 \\ 0 & 5.41 & 103.86 \end{pmatrix} \begin{pmatrix} 0.1 \\ -2 \\ 1 \end{pmatrix} = \begin{pmatrix} 5 \\ -3.15 \\ 93.04 \end{pmatrix}$$

3. Calculez les moyennes empiriques $\bar{x}$, $\bar{z}$ et $\bar{y}$.

   On a:

   $$n\bar{x} = \mathbf{1}'x = [X^TX]_{12} = 0 \qquad n\bar{z} = \mathbf{1}'z = [X^TX]_{13} = 0 \qquad n\bar{y} = \mathbf{1}'y = [X^TY]_{12} = 5$$

   Donc:

   $$\bar{x} = 0 \qquad \bar{z} = 0 \qquad \bar{y} = 0.1$$

4. Donnez l'expression de $\hat{Y}$ et calculez $\left\|\hat{Y}\right\|^2$.

$$\hat{Y} = X\hat{\beta} = 0.1 \times \mathbb{1} + -2 \times x + 1 \times z$$

$$\left\|\hat{Y}\right\|^2 = \left\|X\hat{\beta}\right\|^2 = \hat{\beta}^T X^T X \hat{\beta} = \hat{\beta}^T X^T Y = \begin{pmatrix} 0.1 & -2 & 1 \end{pmatrix} \begin{pmatrix} 5 \\ -3.15 \\ 93.04 \end{pmatrix} = 99.84$$

5. Donnez l'estimateur sans biais de la variance, et calculez-le.

$$\hat{\sigma}^2 = \frac{1}{n-3}\left\|Y - \hat{Y}\right\|^2 = \frac{1}{n-3}\left(\|Y\|^2 - \left\|\hat{Y}\right\|^2\right) = \frac{1}{47}\left(147.24035 - 99.8312\right) = 1.00871$$

6. Sous quelles hypothèses peut-on connaitre la loi de $\hat{\sigma}^2$? Donnez l'expression de cette loi (normalisée).

   Voir le cours pour les hypothèses.

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$$

7. Construisez un intervalle de confiance à 95% pour σ^2.
   On donne les quantiles à 2.5%, 5%, 95% *et* 97.5% *de la loi du* χ^2 *à* 49, 48, *et* 47 *degrés de liberté dans la table suivante :*

	2.5%	*5%*	*95%*	*97.5%*
47	30	32.3	64	67.8
48	30.8	33.1	65.2	69
49	31.6	33.9	66.3	70.2

   On a:

$$S = (n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{47}.$$

   Par définition,

$$\mathbf{P}\left[q^{47}_{2.5\%} \leq S \leq q^{47}_{97.5\%}\right] = 95\%.$$

   Donc:

$$\mathbf{P}\left[(n-p)\hat{\sigma}^2/q^{47}_{97.5\%} \leq \sigma^2 \leq (n-p)\hat{\sigma}^2/q^{47}_{2.5\%}\right] = 95\%.$$

   Et un intervalle de confiance à 95% est donné par:

$$[0.69925; 1.5803]$$

Exercice 4. On dispose des vecteurs $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ et $\boldsymbol{y}$ décrivant les valeurs prises par trois variables X_1, X_2 et Y. On execute les commandes suivantes dans `R` :

```
fit_1 <- lm(y ~ x_1 + x_2)
summary(fit_1)
```

```
##
## Call:
## lm(formula = y ~ x_1 + x_2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.558 -0.807 -0.417  0.397  3.671
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.888      0.354   -5.34  5.4e-05 ***
## x_1           -0.290      0.375   -0.78     0.45
## x_2            2.960      0.251   11.78  1.3e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.58 on 17 degrees of freedom
## Multiple R-squared:  0.891, Adjusted R-squared:  0.878
## F-statistic: 69.4 on 2 and 17 DF,  p-value: 6.66e-09
```

```
fit_2 <- lm(y ~ x_2)
summary(fit_2)
```

```
##
## Call:
## lm(formula = y ~ x_2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.383 -0.873 -0.276  0.424  3.407
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.891      0.350   -5.41  3.9e-05 ***
## x_2            2.951      0.248   11.89  5.9e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.56 on 18 degrees of freedom
## Multiple R-squared:  0.887, Adjusted R-squared:  0.881
## F-statistic:  141 on 1 and 18 DF,  p-value: 5.88e-10
```

1. Donnez les équations (numériques) des deux droites de régression estimées.

   Premier modèle: $\hat{y}_i = -1.888 + -0.29 \times x_{1i} + 2.96 \times x_{2i}$
   Second modèle: $\hat{y}_i = -1.891 + 2.951 \times x_{2i}$

2. Quelle est la longueur des vecteurs $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ et $\boldsymbol{y}$?

   Le premier modèle a 17 degrés de liberté et 3 coefficients estimés, le nombre total d'observations est donc de $17 + 3 = 20$.
   Alternativement, en utilisant le second modèle, le nombre total d'observations est: $18 + 2 = 20$.

3. On s'intéresse au premier modèle. Que dire de la significativité des coefficients ? Justifiez. On décrira en détails le test associé, avec ses hypothèses.

Le coefficient associé à x_1 dans le premier modèle n'est pas significatif. En effet, la p-valeur du test de Student sur ce paramètre est de 0.45, avec un niveau de test $\alpha = 5\%$, on ne peut donc pas rejeter l'hypothèse nulle suivant laquelle ce coefficient est égal à zéro.
L'intercept et le coefficient associé à x_2 sont en revanche significatifs. On peut en effet rejeter l'hypothèse nulle suivant lequel ces coefficient est nul avec des p-valeurs inférieures à 2×10^{-16}.

4. Dans le premier modèle, les intervalles de confiance pour coefficient associé à `x_2` aux niveaux, respectivement, de 90%, 95% et 99% contiennent-ils zéro ?

   Non, car la p-valeur associée au test est plus petite que 1%.

5. Lequel des deux modèles a le R^2 le plus grand ? Est-ce surprenant ?

   Le second modèle est strictement emboîté dans le premier, qui a plus de paramètre. Il est donc attendu que le R^2 du premier modèle soit plus grand que celui du second. Ce n'est cependant pas le cas des R^2 ajustés.

6. Lequel des deux modèles préféreriez-vous ? Justifiez.

   Comme l'on ne peut pas rejeter l'hypothèse suivant laquelle le coefficient associé à x_1 est nul, on préfère le second modèle.
   De plus, le second modèle a un R^2 ajusté plus faible que le premier, ce qui nous le fait préférer.