Contrôle Continu

Durée 1h30. 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 paramètres, X une matrice de taille $n \times p$, et $\epsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$. On note $\hat{\beta}$ l'estimateur des moindres carrés de β. Répondez aux questions suivantes. Une seule réponse est acceptée par question.

1. Le vecteur $\hat{\beta}$ est de taille :
   (a) $n \times 1$ (b) $p \times 1$

   $p \times 1$

2. Dans ce modèle, les Y_i, $1 \leq i \leq n$, sont des variables aléatoires gaussiennes indépendantes.
   (a) Vrai (b) Faux

   Vrai.

3. Dans ce modèle, les $\hat{\beta}_k$, $1 \leq k \leq p$, sont des variables aléatoires gaussiennes indépendantes.
   (a) Vrai (b) Faux

   Faux : non indépendantes.

4. Laquelle de ces formules est correcte ?
   (a) $Y_i = \sum_{k=1}^{p} \beta_k X_{ik} + \epsilon_i$ (b) $Y_i = \sum_{k=1}^{p} \hat{\beta}_k X_{ik} + \epsilon_i$

   $Y_i = \sum_{k=1}^{p} \beta_k X_{ik} + \epsilon_i$

5. Laquelle de ces formules est correcte ?
   (a) $\hat{\beta} = (XX^T)^{-1}X^TY$ (c) $\hat{\beta} = (X^TX)^{-1}X^TY$
   (b) $\hat{\beta} = (XX^T)^{-1}XY$ (d) $\hat{\beta} = (X^TX)^{-1}XY$

   $\hat{\beta} = (X^TX)^{-1}X^TY$

6. On note $\hat{\epsilon} = Y - X\hat{\beta}$. $\hat{\epsilon}$ est la projection de Y sur l'espace engendré par les lignes de X.
   (a) Vrai (b) Faux

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

7. $\hat{\epsilon}$ est la projection de Y sur l'espace engendré par les colonnes de X.

(a) Vrai (b) Faux

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

8. On note $\hat{\sigma}^2 = \|\hat{\epsilon}\|^2/(n-p)$. Laquelle de ces affirmations est correcte ?
   (a) $p\hat{\sigma}^2/\sigma^2$ suit une loi du χ^2 à p degrés de libertés.
   (b) $(n-p)\hat{\sigma}^2/\sigma^2$ suit une loi χ^2 à $n-p$ degrés de libertés.
   (c) $p\hat{\sigma}^2/\sigma^2$ suit une loi de Student à p degrés de libertés.
   (d) $(n-p)\hat{\sigma}^2/\sigma^2$ suit une loi de Student à $n-p$ degrés de libertés.

   $(n-p)\hat{\sigma}^2/\sigma^2$ suit une loi χ^2 à $n-p$ degrés de libertés.

9. Cet estimateur $\hat{\sigma}^2$ est l'estimateur du maximum de vraisemblance du paramètre σ^2.
   (a) Vrai (b) Faux

   Faux. L'estimateur du maximum de vraisemblance est $\|\hat{\epsilon}\|^2/n$.

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

    Vrai (théorème de Cochran)

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 :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 1 \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = -2.99$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0.32 \qquad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 12.16$$

$$s_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 1.61 \qquad \overline{\log(x)} = \frac{1}{n}\sum_{i=1}^{n}\log(x_i) = 0.95$$

$$s_{\log(x)}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\log(x_i) - \overline{\log(x)})^2 = 0.11$$

$$s_{\log(x)y}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\log(x_i) - \overline{\log(x)})(y_i - \bar{y}) = 1.03$$

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:

   $$\hat{\beta}_1 = \frac{s_{xy}^2}{s_x^2} = \frac{1.61}{0.32} = 5.03$$

   et

   $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -8.02$$

   La droite des MC est $y_i = -8.02 + 5.03 \times x_i$.

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:

   $$\begin{aligned}\rho_{x,y} &= \frac{s_{xy}^2}{\sqrt{s_x^2 s_y^2}} \\ &= \frac{1.61}{\sqrt{0.32 \times 12.16}} = 0.82\end{aligned}$$

   D'où $R^2 = \rho_{x,y}^2 = 0.67$.

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

   On fait la régression $y_i = \beta'_0 + \beta'_1 \log(x_i) + \epsilon_i$.

   Les estimateurs des moindres carrés sont:

   $$\hat{\beta}'_1 = \frac{s^2_{\log(x)y}}{s^2_{\log(x)}} = \frac{1.03}{0.11} = 9.36$$

   et

   $$\hat{\beta}'_0 = \bar{y} - \hat{\beta}'_1 \overline{\log(x)} = -11.89$$

   Le coefficient de corrélation entre $\log(x)$ et y est donné par:

   $$\begin{aligned}\rho_{\log(x),y} &= \frac{s^2_{\log(x)y}}{\sqrt{s^2_{\log(x)} s^2_y}} \\ &= \frac{1.03}{\sqrt{0.11 \times 12.16}} = 0.89\end{aligned}$$

   D'où $R^2 = \rho^2_{\log(x),y} = 0.79$.

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 $\log(X)$.

Exercice 3. On examine l'évolution d'une variable Y en fonction de deux variables x et z. On dispose de n 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 :

$$X^T X = \begin{pmatrix} 50 & 0 & 0 \\ ? & 3.57 & 1.12 \\ ? & ? & 114.22 \end{pmatrix} \quad , \quad \hat{\varepsilon}^T \hat{\varepsilon} = \|\hat{\varepsilon}\|^2 = 45.28 \quad , \quad \hat{\beta} = \begin{pmatrix} -1 \\ -0.1 \\ 2 \end{pmatrix}.$$

1. Donnez les valeurs manquantes dans la matrice. Quelle est la valeur de n ?

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

   Puis, par symétrie, on a:

   $$X^T X = \begin{pmatrix} 50 & 0 & 0 \\ 0 & 3.57 & 1.12 \\ 0 & 1.12 & 114.22 \end{pmatrix}$$

2. Rappelez la définition de $\hat{\varepsilon}$. Calculez $\sum_{i=1}^n \hat{\varepsilon}_i$.

   Par définition, $\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\beta}$. C'est aussi la projection de Y sur l'orthogonal de l'espace engendré par les colonnes de X. Comme le vecteur $\mathbb{1}$ est dans l'espace engendré par les colonnes de X, on a :

   $$\sum_{i=1}^n \hat{\varepsilon}_i = \mathbb{1}^T \hat{\varepsilon} = \langle \mathbb{1}, \hat{\varepsilon} \rangle = \langle \mathbb{1}, P^{X^\perp} y \rangle = 0.$$

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

   On a :

   $$n\bar{x} = \mathbf{1}'x = [X^T X]_{12} = 0 \qquad n\bar{z} = \mathbf{1}'z = [X^T X]_{13} = 0$$

   Donc:

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

   Pour $\bar{y}$, on utilise la question précédente:

   $$\mathbb{1}^T \hat{\varepsilon} = \mathbb{1}^T (Y - \hat{Y}) = 0$$

   Donc:

   $$\bar{y} = \frac{1}{n} \mathbb{1}^T \hat{y}$$

   et

   $$\mathbb{1}^T \hat{y} = \mathbb{1}^T (X\hat{\beta}) = \mathbb{1}^T (-1\mathbb{1} + -0.1x + 2z) = -1 \times n + -0.1 \times 0 + 2 \times 0$$

   D'où $\bar{y} = -1$.

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

   $$\hat{\sigma}^2 = \frac{1}{n-3} \|\hat{\varepsilon}\|^2 = \frac{1}{47} 45.28 = 0.9634$$

5. Sous quelles hypothèses pouvez-vous obtenir un intervalle de confiance pour le paramètre β_1 lié à l'intercept ? Donnez l'expression de cet intervalle de confiance à 90%. En donner une valeur numérique approchée.
   On rappelle la formule de la comatrice : pour une matrice A, on a $[A^{-1}]_{ii} = \det(A_{-ii})/\det(A)$, où A_{-ii} est la matrice A d'où on a retiré la ligne et la colone i. On rappelle également que les quantiles de la loi normale centrée réduite à 90%, 95% et 97.5% sont respectivement de 1.28, 1.64 et 1.96.

   Voir le cours pour les hypothèses.

   $$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 [(X^T X)^{-1}]_{11}}} \sim \mathcal{T}(n-3)$$

   En prenant $\alpha = 0.1$, un intervalle de confiance à 90% est donné par:

   $$\beta_1 \in \left[\hat{\beta}_k \pm t_{n-3}(1 - \alpha/2) \sqrt{\hat{\sigma}^2 [(X^T X)^{-1}]_{11}} \right]$$

   Or:

   $$[(X^T X)^{-1}]_{11} = \det \begin{pmatrix} 3.57 & 1.12 \\ 1.12 & 114.22 \end{pmatrix} / \det(X^T X) = 0.02$$

   Donc:

   $$\sqrt{\hat{\sigma}^2 [(X^T X)^{-1}]_{11}} = \sqrt{0.96345 \times 0.02} = 0.13881$$

   Et, puisque n est grand, la loi de Student à n degré de liberté est bien approchée par la loi normale standard, donc $t_{n-3}(1 - \alpha/2) \approx 1.64$. Finalement:

   $$\beta_1 \in [-1 \pm 1.64 \times 0.13881] = [-1.22765; -0.77235]$$

6. Sans faire de calcul, que peut-on conclure sur la nullité de l'intercept ?

   L'intervalle de confiance à 90% ne contient pas zéro, on ne peut donc rejeter l'hypothèse suivant laquelle l'intercept est nul avec un risque de 10%.

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. Répondez aux questions suivantes en utilisant uniquement les sorties `R` suivantes, sans faire de calcul.

```
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
## -2.202 -1.381 -0.385  0.527  5.080
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   -1.947      0.464   -4.19   0.00061 ***
## x_1            0.529      0.421    1.25   0.22648
## x_2            2.975      0.451    6.60 0.0000045 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.99 on 17 degrees of freedom
## Multiple R-squared:  0.769, Adjusted R-squared:  0.742
## F-statistic: 28.3 on 2 and 17 DF,  p-value: 0.00000387

predict(fit_1,
        newdata = data.frame(x_1 = c(0, 0), x_2 = c(1, 2.5)),
        interval = "prediction", level = 0.95)

##      fit      lwr    upr
## 1 1.0276 -3.32984  5.385
## 2 5.4896  0.69022 10.289
```

1. Donnez l'équation (numérique) de la droite de régression estimée.

   Premier modèle: $\hat{y}_i = -1.947 + 0.529 \times x_{1i} + 2.975 \times x_{2i}$

2. 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 0.23, 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 très faibles.

3. Les intervalles de confiance pour le 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%.

4. On suppose que l'on fixe $X_1 = 0$. Si X_2 vaut 1, quelle est la valeur prédite pour Y ? Au niveau de confiance de 95%, peut-on affirmer que la valeur prédite pour Y dans ce cas est positive ?

   La valeur prédite est 1.02759. L'intervalle de prédiction contient zéro, à 95% on ne peut pas affirmer que la valeur prédite pour Y est positive.

5. Toujours avec $X_1 = 0$, on prend maintenant $X_2 = 2.5$. Quelle est la valeur prédite pour Y ? Peut-on affirmer que la valeur prédite pour Y dans ce cas est positive ?

   La valeur prédite est 5.48959. L'intervalle de prédiction ne contient pas zéro, à 95% on peut donc affirmer que la valeur prédite pour Y est positive.