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) $p \times 1$
2. Dans ce modèle, les Y_i, $1 \leq i \leq n$, sont des variables aléatoires gaussiennes indépendantes.
   (a) Vrai (b) Faux
3. Dans ce modèle, les $\hat{\beta}_k$, $1 \leq k \leq p$, sont des variables aléatoires gaussiennes indépendantes.
   (a) Vrai (b) Faux
4. Laquelle de ces formules est correcte ?
   (a) $Y_i = \sum_{k=1}^{p} \beta_k X_{ik} + \epsilon_i$ (b) $Y_i = \sum_{k=1}^{p} \hat{\beta}_k X_{ik} + \epsilon_i$
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) $\hat{\beta} = (X^TX)^{-1}XY$
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
7. $\hat{\epsilon}$ est la projection de Y sur l'espace engendré par les colonnes de X.
   (a) Vrai (b) Faux
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.
9. Cet estimateur $\hat{\sigma}^2$ est l'estimateur du maximum de vraisemblance du paramètre σ^2.
   (a) Vrai (b) Faux
10. Les estimateurs $\hat{\beta}$ et $\hat{\sigma}^2$ sont indépendants.
    (a) Vrai (b) 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 :

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

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

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

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

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

1. Posez un modèle de régression en précisant les hypothèses, et donnez l'expression des estimateurs des coefficients.
2. Calculez les coefficients en utilisant les données fournies.
3. Rappelez la définition du R^2 du modèle et son interprétation.
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.
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.
6. D'après ces résultats, pouvez-vous préférer un modèle plutôt qu'un autre ?

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 ?
2. Rappelez la définition de $\hat{\varepsilon}$. Calculez $\sum_{i=1}^{n}\hat{\varepsilon}_i$.
3. Calculez les moyennes empiriques $\bar{x}$, $\bar{z}$ et $\bar{y}$.
4. Donnez l'estimateur sans biais de la variance, et calculez-le.
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.

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

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.
2. Que dire de la significativité des coefficients ? Justifiez. On décrira en détails le test associé, avec ses hypothèses.
3. Les intervalles de confiance pour le coefficient associé à `x_2` aux niveaux, respectivement, de 90%, 95% et 99% contiennent-ils zéro ?
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 ?
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 ?