

RÉGRESSION LINÉAIRE ET LOGISTIQUE
FEUILLE 2

EXERCICE 1. INTERVALLES DE CONFIANCES ET p -VALUE

Reprendre le vecteur X utilisé dans la feuille 1 Exercice 4 question (5). On considère le modèle

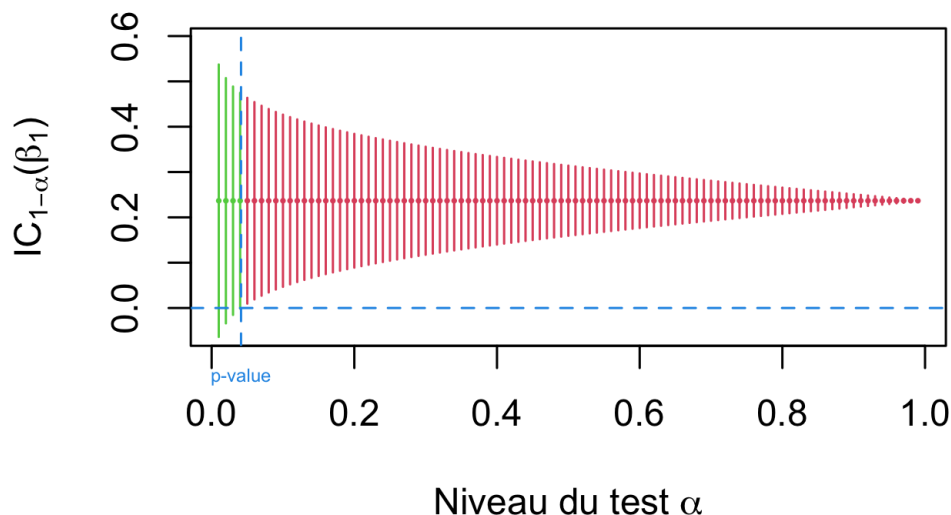
$$y_i = 10 + \beta_1 x_i + \varepsilon_i,$$

où les ε_i sont i.i.d. et $\varepsilon_1 \sim \mathcal{N}(0,1)$. On veut tester $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$.

Partie 1 : On va illustrer sur un exemple le lien entre la p -value et les IC . On considère le jeu de données dans R défini par :

```
set.seed("44")
y<-rnorm(100,10.2*x,1)
```

- (1) Déterminer à l'aide de la fonction `summary` la p -value du test.
- (2) En utilisant la fonction `confint` construire les intervalles de confiance $IC_{1-\alpha}(\beta_1)$ pour α variant dans $[0.01, 0.99]$. On prendra un pas de 0.01.
- (3) Déterminer la plus petite valeur de α telle que $0 \notin IC_{1-\alpha}(\beta_1)$. Interpréter.
- (4) Réaliser le graphique ci-dessous :



Partie 2 : Simulations de la distribution de la p -value sous \mathcal{H}_0 et \mathcal{H}_1 .

- (1) On fixe $\beta_1 = 0$. Simuler un échantillon $(y_i)_{i=1,\dots,n}$ et stocker la valeur de la p -value. Répéter $B = 10000$ fois l'expérience. Tracer l'histogramme de cet échantillon de p -value. Que constate t'on ?
- (2) Refaire la même expérience avec $\beta_1 = 0.2$ et $\beta_1 = 0.5$.
- (3) On fixe $\alpha = 0.5\%, 1\%, 5\%, 10\%$. Donner une estimation de $\mathbb{P}(p\text{-value} < \alpha)$ pour $\beta_1 = 0, 0.2, 0.5$ (on utilisera les deux questions précédentes). A quoi correspond cette probabilité pour ces différentes valeurs de β_1 ? Interpréter les résultats.

EXERCICE 2. RÉGRESSION ORTHOGONALE

On admet que le problème de minimisation associé à la régression orthogonale revient à minimiser

$$S = \sum_{i=1}^n \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{1 + \beta_1^2}$$

On note $\hat{\beta}_{0RO}, \hat{\beta}_{1RO}$ les paramètres du modèle de régression vérifiant ce problème de minimisation

- (1) Vérifier que $S_{XY} - \beta_1 S_{XX} = \sum_{i=1}^n X_i [(Y_i - \hat{Y}) - \beta_1 (X_i - \hat{X})]$.
- (2) Calculer $\frac{\partial dS}{\partial d\beta_0}$ et $\frac{\partial dS}{\partial d\beta_1}$.
- (3) Montrer que $\hat{\beta}_{0RO} = \bar{Y} - \hat{\beta}_{1RO} \bar{X}$. Commenter ce résultat.
- (4) Montrer que $\hat{\beta}_{1RO} = \frac{(S_{YY} - S_{XX}) + \text{sign}(S_{XY}) \sqrt{(S_{YY} - S_{XX})^2 + 4S_{XX}S_{YY}}}{2S_{XY}}$.
- (5) Écrire une fonction R permettant de calculer $\hat{\beta}_{0RO}$ et $\hat{\beta}_{1RO}$.
- (6) En reprenant le schéma de simulation de l'exercice 4 Partie 1 question (5) ajouter à la comparaison la régression orthogonale.

EXERCICE 3. RÉGRESSEURS ALÉATOIRES

On considère le modèle suivant : $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, où les variables ε_i sont i.i.d, centrées, de variance σ^2 et où les régresseurs X_i sont supposés i.i.d. de carré intégrable. On note $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$, $\beta = (\beta_0, \beta_1)^\top$ et $\hat{\beta}$ l'estimateur de β par MCO. On suppose que les vecteurs aléatoires X et ε sont indépendants.

- (1) Exprimer $\hat{\beta} - \beta$ en fonction des vecteurs X et ε .
- (2) En déduire que $\hat{\beta}$ converge presque sûrement vers β lorsque $n \rightarrow \infty$.

EXERCICE 4. MODÈLES DE CHIMIE

Lors d'une expérience chimique, on observe la teneur d'un certain produit à différents instants réguliers allant de 1 à n . Le résultat à l'instant t est noté y_t . On suppose le lien temporel suivant :

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t, \quad t = 1, \dots, n$$

où les variables ε_i représentent les erreurs de mesures supposées aléatoires, centrées, non corrélées et de variance σ^2 . Soit $\hat{\beta}$ l'estimateur de β par MCO.

- (1) Calculer $\text{var}(\hat{\beta})$ et donner sa limite lorsque $n \rightarrow \infty$.
- (2) En déduire le comportement asymptotique en moyenne quadratique de $\hat{\beta}_0$ et $\hat{\beta}_1$.

On se place sous les mêmes hypothèses que la question précédente mais on suppose cette fois-ci que le lien temporel est

$$y_t = \beta_0 + \beta_1 \frac{1}{t} + \varepsilon_t, \quad t = 1, \dots, n.$$

- (1) Calculer $\text{Var}(\hat{\beta})$ et donner sa limite lorsque $n \rightarrow \infty$.
- (2) En déduire le comportement asymptotique en moyenne quadratique de $\hat{\beta}_0$ et $\hat{\beta}_1$.

EXERCICE 5. MODÈLES DE LA FINANCE

On considère un actif financier dont la valeur au temps t est notée X_t . Un modèle classique consiste à supposer que $Z_t = \log(X_t)$ suit un modèle de diffusion.

$$dZ_t = \theta dt + \sigma dW_t$$

où θ et $\sigma > 0$ sont les paramètres à estimer. W_t est un mouvement brownien c'est un processus gaussien centré à accroissements stationnaires et indépendants. On a

$$\mathbb{E}(W_t) = 0 \text{ et } \mathbb{E}(W_t W_s) = \min(t, s).$$

L'équation différentielle stochastique admet pour solution

$$Z_t - Z_0 = \theta t + \sigma W_t$$

Le processus Z_t n'est pas observé en temps continu mais aux temps $t_i = i\Delta$ pour $i = 1, \dots, n$. On pose $t_0 = 0$ et $Y_i = Z_{t_i} - Z_{t_{i-1}}$ pour tout $i = 1, \dots, n$.

- (1) Montre que le vecteur $Y = (Y_1, \dots, Y_n)$ peut être modélisé par un modèle linéaire.
- (2) Montrer que l'estimateur des moindres carrés de θ s'écrit

$$\hat{\theta}_n = (Z_{t_n} - Z_{t_0}) \frac{1}{n\Delta}.$$

- (3) Construire un estimateur sans biais du paramètre σ^2
- (4) Etudier les propriétés asymptotiques de $\hat{\theta}_n$ quand n tend vers l'infini et Δ fixé.

EXERCICE 6. PRÉVISIONS ET INTERVALLES DE PRÉVISION

On reprend les données de l'ozone en fonction de la température à partir du fichier `ozone.txt`.

- (1) Retrouver les estimations des paramètres de la régression en utilisant la fonction `lm`. Commenter les sorties obtenues en faisant un `summary` de l'objet `lm`.
- (2) Ecrire une fonction `R` permettant d'obtenir un intervalle de confiance de la droite de régression. Tracer les données, la droite de régression et son intervalle de confiance au niveau 95%. Commenter.
- (3) À l'aide de la fonction `predict` de `R` (on regardera l'aide) ajouter, sur le graphe précédent, l'intervalle de confiance au niveau 95% des prévisions. Commenter.