Éléments de correction

EXERCICES

Exercise 1

Obs.	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$	$\boldsymbol{X_4}$	$\boldsymbol{Y}$
1	1.0	2.3	-1.1	7.8	4.5
2	0.8	1.9	-0.9	3.2	3.8
3	1.2	2.5	-1.2	6.5	5.1
4	0.9	2.1	-1.0	8.1	4.2
5	1.1	2.4	-1.3	4.9	4.9
6	0.7	1.8	-0.8	5.6	3.7
7	1.3	2.6	-1.4	3.7	5.2
8	1.0	2.2	-1.1	7.1	4.4

Table 1: Données pour les variables explicatives et la variable cible

1. Pourquoi est-il important d'étudier la corrélation entre chaque variable explicative et la variable cible avant de construire un modèle de régression multiple ?
2. Calculer la moyenne mariginale et variance pour X_1 et Y.
3. Calculez la covariance entre X_1 et Y, puis déduisez le coefficient de corrélation linéaire $\rho\,(X_1, Y)$. Interprétez la valeur obtenue en termes de sens, force du lien, et pertinence de X_1 pour modéliser Y.
4. Les quantités suivantes ont été calculées pour les variables explicatives X_2, X_3 et X_4 par rapport à la variable cible Y :

$$\begin{aligned} \mathrm{Cov}(X_2, Y) &= 0.41, & \mathrm{Var}(X_2) &= 0.087 \\ \mathrm{Cov}(X_3, Y) &= -0.37, & \mathrm{Var}(X_3) &= 0.057 \\ \mathrm{Cov}(X_4, Y) &= 0.003, & \mathrm{Var}(X_4) &= 0.86 \end{aligned}$$

- a. Décrivez, pour chaque variable, la nature du lien avec la variable cible Y : sens (positif ou négatif) et force (fort, modéré, faible, quasi-nul).
- b. Justifiez pourquoi il serait **pertinent de ne pas inclure la variable** X_4 dans un modèle de régression multiple visant à expliquer Y.

5. Écrivez l'équation matricielle du modèle de régression multiple reliant Y à X_1, X_2, X_3. Précisez la signification des matrices et vecteurs utilisés.
6.Donnez l'expression matricielle des coefficients estimés $\hat{\beta}$ par la méthode des moindres carrés :
7. Expliquer pourquoi, dans certain cas l'application directe de l'expression matricielle de $\hat{\beta}$ dans les calculs peut poser un problème. Proposer une méthode qui peut être utilisée pour estimer $\hat{\beta}$ de manière plus stable tout en la développant.

Indications pour le calcul de la moyenne, variance et covariance

Pour faciliter les calculs, vous pouvez compléter le tableau suivant à partir des données de X_1 et Y :

Obs.	X_1	Y	X_1^2	Y^2	$X_1 - \bar{X}_1$	$Y - \bar{Y}$	$(X_1 - \bar{X}_1)(Y - \bar{Y})$
1							
2							
3							
4							
5							
6							
7							
8							

Table 2: Tableau à compléter pour le calcul de la moyenne, variance et covariance

Formules à utiliser :

- **Moyenne de** X_1 :

$$\bar{X}_1 = \frac{1}{n}\sum_{i=1}^{n} X_{1i}$$

- **Variance de** X_1 (variance corrigée) :

$$\mathrm{Var}(X_1) = \frac{1}{n-1}\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2$$

- **Covariance entre** X_1 **et** Y :

$$\mathrm{Cov}(X_1, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(Y_i - \bar{Y})$$

Coefficient de corrélation

$$\rho\left(X_1, Y\right) = \frac{\mathrm{Cov}\left(X_1, Y\right)}{\sqrt{\mathrm{Var}\left(X_1\right)} \cdot \sqrt{\mathrm{Var}(Y)}} \approx 0,986.$$

Cette valeur indique une relation linéaire positive très forte entre X_1 et Y. Cela signifie que Lorsque X_1 augmente, Y a tendance à augmenter également. X_1 est donc **une variable explicative très pertinente pour modéliser ou prédire** Y**.**

4. a. Indication : calculez les coefficients de corrélation entre Y et chaque variable explicative. Selon la valeur obtenue, vous pourrez évaluer le sens (positif ou négatif) ainsi que le degré de pertinence de la relation (forte, modérée, faible ou négligeable).

b. **b. Justification de l'exclusion de la variable** X_4 **:**

Le *coefficient de corrélation linéaire* entre X_4 et la variable cible Y, obtenu à partir de la covariance (0,003) et de la variance (0,86), est donné par :

$$\rho(X_4, Y) = \frac{\mathrm{Cov}(X_4, Y)}{\sqrt{\mathrm{Var}(X_4)} \cdot \sqrt{\mathrm{Var}(Y)}} = \frac{0{,}003}{\sqrt{0{,}86} \cdot \sqrt{0{,}77}} \approx 0{,}0037$$

Ce résultat indique une **relation linéaire quasi nulle** entre X_4 et Y.

Ainsi, la variable X_4 n'apporte **aucune information pertinente** pour expliquer ou prédire Y dans un modèle de régression multiple. Il est donc **judicieux de ne pas inclure** X_4 dans le modèle, afin d'éviter d'ajouter du bruit inutile ou de complexifier inutilement l'analyse.

5. Équation matricielle du modèle de régression multiple :

Le modèle de régression multiple reliant la variable cible Y aux variables explicatives X_1, X_2, X_3 s'écrit sous forme matricielle comme suit :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

où :

- $\mathbf{Y} \in \mathbb{R}^{8\times 1}$ est le vecteur des observations de la variable cible ;
- $\mathbf{X} \in \mathbb{R}^{8\times 4}$ est la matrice des données, avec une première colonne de 1 correspondant à l'ordonnée à l'origine (β_0) ;
- $\boldsymbol{\beta} \in \mathbb{R}^{4\times 1}$ est le vecteur des coefficients à estimer ;
- $\boldsymbol{\varepsilon} \in \mathbb{R}^{8\times 1}$ est le vecteur des erreurs aléatoires.

Voici la matrice $\mathbf{X}$ correspondant aux 8 observations données :

$$\mathbf{X} = \begin{bmatrix} 1 & 1.0 & 2.3 & -1.1 \\ 1 & 0.8 & 1.9 & -0.9 \\ 1 & 1.2 & 2.5 & -1.2 \\ 1 & 0.9 & 2.1 & -1.0 \\ 1 & 1.1 & 2.4 & -1.3 \\ 1 & 0.7 & 1.8 & -0.8 \\ 1 & 1.3 & 2.6 & -1.4 \\ 1 & 1.0 & 2.2 & -1.1 \end{bmatrix}$$

6.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

L'estimation du vecteur β par la méthode des moindres carrés est donnée par :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

7. Problèmes liés aux calculs d'inverse de la matrice et de déterminants associés. Solutions suggérées : applications d'une décomposition matricielle (regarder le cours, la décomposition QR est appliquée en details).

Exercise 2

Un service informatique cherche à prédire automatiquement si un ordinateur est en panne (valeur cible y = 1) ou fonctionnel ($y = 0$), en se basant sur deux caractéristiques observées:
- x_1 : taux moyen d'utilisation du processeur (en pourcentage)
- x_2 : présence d'un message d'erreur dans les journaux système ($0 =$ non, $1 =$ oui)
1. Expliquez pourquoi il est approprié d'utiliser une régression logistique dans ce contexte, et donnez l'équation mathématique générale du modèle.
2. La régression logistique utilise la fonction sigmoïde définie par:

$$\sigma(z) = 1/\left(1 + e^{-z}\right)$$

Expliquez pourquoi on utilise cette fonction pour modéliser la probabilité qu'un ordinateur soit en panne.
3. Pour un ordinateur donné, on observe :
- $x_1 = 75$
- $x_2 = 1$

Le modèle de régression logistique estimé par l'équipe technique a pour coefficients :
- $\beta_0 = -50$
- $\beta_1 = 0.6$
- $\beta_2 = 4.5$

- **En vous appuyant sur ce modèle, calculer la probabilité que la machine soit en panne.**
- **Quelle décision le système devrait-il prendre ?**

Réponses

1. 1. Justification de l'utilisation de la régression logistique et équation du modèle

Dans ce contexte, la variable cible y est binaire : elle prend la valeur 1 si l'ordinateur est en panne, et 0 s'il fonctionne normalement. Lorsque la variable à prédire est une variable qualitative à deux modalités (classification binaire), le modèle de régression linéaire classique n'est pas adapté, car il pourrait produire des prédictions en dehors de l'intervalle $[0, 1]$.

La régression logistique est donc plus appropriée. Elle permet de modéliser la probabilité que l'événement $y = 1$ se produise, en s'appuyant sur une fonction logistique (ou sigmoïde) qui garantit des sorties entre 0 et 1.

L'équation générale du modèle logistique est la suivante :

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}} \quad \textbf{avec} \quad z = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

où :

- $\hat{y}$ **est la probabilité estimée que l'ordinateur soit en panne (i.e.** $P(y = 1 \mid x_1, x_2)$**) ;**
- x_1 **: taux moyen d'utilisation du processeur ;**
- x_2 **: présence d'un message d'erreur (0 ou 1) ;**
- β_0**,** β_1**,** β_2 **: coefficients du modèle à estimer.**

Ainsi, ce modèle permet de classifier automatiquement l'état de fonctionnement d'un ordinateur en fonction de deux variables mesurables sous les contraintes probabilistes liées à une variable cible binaire.
2. vous avez la réponse dans le cours

3. Application numérique du modèle

On dispose des valeurs observées $x_1 = 75$, $x_2 = 1$, et des coefficients estimés : $\beta_0 = -50$, $\beta_1 = 0{,}6$, $\beta_2 = 4{,}5$.

Le score est calculé par :

$$z = -50 + 0{,}6 \times 75 + 4{,}5 \times 1 = -0{,}5$$

On obtient la probabilité :

$$\hat{y} = \frac{1}{1 + e^{0{,}5}} \approx 0{,}3775$$

La probabilité que la machine soit en panne est donc d'environ $37,75\%$. Décision à prendre :

Dans une classification binaire standard, le seuil de décision est généralement fixé à $0{,}5$.

$$\hat{y} < 0{,}5 \Rightarrow \textbf{la machine est considérée comme fonctionnelle.}$$

Décision : Le système devrait conclure que la machine est fonctionnelle.

Exercise 3

Un ingénieur souhaite tester si la durée de vie moyenne d'un nouveau composant électronique est différente de 1000 heures, durée garantie par le fabricant. Il prélève un échantillon aléatoire de 12 composants et mesure leurs durées de vie (en heures):

$$1012 \quad 995 \quad 980 \quad 1005 \quad 1023 \quad 988 \quad 1000 \quad 1010 \quad 978 \quad 993 \quad 1002 \quad 997$$

On suppose que les durées de vie suivent une loi normale.
1. Formulez les hypothèses du test
2. Calculez la moyenne empirique $\bar{x}$ de l'échantillon. 3. Proposer une statistique de test en précisant la loi qu'elle suit et le type de test proposé.

4. Donner l'intervalle de confiance sur la moyenne sous l'hypotèse nulle. 5.Que concluez-vous du test? Rejetez-vous ou non H_0 au seuil de 5%?
7. Calculez la valeur p associée à la statistique T. Interprétez cette valeur.

Réponses

Formulation des hypothèses

: Pour déterminer si la durée de vie moyenne du nouveau composant électronique est différente de 1000 heures, nous allons effectuer un test d'hypothèse pour la moyenne.
- Hypothèse nulle (H_0): La durée de vie moyenne est égale à 1000 heures.

$$H_0 : \mu = 1000$$

- Hypothèse alternative (H_1) : La durée de vie moyenne est différente de 1000 heures.

$$H_1 : \mu \neq 1000$$

II s'agit d'un test bilatéral, car nous testons une différence dans les deux sens (supérieure ou inférieure à 1000 heures).

Moyenne de l'échantillon

:

$$\bar{x} = \frac{\sum^{12} x_i}{n} = \approx 998.583 \text{ heures}$$

3. Statistique de test :

Nous souhaitons tester la moyenne d'un échantillon issu d'une loi normale, avec un petit échantillon ($n = 12 < 30$) et une variance inconnue. Dans ce cas, la statistique de test appropriée est le T Score donnée par :

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

qui suit la loi de student de degré de liberté $n - 1 = 11$.
N.B: S^2 est l'estimation de la variance donnée par

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = S^2 = \frac{1}{11}\sum_{i=1}^{12}\left(X_i - \bar{X}\right)^2 \approx 171.63,$$

c à d

$$S = \sqrt{S^2} = \sqrt{171.63} \approx 13.10 \text{ heures}$$

4. Intervalle de confiance pour la moyenne sous H_0 Dans un cas bilatéral sous un test T la forme génerale de la l intervalle de confiance de la moyenne est donnée par

$$IC_{1-\alpha} = \left[\mu_0 - |t_{1-\alpha/2,n-1}| \cdot \frac{S}{\sqrt{n}}, \mu_0 - |t_{1-\alpha/2,n-1}| \cdot \frac{S}{\sqrt{n}}\right]$$

c à d

$$IC_{1-\alpha} = \left[1000 - |t_{1-\alpha/2,11}| \cdot \frac{13,1}{\sqrt{12}}, \mu_0 - |t_{1-\alpha/2,11}| \cdot \frac{13,1}{\sqrt{12}}\right]$$

où $t_{1-\alpha/2,n-1} = t_{0.95,11}$ est la valeur critique de la loi de Student à $n - 1 = 11$ degrés de liberté pour un seuil $\alpha = 5\%$.
N.B: Si 0.95 ne figure pas sur l'une des colonne de votre tableau ca veut dire si ça dépasse l'intervalle traité dans votre tableau, vous pouvez utiliser la relation $t1 - \alpha = -t_\alpha$
On trouve $t_{0.025,11} \approx 2.201$ cela implique que $t_{0.95,11} \approx -2.201$. Finalement, l'intervalle de confiance est

$$\begin{aligned} IC_{95\%} &= \left[1000 \pm 2.201 \cdot \frac{13.10}{\sqrt{12}}\right] \\ &= [1000 \pm 2.201 \cdot 3.78] \\ &= [1000 \pm 8.32] \\ &= [991.68, 1008.32] \end{aligned}$$

Décision: puisque la moyenne empirique calculée de l'échantillon appartient à cet intervalle de confiance, alors la décision serait de ne pas rejeter H_0.
7. Dans l'expression de T score vous remplacer $\bar{X}$ par la valeur de la moyenne de l'échantillon. Vous alez obtenir une valeur de décision:: si elle apartient à l'intervalle de confiance $[-|t_{0.95,11}|, |t_{0.95,11}|$ la décision serait le non rejet de $H0$ sinon on décide de rejeter $H0$.

Exercise 4

Une entreprise affirme que ses employés travaillent en moyenne au moins 42 heures par semaine. Un syndicat souhaite remettre en question cette affirmation. Un échantillon aléatoire de 64 employés a donné une moyenne hebdomadaire de 41

heures, avec un écart-type connu de $\sigma = 4$ heures.

On veut tester cette affirmation au niveau de signification $\alpha = 5\%$.
1. Formuler les hypothèse du test en indiquant sa nature.
2. Proposer une statistique de test en précisant la loi qu'elle suit et le type de test proposé.
3. Déterminer la valeur critique
4.Donner l'air de la région de non rejet de (H_0)
5. Quelle décision à prendre. Interpretez le résultat

Exercise 5

Une entreprise affirme que le poids de ses colis distribué normalement est d'une moyenne de 10 kg . Pour vérifier cette affirmation, un contrôleur prélève un échantillon de n = 20 colis. Les poids (en kg) observés sont les suivants :

$$\{9.8, 10.1, 10.5, 9.9, 10.3, 9.7, 9.8, 10.2, 9.9, 10.1$$
$$9.6, 10.4, 9.7, 10.0, 10.3, 9.5, 9.9, 10.0, 10.2, 9.8\}$$

On souhaite tester l'hypothèse suivante au niveau de signification $\alpha = 0,10$
1.Formuler les hypothèses du test
2. Calculer la moyenne empirique et l'écart-type empirique corrigé de l'échantillon.
3. Indiquer si on doit utiliser un test t ou un test z . Justifier.
4. Calculer la statistique de test correspondante.
5. Calcule la valeur p.
6. Quelle décison à prendre ?
7. Interprèter le résultat en langage courant.

Exercise 6

Une entreprise de livraison affirme que 90% de ses colis arrivent à l'heure. Un client doute de cette affirmation et décide de vérifier en observant un échantillon de 80 livraisons, dont 66 sont arrivées à I'heure.

On souhaite vérifier si la proportion de livraisons à l'heure est inférieure à ce que prétend l'entreprise, au seuil de signification $\alpha = 5\%$.
1. Formulez les hypothèses H_0 et H_1.
2. Vérifiez que les conditions pour utiliser l'approximation normale sont remplies.
3. Calculez la statistique de test (Z-score).
4. Déterminez la valeur critique.
5. Calculez la valeur p.
6. Quelle est la décision ? Justifiez.
7. Interprétez le résultat.

Réponse

1. Formulation des hypothèses

On note p la proportion réelle de colis arrivant à l'heure.

L'entreprise affirme que $p = 0{,}90$, tandis que le client pense que cette proportion est inférieure. Il s'agit donc d'un test unilatéral à gauche.

Les hypothèses sont formulées comme suit :

$$\begin{cases} H_0 : p = 0{,}90 & \textbf{(la proportion de livraisons à l'heure est conforme à l'affirmation)} \\ H_1 : p < 0{,}90 & \textbf{(la proportion réelle est inférieure à celle annoncée)} \end{cases}$$

2. Vérification des conditions d'approximation normale

Pour utiliser l'approximation normale dans un test sur une proportion, il faut vérifier que les deux quantités suivantes sont supérieures ou égales à 5 sous l'hypothèse nulle H_0 :

$$np_0 \geq 5 \quad \textbf{et} \quad n(1-p_0) \geq 5$$

Dans notre cas :

$$n = 80, \quad p_0 = 0{,}90 \Rightarrow \begin{cases} np_0 = 80 \times 0{,}90 = 72 \geq 5 \\ n(1-p_0) = 80 \times 0{,}10 = 8 \geq 5 \end{cases}$$

Les deux conditions sont remplies. On peut donc utiliser l'approximation normale dans le cadre de ce test.

3. Calcul de la statistique de test

Définition de la staitistique Z score :

Sous l'hypothèse nulle $H_0 : p = p_0 = 0{,}90$, la statistique de test est définie par la variable aléatoire :

$$Z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$

où $\hat{p}$ est la proportion de l'échantillon.

Sous H_0, et si les conditions d'approximation sont remplies, on a :

$$Z \sim \mathcal{N}(0,1) \quad \textbf{(loi normale centrée réduite)}$$

4 Valeur observée de la statistique de test:

$$\hat{p} = \frac{66}{80} = 0{,}825, \quad n = 80, \quad p_0 = 0{,}90$$

$$Z_{\textbf{obs}} = \frac{0{,}825 - 0{,}90}{\sqrt{\dfrac{0{,}90 \times 0{,}10}{80}}} = \frac{-0{,}075}{\sqrt{0{,}001125}} \approx \frac{-0{,}075}{0{,}0335} \approx -2{,}24$$

La valeur observée de la statistique de test est donc $Z_{\textbf{obs}} \approx \boxed{-2{,}24}$.

5. Détermination de la valeur critique

Dans un test unilatéral à gauche au seuil $\alpha = 0{,}05$, on rejette H_0 si la statistique observée $Z_{\textbf{obs}}$ est inférieure à la valeur critique z_α telle que :

$$P(Z \leq z_\alpha) = \alpha = 0{,}05$$

et on sait que $z_\alpha = z_{1-\alpha}$ ce qui implique que $Z_{0.005} = -Z0.975$

D'après la table de la loi normale centrée réduite :

$$z_{0.95} \approx 1,645$$

d'où

$$z_{0.05} \approx \boxed{-1{,}645}$$

Autrement dit, on rejette H_0 si $Z_{\text{obs}} < -1{,}645$ ce qui est bien le cas ici.
5. Calcul de la valeur p, décision et interprétation

(a) Calcul de la valeur p :

La valeur observée de la statistique de test est $Z_{\text{obs}} \approx -2{,}24$. La valeur p est la probabilité, sous H_0, d'observer une valeur aussi extrême ou plus extrême que celle-ci, dans le sens du test (ici, à gauche) :

$$p\text{-value} = P(Z \leq -2{,}24) \approx \boxed{0{,}0125}$$

(d'après la table de la loi normale centrée réduite suivant une lecture directe.)

Décision :

On compare la valeur p au seuil de signification $\alpha = 0{,}05$:

$$\text{p-value} = 0{,}0125 < 0{,}05 = \alpha \quad \Rightarrow \quad \text{On rejette l'hypothèse nulle } H_0$$

Interprétation :

Les résultats de l'échantillon suggèrent que la proportion réelle de livraisons arrivant à l'heure est significativement inférieure à 90%. Le doute du client semble donc justifié, et l'affirmation de l'entreprise peut être remise en question.

Exercise 7

Une entreprise affirme que sa machine produit des bouteilles d'eau de 500 mL . Un responsable qualité pense que la machine ne respecte plus cette moyenne. Il prélève un échantillon de 25 bouteilles. On sait que le volume suit une loi normale avec un écart-type connu $\sigma = 5$ mL. Il souhaite effectuer un test au seuil de $\alpha = 5\%$.
1. Formulez les hypothèses H_0 et H_1.
2. Donnez la statistique de test et la région de rejet au seuil α.
3. Supposons que la vraie moyenne soit en réalité $\mu = 497$ mL. Calculez la probabilité de commettre une erreur de 2^{e} espèce (β).
4. En déduire la puissance du test.
5. Interprétez les résultats.