Travaux dirigés - Statistiques

Paul MINCHELLA, Stéphane CHRÉTIEN

Second Semestre 2025-2026

Licence MIASHS

1 Régression linéaire

Exercice 1 – Régression linéaire simple (TD + TP)

On cherche à étudier la relation entre une variable explicative quantitative X et une variable réponse quantitative Y à l'aide d'un modèle de régression linéaire simple. Dans cet exercice, on utilisera le jeu de données `cars`, disponible nativement dans `R`. Il contient un certain nombre d'observations issues de mesures expérimentales :

- `speed` : vitesse d'un véhicule (en miles par heure),
- `dist` : distance de freinage correspondante (en pieds).

Notre objectif ici consiste à modéliser la distance de freinage en fonction de la vitesse.

Question 1 – Statistique descriptive du jeu de données (TP) Charger le jeu de données `cars` et construire le data frame de travail contenant les variables `speed` et `dist` comme indiqué dans le code ci-dessous :

```r
data(cars)

df <- data.frame(
  speed = cars$speed,
  dist  = cars$dist
)
```

1. Afficher la structure du jeu de données à l'aide des commandes `str` et `summary`. Quel est le **nombre d'observation**, de **variables** ? Afficher les premières lignes du DataFrame grâce à la commande `head(df)`.
2. Commenter la nature des variables (quantitatives, unités, ordre de grandeur). Selon la nature, relever **la moyenne, l'écart-type, la médiane, le** min**, le** max**.**
3. Repérer d'éventuelles valeurs atypiques ou déséquilibres visibles. On représente la distribution marginale de chacune des variables à l'aide d'un **histogramme**. Comparer les étendues et la dispersion des variables `speed` et `dist`. Commenter la forme des distributions (symétrie, asymétrie, concentration).

Question 2 – Visualisation des données (TP)

```r
# Nuage de points
plot(df$speed, df$dist, pch = 19, col = "darkgreen",
     xlab = "Vitesse (mph)", ylab = "Distance de freinage (ft)",
     main = "Distance de freinage en fonction de la vitesse")
```

Exécuter le code fourni pour représenter le nuage de points (`speed`, `dist`) à l'aide d'un *scatter plot*.

1. Quelle tendance globale observez-vous ?
2. La relation semble-t-elle parfaitement linéaire ?

Question 3 – Pertinence du modèle (TD)

1. Pourquoi un modèle de régression linéaire semble-t-il pertinent pour décrire la relation entre la vitesse et la distance de freinage ?
2. Rappeler la forme mathématique du modèle de régression linéaire simple.
3. Donner une interprétation concrète des paramètres β_0 et β_1 dans le contexte de cet exercice.

Question 4 – Hypothèses et estimation (TD)

1. Rappeler les principales hypothèses du modèle de régression linéaire.
2. Quelle est la fonction de perte minimisée dans la méthode des moindres carrés ?
3. Rappeler les formules explicites des estimateurs des moindres carrés.

Question 5 – Ajustement du modèle et interprétation (TP + TD) Ajuster le modèle de régression linéaire à l'aide de la fonction `lm`, puis afficher un résumé du modèle.

```
model <- lm(dist ~ speed, data = cars)
summary(model)
```

1. Relever les coefficients estimés. Les interpréter.
2. Afficher grâce à la ligne de code ci-dessous l'intervalle de confiance associé à chaque coefficients estimés. Le modèle est-il significatif ? Rappeler l'expression litérale et les hypothèses nécéssaires à ce dernier.

```
confint(model, level = 0.95)
```

3. Quelles métriques globales de qualité d'ajustement observez-vous ? Rappeler les formules explicites. Relever leur valeur.
4. Le modèle vous semble-t-il expliquer correctement la variabilité de la distance de freinage ? *Justifier en commentant l'intervalle de confiance pour $\widehat{\beta}_1$ et les métriques exploitées.*

Question 6 – Visualisation du modèle et analyse des résidus (TP + TD)

1. Exécuter et commenter le code suivant.

```
# Droite de regression
abline(model, col = "red", lwd = 2)

# Residus (distances verticales a la droite)
segments(x0 = df$speed, y0 = fitted(model), x1 = df$speed, y1 = df$dist,
         col = "purple",lty = 2)

# Legende
legend("topleft", legend = c("Observations", "Droite de régression", "Résidus"),
       col = c("darkgreen", "red", "purple"),
       pch = c(19, NA, NA), lty = c(NA, 1, 2), lwd = c(NA, 2, 1), bty = "n")
```

2. Extraire les résidus du modèle et tracer leur histogramme.

```
res <- residuals(model)

hist(res, breaks = 10, main = "Histogramme des résidus",
     xlab = "Résidus", col = "lightgray")
```

3. La distribution des résidus semble-t-elle centrée ? La forme est-elle compatible avec une loi normale ? Que suggère cette analyse quant à la validité des hypothèses du modèle ?

Fin Exercice 1

Exercice 2 – Régression linéaire simple et intervalle de confiance

On cherche à étudier la relation entre une variable explicative quantitative X et une variable réponse quantitative Y à l'aide d'un modèle de régression linéaire simple. On utilise ici le jeu de données `faithful`, disponible nativement dans `R`. Ce jeu de données contient des mesures effectuées au geyser *Old Faithful* :

— `eruptions` : durée des éruptions (en minutes),

— `waiting` : temps d'attente avant l'éruption suivante (en minutes).

On cherchera à expliquer la durée d'une éruption en fonction du temps d'attente précédent.

Question 1 – Statistique descriptive (TP) Charger le jeu de données `faithful` et construire un data frame `df` contenant les variables `waiting` et `eruptions`.

1. Afficher les premières lignes du DataFrame grâce à la commande `head()`.
2. Décrire les données à l'aide :
   — de la structure du jeu de données,
   — du résumé statistique (`summary`),
   — d'histogrammes pour chacune des variables.
3. Commenter les ordres de grandeur, la dispersion et la forme des distributions.

Question 2 – Visualisation bivariée (TP) Représenter le nuage de points (`waiting`, `eruptions`) à l'aide d'un *scatter plot*.

1. Décrire la tendance observée.
2. Discuter la pertinence d'un modèle linéaire pour décrire la relation entre les deux variables.

Question 3 – Ajustement du modèle et estimation des coefficients (TD + TP) On considère le modèle

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

1. Discuter si les hypothèses du modèle de régression linéaire semblent raisonnables dans le contexte des données observées.
2. Ajuster le modèle de régression linéaire expliquant `eruptions` par `waiting`.
3. Donner les valeurs estimées de $\hat{\beta}_0$ et $\hat{\beta}_1$ et les interpréter.
4. Donner l'intervalle de confiance à 95 % associé au coefficient β_1 et interpréter cet intervalle (*et donc, conclure sur la significativité du modèle*).

Question 4 – Validité du modèle : analyse des résidus (TD + TP) Étudier les résidus du modèle ajusté.

1. Commenter leur distribution (via histogramme et QQ-plot).
2. Vérifier s'ils sont centrés autour de zéro. Commenter.
3. Commenter la validité des hypothèses du modèle linéaire.
4. Commenter également les indicateurs globaux fournis par `summary(model)` :
   - Erreur standard des résidus $\widehat{\sigma} = \sqrt{\frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$;
   - Coefficient de détermination R^2 ;
   - Son penchant ajusté R^2_{ajd}.

Question 5 – Conclusion (TD) Conclure sur l'effet de la variable explicative `waiting` sur la variable réponse `eruptions`.

1. L'effet est-il statistiquement significatif ?
2. Le modèle linéaire fournit-il une description satisfaisante de la relation observée ?

——— *Fin Exercice 2* ———

2 Régression linéaire multiple

Exercice 1

On rappelle que le modèle de régression linéaire multiple peut s'écrire sous la forme

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{RLM}$$

pour n observations $\big\{(X_i^1, \dots, X_i^p, Y_i)\big\}$, $i = 1, \dots, n$, où :

- $\mathbb{X} \in \mathbb{R}^{n\times(p+1)}$ désigne la matrice de design *(la première colonne est constituée de 1 afin de modéliser l'ordonnée à l'origine)* ;
- $\boldsymbol{\beta} \in \mathbb{R}^{(p+1)\times 1}$ est le vecteur des paramètres ;
- $\boldsymbol{Y} \in \mathbb{R}^{n\times 1}$ est le vecteur des observations ;
- $\boldsymbol{\varepsilon} \in \mathbb{R}^{n\times 1}$ est le vecteur d'erreurs.

On suppose que le vecteur d'erreurs $\boldsymbol{\varepsilon}$ vérifie

$$\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0} \qquad \text{et} \qquad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n,$$

c'est-à-dire que ses composantes sont centrées, de variance σ^2, et non corrélées.

1. Retrouver l'expression de l'estimateur des moindres carrés $\hat{\boldsymbol{\beta}}$ pour le modèle linéaire multiple, à partir du principe de minimisation de la somme des carrés des résidus.

On se place désormais dans le cas particulier $p = 1$, correspondant au modèle de régression linéaire simple étudié au Chapitre 1. Le modèle s'écrit alors

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \dots, n. \tag{RLS}$$

L'objectif est de montrer que le modèle (RLS) peut être naturellement interprété comme un cas particulier du modèle de régression linéaire multiple (RLM), et de retrouver les résultats du Chapitre 1 à l'aide du formalisme matriciel.

2. En utilisant l'expression générale de l'estimateur des moindres carrés dans le modèle matriciel, retrouver explicitement les formules des estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ du modèle de régression linéaire simple.

——— *Fin Exercice 1* ———

Exercice 2

On considère toujours un modèle de régression linéaire multiple de la forme

$$Y_i = \beta_0 + \boldsymbol{\beta}^\top X_i + \varepsilon_i, \qquad i = 1, \dots, n,$$

où $X_i \in \mathbb{R}^p$ désigne le vecteur des variables explicatives associé à la i-ème observation.

1. Montrer qu'il est possible de transformer les données X_i en de nouveaux vecteurs $\boldsymbol{x}_{i,*} \in \mathbb{R}^{p+1}$ de telle sorte que le modèle puisse s'écrire sous la forme

   $$Y_i = \boldsymbol{\beta}_*^\top \boldsymbol{x}_{i,*} + \varepsilon_i,$$

   pour un certain vecteur de paramètres $\boldsymbol{\beta}_* \in \mathbb{R}^{p+1}$. On pourra alors supposer, sans perte de généralité, que tous les modèles linéaires considérés sont de ce type.

   Dans la suite, on notera simplement $\boldsymbol{\beta} = \boldsymbol{\beta}_*$ pour alléger les notations.

2. Rappeler la solution du problème d'estimation des moindres carrés dans le modèle linéaire gaussien, notée $\hat{\boldsymbol{\beta}}$. Calculer $\mathbb{E}[\hat{\boldsymbol{\beta}}]$ et interpréter ce résultat.

3. Montrer que la matrice de covariance de l'estimateur $\hat{\boldsymbol{\beta}}$, définie par

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathbb{E}\Big[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\Big],$$

peut s'exprimer explicitement sous la forme

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbb{X}^\top\mathbb{X})^{-1},$$

où $\mathbb{X}$ désigne la matrice de design associée aux vecteurs $\boldsymbol{x}_{i,*}$.

4. En déduire que, dans le cas de la régression linéaire simple, on a

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right),$$

et

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^n X_i^2}{n\sum_{i=1}^n (X_i - \bar{X})^2}\right),$$

où $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

——— *Fin Exercice 2* ———

Exercice 3

On s'intéresse à l'étude du taux de cholestérol sanguin en fonction de plusieurs caractéristiques morphologiques et démographiques. On dispose des données suivantes, mesurées sur $n = 15$ individus :

Cholestérol (ml/100 ml)	**Poids** (kg)	**Âge** (ans)	**Taille** (cm)
354	84	46	180
190	73	20	190
405	65	52	160
263	70	30	155
451	76	57	165
302	69	25	170
288	63	28	175
385	72	36	180
402	79	57	150
365	75	44	165
209	47	24	160
290	89	31	165
346	65	52	165
254	57	23	170
395	59	60	175

On note :

$$Y = \text{Cholestérol}, \qquad X^1 = \text{Poids}, \qquad X^2 = \text{Âge}, \qquad X^3 = \text{Taille}.$$

1) Modélisation Pour tout observation $i = 1, \dots, n$, on considère le modèle de régression linéaire multiple :

$$Y_i = \beta_0 + \beta_1 X_i^{(1)} + \beta_2 X_i^{(2)} + \beta_3 X_i^{(3)} + \varepsilon_i. \tag{M1}$$

1. Rappeler les hypothèses du modèle linéaire.
2. Écrire ce modèle sous forme matricielle $Y = X\beta + \varepsilon$, en précisant la dimension de chaque objet.
3. Rappeler l'expression de l'estimateur des moindres carrés $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p+1}$ avec $p = 3$.

2) Estimation et significativité des coefficients (TP) À l'aide des commandes `lm` et `summary`, estimer les coefficients du modèle.

1. Donner les valeurs estimées de $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$.
2. Donner l'estimation $\hat{\sigma}^2$ de la variance des erreurs.
3. Parmi les variables explicatives X^1, X^2, X^3, indiquer la ou les variables statistiquement significatives au seuil $\alpha = 0.05$, en justifiant à l'aide des p-values.

3) Intervalles de confiance Donner les intervalles de confiance à niveau $1 - \alpha$ avec $\alpha = 0.05$ pour chacun des coefficients du modèle. Interpréter.

4) Significativité globale du modèle À partir des résultats du `summary`, discuter la significativité globale du modèle linéaire multiple *(il faudra également vérifier le caractère centré des résidus, et étudier sa (non)-normalité)*. On précisera l'hypothèse nulle du test de Fisher ; la statistique de test et la conclusion au seuil $\alpha = 0.05$.

5) Discussion et prolongements Le modèle linéaire multiple est-il satisfaisant pour expliquer la variabilité du taux de cholestérol ? Proposer un autre type de modèle ou une modification du modèle existant (transformation de variables, sélection de variables, interaction, etc.) et justifier brièvement ce choix, notamment par les critères qui comparent $R^2, R^2_{\text{adj}}, \text{AIC}, \text{BIC}$. Le comparer au modèle (M1).

——— *Fin Exercice 3* ———