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 $\widehat{\sigma} = \sqrt{\frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$;
   - 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* ———