TP 4 - Analyse des scores de l'étude PISA

Contexte du TP

Les données utilisées dans ces travaux pratiques proviennent de l'étude **PISA 2018** (Programme International pour le Suivi des Acquis des élèves), conduite par l'**OCDE**. Cette enquête évalue les compétences des élèves de 15 ans dans plus de 100 pays, dont la France, en mathématiques, lecture et sciences.

L'objectif de ce TP est d'analyser les **relations entre les performances scolaires** (scores en mathématiques et en lecture), **l'environnement socio-économique** des élèves et leur **genre**. L'étude portera sur un échantillon restreint de pays d'Europe de l'Ouest.

Données utilisées

Nous utiliserons un sous-échantillon des données PISA 2018 disponible sous le nom de fichier `PISA2018subset.csv`, à télécharger depuis Chamilo et à placer dans le répertoire local `data`.

```
PISA2018 <- read.csv("./data/PISA2018subset.csv", stringsAsFactors = TRUE)

# Sélection des pays d'Europe de l'Ouest
pisa <- subset(PISA2018, country %in% c("FRA", "DEU", "BEL", "ITA", "GBR"))

# Sélection des variables d'intérêt
pisa <- pisa[, c("country", "gender", "read", "math", "escs")]

# Suppression des observations avec valeurs manquantes
pisa <- pisa[!(is.na(pisa$read) | is.na(pisa$escs)),]

# Réduction des niveaux de facteur
pisa$country <- droplevels(pisa$country)
```

Pour vérifier le contenu du jeu de données, on peut afficher les premières lignes :

```
head(pisa)
```

```
##     country gender    read    math    escs
## 252     BEL   male 314.349 407.476  0.9249
## 253     BEL female 455.710 495.361 -1.0257
## 254     BEL female 399.831 420.729 -0.7299
## 255     BEL   male 455.316 434.091  0.1347
## 256     BEL   male 738.190 741.595  0.5109
## 257     BEL   male 509.440 530.670  0.6815
```

Description des variables

Le tableau `pisa` contient les variables suivantes :

Variable	Description	Type
`country`	Code à trois lettres du pays d'origine de l'élève (certaines régions sont codées comme pays).	Facteur
`gender`	Sexe de l'élève ("male" ou "female").	Facteur
`math`	Score en **mathématiques**.	Numérique
`read`	Score en **lecture** (compréhension de l'écrit).	Numérique
`escs`	Indice **économique, social et culturel** de l'élève (plus élevé = milieu favorisé).	Numérique

---

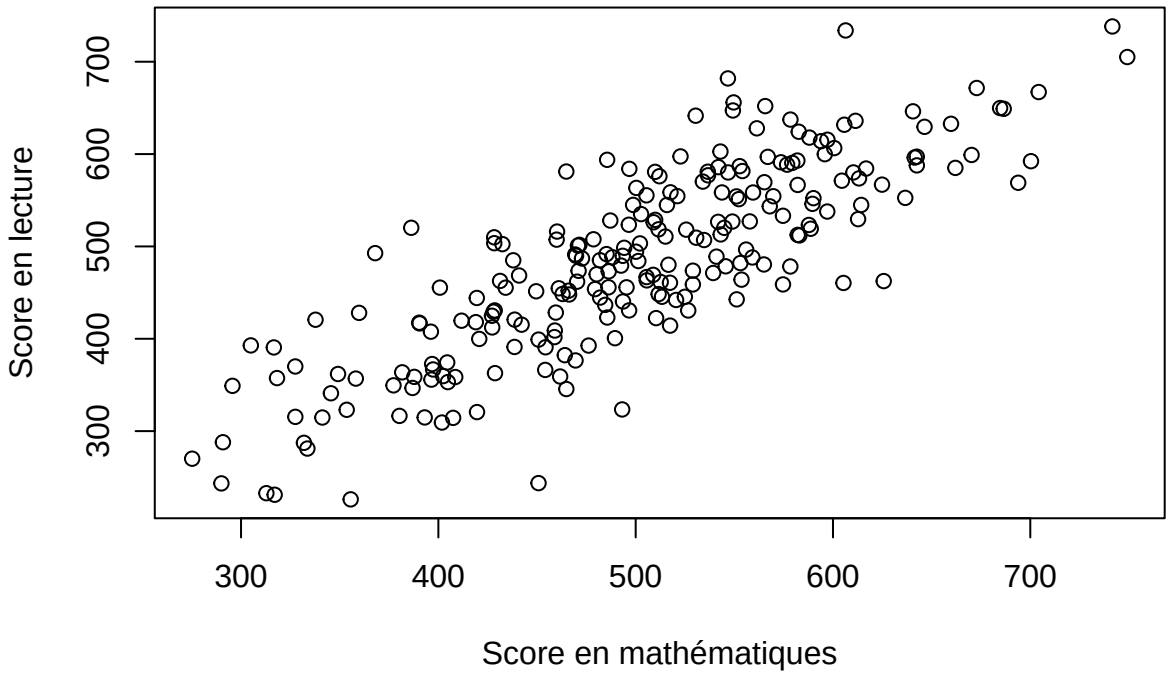
Partie I

Question 1

Représentez graphiquement la relation entre les scores en mathématiques (`math`) et en lecture (`read`) à partir du jeu de données `pisa`.

Réponse. Choisissez une visualisation adaptée permettant de bien comprendre le contenu de la figure et d'observer la corrélation entre ces deux variables.

```
plot(x = pisa$math, y = pisa$read, xlab = "Score en mathématiques", ylab = "Score en lecture")
```



Question 2

Ajustez un modèle de régression linéaire expliquant le score en mathématiques (`math`) par le score en lecture (`read`), à l'aide de la fonction R `lm()`. Présentez le résumé du modèle (commande summary()), et commentez brièvement la relation estimée entre les deux variables.

Vérifiez si les hypothèses du modèle linéaire (en particulier la normalité des résidus) vous paraissent satisfaites, à l'aide d'un graphique (qqnorm() et qqline()).

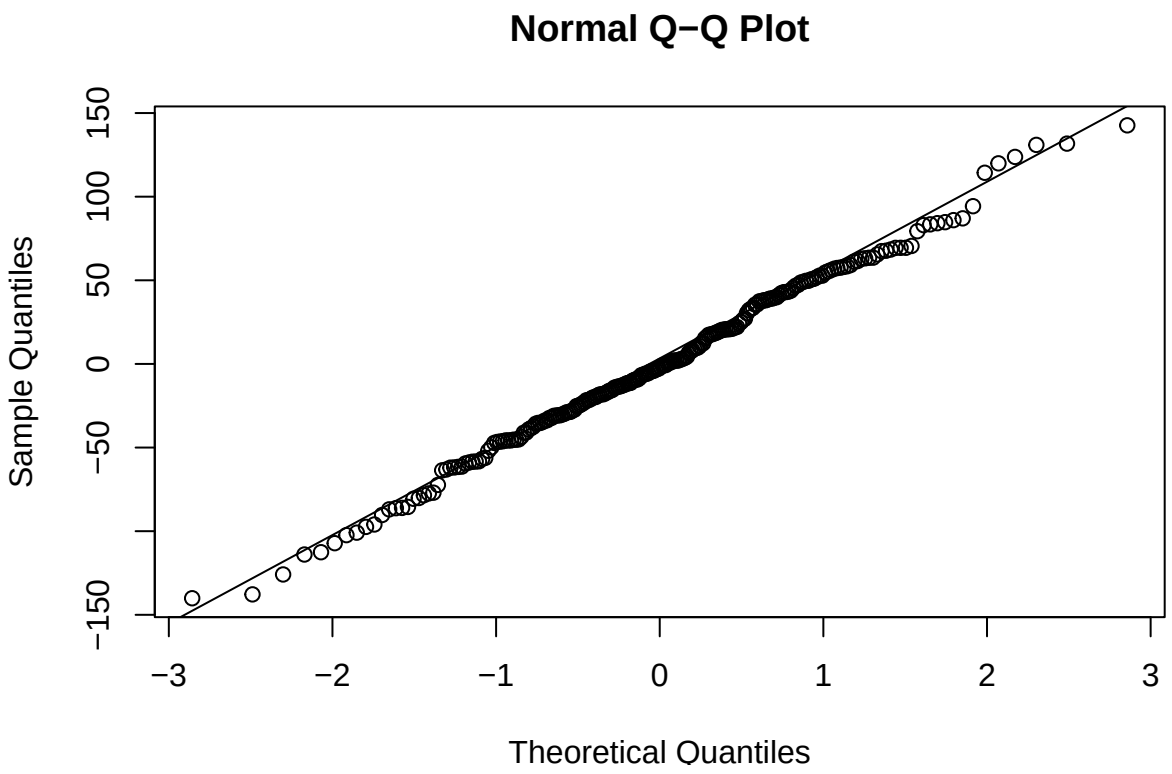
Réponse.

```
mod = lm(formula = pisa$math ~ pisa$read)
summary(mod)
```

```
##
## Call:
## lm(formula = pisa$math ~ pisa$read)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.089  -32.306   -2.239   38.939  142.652
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 136.29762   16.66372   8.179 1.87e-14 ***
## pisa$read     0.74983    0.03364  22.288  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.31 on 232 degrees of freedom
## Multiple R-squared:  0.6817, Adjusted R-squared:  0.6803
## F-statistic: 496.8 on 1 and 232 DF,  p-value: < 2.2e-16
```

L'analyse de corrélation met en évidence une association positive forte ($\rho = 0.83$), entre le score en mathématiques d'un étudiant et son score en lecture. ($R^2 = 0.68$). Cette association est statistiquement significative p-value $2.2 * 10^{-16}$, statistique $= 496.8$; test de Student à 232 degrés de liberté.

```
#On extrait les coefficients déterminé grâce au modèle linéaire.
coef = coefficients(mod)

#On extrait les résidus
residus = mod$residuals

#On affiche les quantiles-quantiles d'une loi normale par apport aux résidus
qqnorm(residus)

#On affiche la droite représentant la droite quantile quantile attendu
qqline(residus)
```



On remarque que l'hypothèse de normalité des résidus semblent être satisfaite au vu du graphe quantile quantile. L'hypothèse de linéarité de même au vu du coefficient de corrélation calculé dans le sommaire du modèle.

Question 3

À l'aide d'une formule de cours, calculez le pourcentage de variance du score en mathématiques (`math`) expliqué par le score en lecture (`read`). Vérifiez ensuite que vous retrouvez le même résultat à l'aide de la commande `summary()` appliquée au modèle linéaire. Reprenez la représentation graphique obtenue à la question 1 et superposez la droite de régression estimée. Ajoutez un encadré sur le graphique indiquant le pourcentage de variance expliquée.

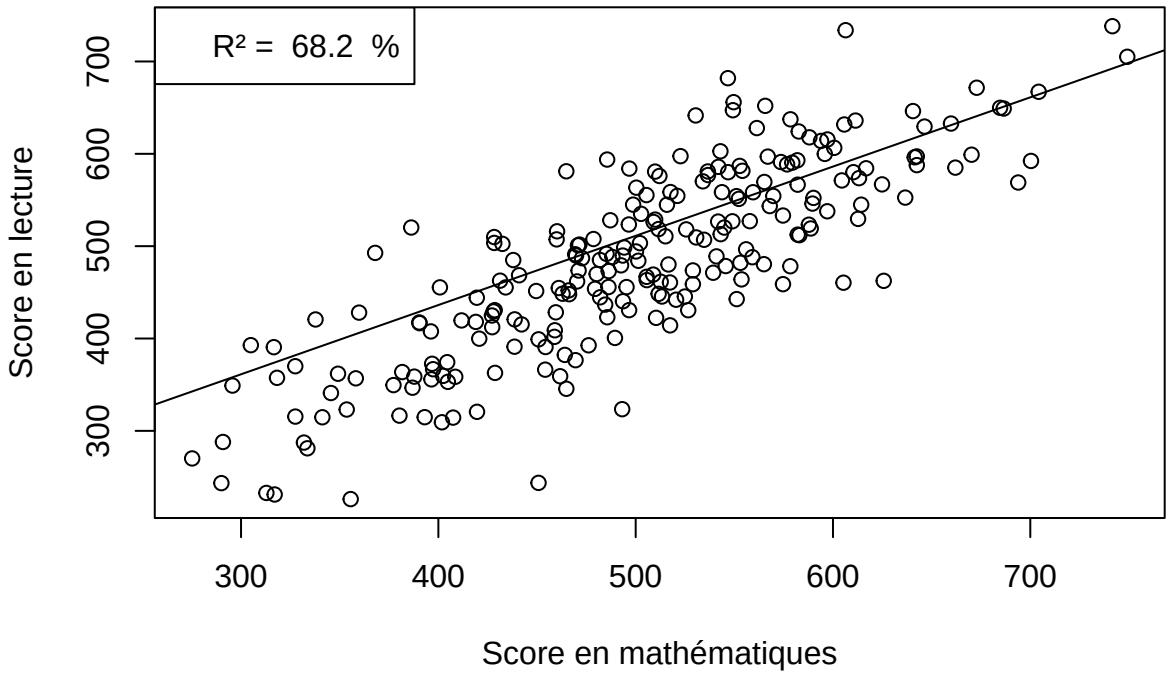
Réponse.

```
y<-pisa$math

#Calcul de R^2 à partir de la formule du cours.
RSS = sum((mod$residuals)^2)
TSS = sum((y-mean(y))^2)

R_2_hat = 1 -  RSS/TSS

#Récupération de la valeur de R^2 à partir du sommaire;
R_2_mod = summary(mod)$r.squared

#On teste l'égalité entre les deux valeurs.
round(R_2_hat,4) == round(R_2_mod,4)
```

```
## [1] TRUE
```

```
plot(x = pisa$math, y = pisa$read, xlab = "Score en mathématiques", ylab = "Score en lecture")
abline(mod)
```

```
legend(x = "topleft" , legend = paste("R² = ", 100 * round(summary(mod)$r.squared,3), " %"))
```



Question 4

À l'aide d'une formule de cours, calculez la pente de la droite de régression du modèle expliquant `math` par `read`. Vérifiez que vous retrouvez le même résultat à l'aide de la commande `coefficients()` appliquée au modèle linéaire.

Réponse.

D'après le cours, la pente de la droite de régression du modèle expliquant `math` par `read` est donnée par la formule : $\frac{Cov(Math,Read)}{Var(Read)}$.

```
#b1, le coefficient de la droite théorique.
b1 = cov(pisa$math, pisa$read)/var(pisa$read)

#Le coefficient déterminé par le la fonction lm.
b1_hat = coefficients(summary(mod))[2]

#On affiche la valeur théorique et on vérifie l'égalité des valeurs.
b1
```

```
## [1] 0.7498265
```

```
cat("La valeur attendu correspond à la valeur donné par le modèle :", round(b1,4) == round(b1_hat,4), "
```

```
## La valeur attendu correspond à la valeur donné par le modèle : TRUE .
```

Question 5

Un groupe d'élèves obtient en moyenne 10 points de plus en lecture que la moyenne générale de la population. Selon le modèle ajusté, de combien de points supplémentaires peut-on prévoir que leur score moyen en mathématiques sera supérieur à la moyenne générale ?

Réponse. Soit $b_0$la valeur de l'intercept et $b_1$le coefficient de la droite de régression. Soit x le score en lecture

de la population moyenne, y celui de mathématique, on s'attend à avoir $y = b_0 + b_1 \times x$. On cherche alors y' - y tel que $y' = b_0 + b_1 \times x'$et $x' = x + 10$. On a donc $y' - y = b_1 \times (x' - x) = b_1 \times 10$

```
b1 = coefficients(summary(mod))[2]

diff_attendue = b1 * 10

cat("On peut prévoir que le score moyen en mathématiques sera supérieur de ", diff_attendue, "points pa
```

```
## On peut prévoir que le score moyen en mathématiques sera supérieur de  7.498265 points par apport à
```

Question 6

Déterminez l'intervalle de confiance à 95 % de l'estimation de la pente de la droite de régression, à l'aide de la commande `confint()`. Indiquez la p-valeur associée au test de significativité de ce coefficient, et interprétez brièvement son sens.

Réponse.

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

```
##                   2.5 %      97.5 %
## (Intercept) 103.4660671 169.1291697
## pisa$read     0.6835434   0.8161096
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = pisa$math ~ pisa$read)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.089  -32.306   -2.239   38.939  142.652
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 136.29762   16.66372   8.179 1.87e-14 ***
## pisa$read     0.74983    0.03364  22.288  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.31 on 232 degrees of freedom
## Multiple R-squared:  0.6817, Adjusted R-squared:  0.6803
## F-statistic: 496.8 on 1 and 232 DF,  p-value: < 2.2e-16
```

L'intervalle de confiance à 95% centré de l'estimation de la pente de la droite est **[0.6835,0.8161]**

La p-valeur du test de significativité du coefficient de la pente de la droite de régression est inférieur à 2×10^6. Le sens de ce test est de vérifier si le score pisa en lecture à un impact significatif sur le score pisa en mathématiques, la p-valeur étant très inférieur au seuil des 5%, on déduit que oui le score pisa en lecture à un impact significatif sur le score pisa en mathématiques.

Question 7

Comment interpréteriez-vous une relation causale entre le score en lecture (`read`) et le score en mathématiques (`math`) ? Cette relation causale est-elle la seule explication possible à la corrélation observée entre les deux

scores ? Expliquez en quoi le statut socio-économique et culturel de la famille de l'élève (`escs`) pourrait influencer simultanément les performances en lecture et en mathématiques.

Réponse. Nous avons montré qu'il existe un lien de corrélation forte entre le score en mathématique et en lecture. Cependant le lien de causalité n'est pas établi. Si il existait un lien de causalité entre les deux, il se pourrait que cela s'explique par la nécessité de comprendre les énoncés/problèmes formulés en français pour résoudre les exercices de mathématiques.

Non, d'autres explications sont possible, notamment que d'autres indicateurs agissent comme facteurs de confusion, en particulier le statut socio-économique.

Le statut socio-économique et culturel de la famille de l'élève (escs) peut influencer les performances en lecture et en mathématiques de l'élèves, que ce soit par la possibilité par les parents d'engager des professeurs particuliers ou par la capacité des parents à aider les enfants dans leurs devoirs. Il existe en effet un lien de **corrélation positive** entre niveau d'études et escs. Certains parents qui n'auraient donc pas le brevet/bac, ne pourront donc pas aider leurs enfants lorsque des difficultés surviennent.

Partie II — Facteur de confusion

On appelle **facteur de confusion** une variable qui influence à la fois la **variable explicative** (ou prédicteur) et la **variable réponse**. Une telle variable peut **fausser l'interprétation** d'une relation causale apparente entre la variable explicative et la variable réponse.

Considérons une variable aléatoire Yexpliquée par deux prédicteurs X_1et X_2selon le modèle suivant :

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \varepsilon,$$

où $b_1 \neq 0$, $b_2 \neq 0$, et où le terme aléatoire ε est indépendant des prédicteurs (X_1, X_2). On suppose en outre que les variables explicatives sont **corrélées** :

$$\text{cor}(X_1, X_2) = \rho_x \neq 0.$$

Dans ce contexte, la variable X_2peut être considérée comme un **facteur de confusion** pour la relation entre Yet X_1, puisqu'elle possède un effet à la fois sur Yet sur X_1, rendant leur lien difficile à interpréter comme strictement causal.

Question 1

On considère comme **vrai** le **modèle linéaire complet** :

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \varepsilon,$$

où εest une variable aléatoire indépendante de X_1et X_2. Supposons que l'on ajuste à la place un **modèle linéaire incomplet** ne tenant pas compte de x_2:

$$Y = a_0 + a_1 X_1 + \varepsilon',$$

où ε'est indépendante de X_1.

On appelle **biais de confusion** la différence entre le coefficient $a_1$dans le modèle incomplet et la vraie valeur $b_1$dans le modèle complet :

$$\text{biais de confusion} = a_1 - b_1.$$

1. Calculez la covariance entre Yet X_1dans le vrai modèle.
2. Déduisez une expression du biais de confusion et montrez qu'il peut s'écrire :

$$\text{biais de confusion} = b_2 \rho_{\text{x}} \frac{\text{sd}(X_2)}{\text{sd}(X_1)},$$

où ρ_{x}est la corrélation entre X_1et X_2.

3. Interprétez cette formule :

- Expliquez dans quelles conditions le biais de confusion est nul.
- Précisez comment le signe du biais dépend de la corrélation entre X_1et X_2et du signe de b_2.
- Expliquez comment varie l'ampleur du biais lorsque les prédicteurs sont standardisés.

Réponse.

1.
$$\begin{aligned} Cov(X_1, Y) &= Cov(X_1, b_0 + b_1 X_1 + b_2 X_2) \\ &= Cov(X_1, b_0) + b_1 Cov(X_1, X_1) + b_2 Cov(X_1, X_2) \\ &= b_1 V(X_1) + b_2 Cov(X_1, X_2) \end{aligned}$$

2. En reprenant la formule du cours pour a_1, et la formule de la question 1 pour $b_1$on a :

$$\begin{aligned} a_1 - b_1 &= \frac{Cov(X_1, Y)}{Var(X_1)} - \frac{Cov(X_1, Y) - b_2 Cov(X_1, X_2)}{Var(X_1)} \\ &= b_2 \frac{Cov(X_1, X_2)}{Var(X_1)} \\ &= b_2 \frac{Cor(X_1, X_2)\sqrt{Var(X_1)}\sqrt{Var(X_2)}}{Var(X_1)} \\ &= b_2\, Cor(X_1, X_2)\, \frac{sd(X_2)}{sd(X_1)} \end{aligned}$$

3. Le biais de confusion si et seulement l'une des trois valeurs suivantes est nulle : $b_2, Cor(X_1, X_2), sd(X_2)$

Si $sd(X_2)$est nulle, l'échantillon n'a pas beaucoup d'intéret.

Si la valeure de $b_2$est nulle, cela signifie que X_2n'a aucun impact sur Y dans le modèle linéaire incomplet est enfaite un modèle linéaire complet par apport à la variable X_2.

Si $Cor(X_1, X_2)$est nulle, cela signifie qu'il n'existe pas de corrélation linéaire entre X_1et X_2, donc ce n'est pas un facteur de confusion dans le cadre d'un modélisation linéaire.

Le signe du biais de confusion dépend du signe de $b_2$et de $Cor(X_1, X_2)$:

Biais de confusion	$b_2 > 0$	$b_2 < 0$
$Cor(X_1, X_2) > 0$	> 0	< 0
$Cor(X_1, X_2) < 0$	< 0	> 0

Le biais de confusion est donc positif si l'impact de X_2sur la variable résultat et du même signe que l'impact de X_1sur X_2. Inversement, si les deux corrélations sont de sens différents, le biais de confusion est négatif.

Si les variables sont standardisés, c'est à dire $Var(X_1) = 1$et $Var(X_2) = 1$on a :

$$\begin{aligned} biais' &= b_2 Cor(X_1, X_2) \frac{sd(X_2)}{sd(X_1)} \\ &= b_2 Cov(X_1, X_2) \end{aligned}$$

Le biais de confusion est donc proportiel à b2, le coefficient de régression entre X_2et la variable résultat et proportionel à $Cov(X_1, X_2)$ ### Question 2

On suppose que la variable `escs` (niveau socio-économique de la famille) pourrait influencer à la fois les scores de mathématiques `math` et de lecture `read`. Peut-on conclure que `escs` agit comme un facteur de confusion entre ces deux variables ? Expliquez votre raisonnement et testez les conditions nécessaires pour établir la présence d'un facteur de confusion.

Réponse.

On pense que la variable `escs` puisse agir comme un facteur de confusion de la relation entre score Pisa en lecture et et en mathématiques. Théoriquement, c'est un bon candidat, tout d'abord car il ne peut pas être la conséquences des bons résultats mais aussi car il parait cohérent qu'il soit une cause.

Empiriquement, on a vérifié dans la première partie la corrélation entre `read` et `math`.

On teste alors les deux conditions de confusion suivantes : `-escs` corrélé à `read` `-escs` corrélé à `math`

Enfin, on compare les deux modèles, en particulier on observera le coefficient de read dans les deux cas.

Corrélation entre `escs` et `math` :

```
cor.test(pisa$math,pisa$escs)
```

```
##
##  Pearson's product-moment correlation
##
## data:  pisa$math and pisa$escs
## t = 6.5176, df = 232, p-value = 4.404e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2792411 0.4965910
## sample estimates:
##       cor
## 0.3933989
```

Corrélation entre `escs` et `read` :

```
cor.test(pisa$read,pisa$escs)
```

```
##
##  Pearson's product-moment correlation
##
## data:  pisa$read and pisa$escs
## t = 5.8307, df = 232, p-value = 1.838e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2402762 0.4644572
## sample estimates:
##       cor
## 0.3575059
```

L'analyse de corrélation met en évidence une association positive entre le statut social d'un étudiant et son score en lecture et en mathématiques.

Comparaison des deux modèles :

```
m1 <- lm(pisa$math ~ pisa$read)          # modèle sans escs
m2 <- lm(pisa$math ~ pisa$read + pisa$escs)   # modèle ajusté sur escs

summary(m1)
```

```
##
## Call:
## lm(formula = pisa$math ~ pisa$read)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.089  -32.306   -2.239   38.939  142.652
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 136.29762   16.66372   8.179 1.87e-14 ***
## pisa$read     0.74983    0.03364  22.288  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.31 on 232 degrees of freedom
## Multiple R-squared:  0.6817, Adjusted R-squared:  0.6803
## F-statistic: 496.8 on 1 and 232 DF,  p-value: < 2.2e-16
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = pisa$math ~ pisa$read + pisa$escs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -138.597  -33.856    0.031   35.971  143.135
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 153.39752   17.44550   8.793 3.41e-16 ***
## pisa$read     0.71326    0.03547  20.110  < 2e-16 ***
## pisa$escs    11.97536    4.15237   2.884   0.0043 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51.5 on 231 degrees of freedom
## Multiple R-squared:  0.6927, Adjusted R-squared:  0.6901
## F-statistic: 260.4 on 2 and 231 DF,  p-value: < 2.2e-16
```

Etant donné la p-valeur de l'impact de pisa$escs (0.0043) ainsi que la diminution du coefficients devant `read` (de 0.7498 à 0.71326), on conclut que `escs` satisfait les conditions statistiques d'un facteur de confusion.

Question 3

Ajustez un modèle de régression multiple où le score en mathématiques (`math`) est expliqué par le score en lecture (`read`) et le niveau socio-économique (`escs`).

1. Estimez l'effet partiel de `read` sur `math` une fois `escs` pris en compte.
2. Comparez cette estimation à celle obtenue dans le modèle simple (`math` ~ `read`) pour évaluer le biais de confusion introduit par l'omission de `escs`.
3. Interprétez : le biais est-il faible, modéré ou élevé ?

Réponse.

Partie III

Question 1

À l'aide du jeu de données `pisa` :

1. Comparez les distributions des scores en mathématiques (`math`) selon le sexe (`gender`) à l'aide d'un boxplot.
2. Testez si la différence de moyenne entre filles et garçons est significative (tests de normalité et test de comparaison de moyennes).

Commandes utiles : `boxplot`, `shapiro.test`, `qqnorm`, `t.test`.

Présentez les résultats et commentez les différences observées.

Réponse.

Question 2

1. Représentez la distribution des scores en lecture (`read`) selon le sexe des élèves (`gender`) à l'aide d'un boxplot.
2. Comparez visuellement les scores moyens des filles et des garçons.
3. Vérifiez, à l'aide des tests appropriés si la différence de moyenne entre les deux groupes est statistiquement significative.
4. Commentez vos résultats : la différence observée est-elle importante sur le plan statistique et/ou sur le plan pratique ?

Réponse.

Question 3

L'analyse du jeu de données met en évidence un apparent paradoxe :

- Les filles obtiennent en moyenne de meilleurs scores que les garçons en compréhension de l'écrit.
- Les scores en lecture sont positivement corrélés aux scores en mathématiques.

On pourrait donc s'attendre à ce que les filles réussissent également mieux en mathématiques, ce qui n'est pas observé dans les données. Comment peut-on expliquer ce paradoxe ?

Réponse.

Partie IV

Une explication possible du paradoxe observé précédemment est que **l'effet du genre sur le score en mathématiques passe par la compréhension de l'écrit**. Autrement dit :

- Les filles réussissent mieux lorsque les énoncés sont complexes (comme dans ce TP),
- tandis que les garçons réussissent mieux lorsque la tâche consiste à résoudre directement une équation difficile (comme dans d'autres contextes).

Pour simplifier les notations, on pose :

- Y: score en mathématiques,
- X: genre,
- M: score en lecture (médiateur).

On considère alors trois modèles linéaires :

1. Score en lecture expliqué par le genre :

$$M = \alpha_1 + aX + \epsilon_1$$

2. Score en mathématiques expliqué par le score en lecture et par le genre :

$$Y = \alpha_2 + bM + cX + \epsilon_2$$

3. Score en mathématiques expliqué par le genre seulement :

$$Y = \alpha_3 + dX + \epsilon_3$$

Question 1

Montrer que l'**effet total** du genre sur le score en mathématiques, d, se décompose ainsi :

$$d = ab + c$$

- Le coefficient $c$représente l'**effet direct** de Xsur Y,
- Le produit abreprésente l'**effet indirect**, passant par le médiateur M.

Ainsi, on dit que Mest un **médiateur** de l'effet du genre sur le score en mathématiques.

Réponse.

Question 2

À l'aide du jeu de données `pisa` :

1. Estimez les coefficients a, b, $c$et $d$des modèles présentés précédemment.
2. À partir de ces estimations, calculez l'effet direct (c)et l'effet indirect (ab)du genre sur le score en mathématiques.
3. Vérifiez que la relation $d = ab + c$est bien respectée dans vos résultats.
4. Analysez et commentez : les effets direct et indirect se compensent-ils ? Quelle est l'interprétation de ce résultat ?

Réponse.

Question 3

L'effet direct $c$est-il statistiquement significatif au niveau de 5 % ? Calculez également un intervalle de confiance à 95 % pour cet effet et interprétez-le.

Réponse.

Question 4

À l'aide de la méthode du bootstrap, estimez un intervalle de confiance à 95 % pour l'effet indirect abdu genre sur le score en mathématiques. Interprétez ce résultat.

Réponse.