TP3: tests d'hypothèse et régression linéaire simple
Statistiques pour Sciences Humaines II

Baptiste Perez et Anna Kiriliouk

Année 2021–2022

Commandes utiles pour le TP3

Commande	Description de la commande
`str()`	donne la structure d'un jeu de données
`table()`	crée un table de contingence pour une ou deux variable(s) qualitative(s)
`subset()`	sélectionne une sous-collection de variables et/ou observations d'un tableau
`mean()`	donne la moyenne d'une variable quantitative
`plot()`	crée un graphique (linéaire ou en nuage de points)
`t.test()`	calcule un intervalle de confiance et la p-valeur d'un test d'hypothèse (pour la moyenne) ; arguments utiles : `conf.level`, `mu`, `alternative`
`sd()`	donne l'écart-type d'une variable quantitative
`qt()`	calcule un quantile de la distribution student t
`pt()`	calcule la fonction de répartition de la distribution student t
`abline()`	trace une ligne sur un graphique
`cor()`	donne la corrélation entre deux variables quantitatives
`lm()`	ajuste un modèle de régression linéaire
`summary()`	donne le résumé d'une variable ou d'un objet créé par `lm`
`predict()`	prédit une valeur à partir d'une droite de régression estimée (c'est-à-dire, à partir d'un objet crée par la fonction `lm`)
`cov()`	donne la covariance entre deux variables quantitatives
`var()`	donne la variance d'une variable quantitative
`points()`	ajoute des points à un graphique déjà réalisé
`which(x == a)`	donne le(s) indice(s) `i` d'un vecteur `x` pour lesquelles `x[i]` est égal à `a`
`fitted()`	donne les valeurs ajustées $\widehat{y}_1, \ldots, \widehat{y}_n$ à partir d'un objet crée par `lm`
`resid()`	donne les résidus $e_1, \ldots, e_n$ à partir d'un objet crée par `lm`

Exemple de la commande `subset` : si `data` est un tableau de données avec variables `y`, `x1` et `x2`, on peut définir un nouveau tableau de données qui ne contient que les variables `y` et `x1` par

```
data2 <- subset(data, select = c(y, x1))
```

Avec `str(data)` on peut voir un résumé des données qui montre également le type de chaque variable : `num` pour numeric (quantitative continue), `int` pour integer (quantitative discrète) ou `factor` (qualitative).

Supposons que `x1` soit une variable qualitative avec `levels` (catégories) `H` et `F` (par exemple, homme et femme), et que `x2` soit une variable quantitative. On peut définir des nouveaux tableaux de données en ne sélectionnant que certaines observations,

```
data3 <- subset(data, x1 == "H")
data4 <- subset(data, x2 > 10)
```

N'oubliez pas que pour obtenir de l'aide sur n'importe quelle fonction, vous pouvez taper `?nomdelafonction` dans la console même.

Données "grades" : lien entre les notes moyennes et les QIs des étudiants

Récupérez le fichier `grades.RDS`, disponible sur webcampus, et enregistrez-le sur votre ordinateur dans un dossier "TP3". Créez un nouveau script `R` et enregistrez-le dans le même dossier. Ouvrez le fichier sur Rstudio en utilisant la commande suivante :

```
data <- readRDS("grades.RDS")
```

Le jeu de données contient les notes moyennes, les QIs et le genre de $n = 76$ étudiants en première année de leurs études.

Exercices

1. Quelle sont les proportions hommes/femmes dans ce jeu de données ? Est-ce que les hommes de cet échantillon ont en général un QI plus ou moins élevée que les femmes ? Répondez à cette question en faisant une graphe et en calculant le QI moyen pour les hommes et pour les femmes. *NB : il peut être pratique de définir un jeu de données ne contenant que les hommes, disons* `dataH`*, et un ne contenant que les femmes, disons* `dataF`*. Utilisez la fonction* `subset()` *décrite ci-dessus.*
2. Comparer les intervalles de confiance de 99% des notes moyennes pour les hommes et pour les femmes.
3. On se demande si le QI moyen des femmes est égal à 110. On suppose que l'écart-type du QI est inconnu. Faites un test d'hypothèse bilatéral avec niveau $\alpha = 0.05$ pour répondre à cette question : écrivez bien l'hypothèse nulle et l'hypothèse alternative, calculez ensuite la région de rejet associée à ce test, et concluez.
4. Pour le test d'hypothèse précédent, quelle est la probabilité de commettre un erreur de première espèce ? Quelle est la probabilité de commettre un erreur de deuxième espèce si, en réalité, le QI moyen des femmes est égal à 105 ?
5. Pour le test d'hypothèse précédent, écrivez la formule pour la p-valeur et calculez-le. Interprétez le résultat, puis vérifier votre réponse grâce à la fonction `t.test()`. *NB : il faut spécifier l'argument* `mu`*, égal à la valeur supposée dans* H_0.
6. Quelles auraient été les hypothèses d'un test unilatéral ? Et quelle est la p-valeur associée ? Vérifier de nouveau votre réponse grâce à la fonction `t.test()`. *NB : il faut également spécifier l'argument* `alternative`.

Pour la suite, on utilise de nouveau le jeu de données complet, regroupant les hommes et les femmes.

7. Quel type de graphe utiliseriez-vous pour visualisez l'association entre les variables `iq` et `grade_average` ? Faites ce graphique en utilisant `iq` comme variable explicative et `grade_average` comme variable à expliquer. L'association visualisé semble-t-elle linéaire ? *NB : chaque fois que vous faites un graphique, n'oubliez pas de donner des noms aux axes et de changer la graduation des axes si nécessaire. Par exemple, ces données sont plus lisibles en mettant* $(0, 20)$ *pour l'échelle des notes moyennes.*

8. Ajoutez au graphique une ligne horizontale représentant la moyenne des `grade_average` et une ligne verticale représentant la moyenne de `iq`. Est-ce que le lien entre les deux variables est positif ou négatif ? Quantifiez la force du lien entre les `iq` et `grade_average`. Interprétez ce coefficient.

9. Créez un nouvel objet appelé `m1` contenant le résultat d'une régression linéaire de `grade_average` en fonction de `iq`. *NB : utilisez le caractère* $\sim$ *, qui peut être lu "par rapport à" ou "en fonction de".*

10. L'objet `m1` contient toutes les informations du modèle linéaire qui vient d'être ajusté. Écrivez l'équation de ce modèle linéaire, c'est-à-dire, donnez les coefficients estimées $\widehat{\beta}_0$, $\widehat{\beta}_1$ de la droite
$$\texttt{grade\_average} = \widehat{\beta}_0 + \widehat{\beta}_1 \times \texttt{iq}$$
Est-ce que l'on peut interpréter le coefficient $\widehat{\beta}_0$?

11. De combien de points est-ce que la note moyenne augmente si le QI augmente de 95 à 100 points ? Et s'il augmente 108 à 113 points ? Est-ce une coïncidence ? Interprétez le coefficient $\widehat{\beta}_1$.

12. Calculez les coefficients $\widehat{\beta}_0$, $\widehat{\beta}_1$ avec les formules du slide 25 (Chapitre 2) et vérifiez que vous obtenez la même chose que pour la question précédente.

13. Créez à nouveau un nuage de points entre `iq` et `grade_average`. Ajoutez-y la droite de régression estimée. Faites-aussi apparaitre les valeurs estimées $\widehat{y}_1, \ldots, \widehat{y}_n$ en rouge.

14. Si un étudiant ne voyait que la droite de régression et non les données réelles, à quelle moyenne de notes pourrait-il s'attendre étant donné que son QI est de 74 ? S'agit-il d'une surestimation ou d'une sous-estimation, et de combien ? En d'autres termes, quel est le résidu de cette prédiction ?