M2-Mathématiques et Applications 2022-2023, Paris-Saclay
Remise à niveau Statistiques
christine.keribin@universite-paris-saclay.fr

Applications

1 Comparaison de moyennes

Charger le package `MASS` (`library`) pour accéder au jeu de données `birthwt`. Sa description est accessible dans la documentation.

1. Scinder (`split`) le jeu de données en deux parties, l'une concernant les mères fumeuses (`smoke=1`), l'autre les non fumeuses (`smoke=0`).
2. Pour chacune de ces deux catégories, représenter l'histogramme (`hist`) du poids des nouveaux-nés (`bwt`), puis tracer les graphes quantiles/quantiles pour la loi normale. Effectuer un test de Kolmogorov-Smirnov (`ks.test`) et un test de Shapiro-Wilks (`shapiro.test`) au niveau 5% et interpréter.
3. Calculer un intervalle de confiance du poids de nouveaux-nés des mères fumeuses de niveau 95%. Vérifier les calculs avec la fonction `t.test` et interpréter les différentes sorties de cette fonction.
4. Effectuer un test d'égalité des variances des deux échantillons, et calculer sa p-value. Construire un intervalle de confiance de niveau 95% pour le rapport des variances. Comparer avec la sortie de la fonction `var.test` et interpréter.
5. Tester l'égalité des poids moyens des nouveaux-nés en fonction du statut fumeur/non fumeur de leur mère. Calculer la p-value du test. Donner un intervalle de confiance de la différence des poids moyens. Vérifier avec une option du logiciel et interpréter.
6. On suspecte que le tabac ait une influence négative sur le poids des nouveaux-nés, ceux de mère fumeuse étant plus chétifs. Que répondez-vous?
7. Vu sous un angle régression, quel modèle reconnaissez-vous? Utiliser la fonction `lm` pour répondre aux questions précédentes.

2 Régression linéaire multiple

Un ingénieur d'une entreprise de semi-conducteurs souhaite modéliser le `gain` d'un appareil électronique en fonction de trois paramètres : résistance `x1` de l'émetteur, résistance `x2` de la base, résistance `x3` de l'émetteur à la base [1].

1. Importez les données depuis le fichier `hFE.csv` dans le data.frame `df`. Interpréter le code suivant

```
names(df)=c("y",paste("x",1:3,sep=""))
```

2. Combien y a-t-il d'observations (`dim`)? Effectuez une analyse univariée (`summary`, `sd`), puis bivariée (`pairs`, `cor`). Vous pourrez représenter graphiquement les corrélations entre variables en utilisant la fonction `corrplot` du package `corrplot`.

[1] R. Myers, D. Montgomery, G. Vining: Generalized linear models (Wiley 2002)

3. Ajustez le modèle avec une régression linéaire, d'abord en calculant de façon directe l'estimateur, puis en utilisant la fonction `lm`.

   Remarque: l'opérateur de multiplication matricielle est `%*%`, la transposée s'effectue grâce à la fonction `t()` et l'inverse grâce à la fonction `solve()`.

   (a) Identifiez les différentes sorties de la fonction `summary` associée au résultat de `lm`.

   (b) La régression est-elle significative ?

   (c) Partagez l'écran en quatre parties (`par(mfrow=c(2,2))`). Tracez le graphe des observations en fonction des valeurs ajustées.

   (d) Tracez le graphe des résidus en fonction des valeurs ajustées: superposez les résidus bruts (`resid`), les résidus standardisés (`stdres`) et studentisés (`studres`), en choisissant un type de point et de couleur par type de résidu. Superposez les lignes horizontales d'ordonnées 2 et -2 (`abline`) et ajoutez une légende (`legend`).

   (e) Tracez le graphe quantile/quantile gaussien (`qqnorm`). Superposez la droite quantile/quantile (`qqline`).

   (f) Donnez un titre à chaque graphe, et un titre à l'ensemble des graphes (`title`).

   (g) Quelle est l'estimation de l'écart-type résiduel? Le recalculer directement à partir des ajustements.

4. Inférence:

   (a) Calculez un intervalle de confiance de niveau 95% pour chacun des paramètres du modèle (`vcov`, `qt`).

   (b) La résistance de la base est-elle significative?

   (c) Prédire (`predict`) par intervalle de confiance de niveau 95% un gain moyen quand `x1=14.5`, `x2=220`, `x3=5.0`. Retrouvez ce résultat par un calcul direct des formules du cours.

   (d) Comparez avec un intervalle de niveau 95% pour une valeur individuelle de gain sous les mêmes conditions.

5. Sélection de variables

   (a) Recherchez le plus petit sous-modèle adéquat (`anova`).

   (b) Tracez sur un même graphe les observations en fonction de la covariable, et superposez la droite de régression. Précisez dans le titre l'expression de la droite d'ajustement.