L2 MIASHS — Statistiques Descriptives 2
Année 2021-22

TD 1 de Statistiques Descriptives 2

Exercice 1 Un père a deux garçons, et s'inquiète de la croissance de son cadet qu'il trouve petit. Il décide de faire un modèle familial à partir des mesures de taille en fonction de l'âge de l'aîné :

age	3	4	5	6	7	8	9	10	11	12
taille	96	104.8	110.3	115.3	121.9	127.4	130.8	136	139.7	144.5

i. Représenter les données sur un graphique et justifier visuellement l'utilisation d'un modèle de régression linéaire simple. Discuter les hypothèses nécessaires dans le cas où on souhaite modéliser par un modèle linéaire Gaussien.

ii. Estimer les coefficients de la régression et tracez sur le graphique la droite de régression estimée.

iii. Représenter les résidus.

iv. Calculer le coefficient R^2 et le coefficient R^2-ajusté. La régression semble-t-elle valable ?

Pour information, les données proviennent des études auxologique du Docteur Sempé dont une partie a été publiée par Abidi et al (1996). Ces données mesurées sur des milliers d'enfants (de 1 mois à 19 ans) ont permis d'établir un modèle de croissance humaine qui fournit les prédictions du carnet de santé. Il s'écrit de la manière suivante :

$$Y = \theta_1 \left[1 - \frac{1}{1 + \left((x + \theta_8)/\theta_2 \right)^{\theta_3} + \left((x + \theta_8)/\theta_4 \right)^{\theta_5} + \left((x + \theta_8)/\theta_6 \right)^{\theta_7}} \right]$$

où θ_1 représente la taille adulte, θ_{x} le temps de grossesse, et les couples $(\theta_2, \theta_3), (\theta_4, \theta_5)$ et (θ_6, θ_7) permettent de modéliser respectivement la phase de croissance initiale (juste après la naissance), la phase de croissance centrale (pré-adolescente) et la phase finale.

Exercice 2 La tableau suivant contient la liste de 14 pays d'Amérique du Nord et d'Amérique Centrale, dont la population dépassait le million d'habitants en 1985.

Observations	pays	taux d'urbanisation	taux de natalité
1	Canada	55.0	16.2
2	costa-Rica	27.3	30.5
3	Cuba	33.3	16.9
4	USA	56.5	16.0
5	El Salvador	11.5	40.2
6	Guatemala	14.2	38.4
7	Haiti	13.9	41.3
8	Honduras	19.0	43.9
9	Jamaique	33.1	28.3
10	Mexique	43.2	33.9
11	Nicaragua	28.5	44.2
12	Trinidade/Tobago	6.8	24.6
13	Panama	37.7	28.0
14	Rep. Dominicaine	37.1	33.1

Pour chaque pays, on mesure le taux de natalité y_i (nombre de naissances annuel pour 1000 habitants) ainsi que le taux d'urbanisation x_i (pourcentage de la population vivant dans des villes de plus de 100000 habitants). On fait l'hypothèse d'un modèle de régression linéaire simple du type $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, c'est-d-dire que le taux de natalité dépend linéairement du taux d'urbanisation.

1. Représenter graphiquement les données.
2. Estimer les paramètres β_0 et β_1 du modèle et tracer la droite de régression correspondante.
3. Calculer le coefficient R^2 de la variance expliquée par le modèle de régression linéaire sur la variance totale.

Exercice 3 Nous allons traiter les 50 données journalières de la concentration en ozone en fonction de la température. Les données se trouvent dans le fichier "ozone.txt". La variable à expliquer est la concentration en ozone, notée "maxO3", et la variable explicative est la température à midi, notée "T12".

1. Commencer par représenter les données à l'aide des commandes suivantes :

   ```
   ozone<-read.table("ozone.txt",header=T)
   plot(maxO3~T12,data=ozone)
   ```

   Une régression linéaire simple semble-t-elle justifiée graphiquement ?
2. Effectuer la régression linéaire à l'aide de la commande

   ```
   reg<-lm(maxO3~T12,data=ozone)
   ```

   et consulter les résultats à l'aide de la commande

   ```
   resume<-summary(reg)
   ```

   Que représente les coefficients de la matrice coefficients ?
3. Tracer l'estimation de la droite de régression, ainsi qu'un intervalle de confiance à 95% de celle-ci grâce aux commandes suivantes :

   ```
   plot(maxO3~T12,data=ozone)
   T12=seq(min(ozone[,"T12"]),max(ozone[,"T12"]),length=100)
   grille<-data.frame(T12)
   ICdte<-predict(reg,new=grille,interval="confidence",level=0.95)
   matlines(grille$T12,cbind(ICdte),lty=c(1,2,2),col=1)
   ```

   Ce graphique permet de vérifier visuellement l'ajustement des données au modèle de régression proposé. Que remarquez-vous ?
4. Représentez le vecteur des résidus.
5. On s'intéresse à présent à la qualité de prévision du modèle. Pour cela, on va tracer un intervalle de confiance des prévisions de la manière suivante :

   ```
   >plot(maxO3~T12,data=ozone)
   >T12=seq(min(ozone[,"T12"]),max(ozone[,"T12"]),length=100)
   >grille<-data.frame(T12)
   >ICprev<-predict(reg,new=grille,interval="pred",level=0.95)
   >matlines(grille$T12,cbind(ICprev),lty=c(1,2,2),col=1)
   ```

Demander à votre professeur de vous expliquer cette notion d'intervalle de confiance intuitivement !

Exercice 4 On veut expliquer la hauteur des eucalyptus en fonction de leur circonférence à partir d'une régression linéaire simple. On dispose de 1737 couples circonférence-hauteur qui se trouvent dans le fichier "eucalyptus.txt".

1. Extraire et représenter les données dans le plan.
2. Effectuer la régression et commenter les résultats obtenus.
3. Calculer le coefficient R^2 et le coefficient R^2-ajusté. Quelle est la qualité de la modélisation par une droite de régression selon les résultats trouvés pour ces coefficients ?

Exercice 5 Essayer la séance https://avehtari.github.io/ROS-Examples/ElectionsEconomy/hibbs.html