Alexandre Verzura Victor Vannobel

Compte rendu projet de modèle statistique

Avril 2023

Base de données Orange

Introduction

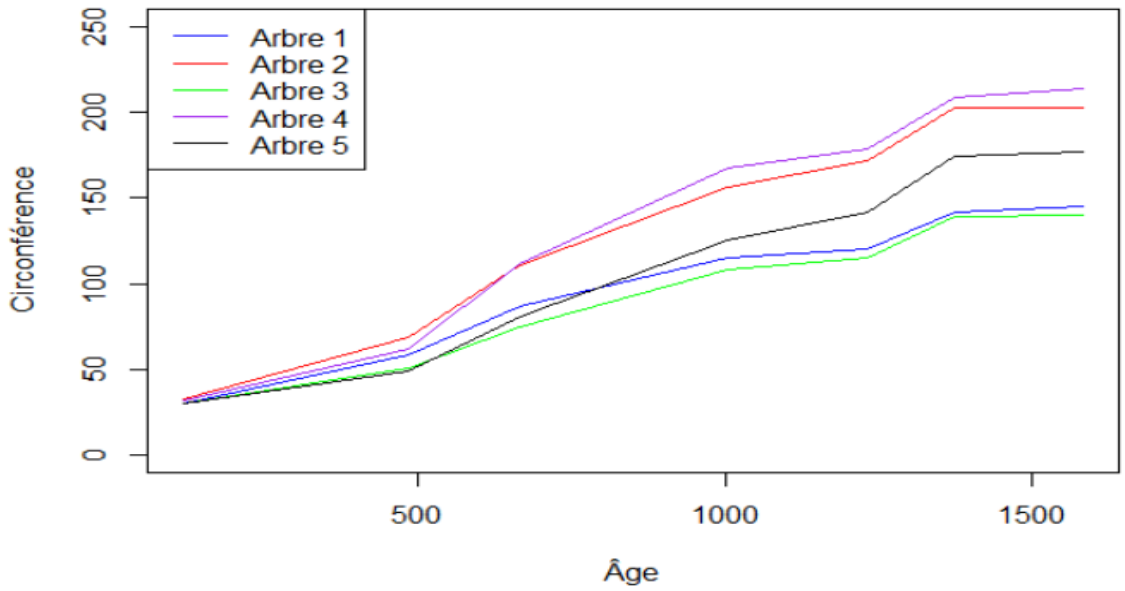
Cette partie traite de l'analyse d'une base de données (contenue dans R) nommée Orange. L'objectif de cette étude est de tester l'indépendance des vitesses de croissance de différents arbres en fonction de leur type.

Question 1 :

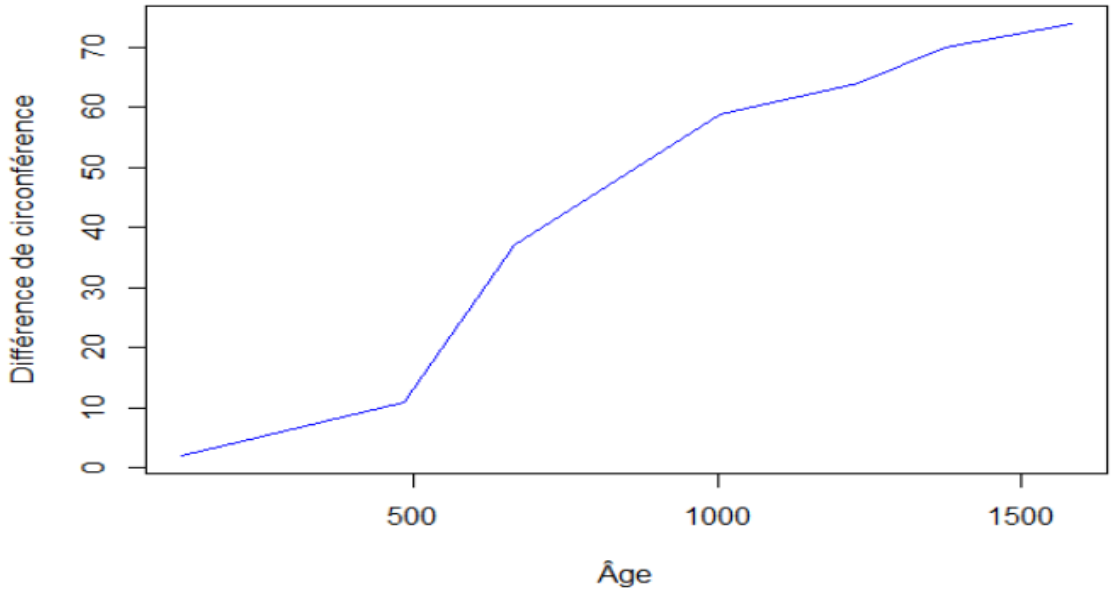
On s'intéresse à la base de données nommée "Orange" inclue dans le language R. Cette base de données est de type "dataframe", elle comporte trois variables, deux variables quantitatives (nommées circonférence et âge qui représentent les circonférences d'un arbre à des âges spécifiques) et une varibale qualitative (nommée Tree qui représente l'identifiant de l'arbre étudié).
Ainsi, le data frame Orange représente l'évolution de la circonférence des arbres en fonction de leur âge. Les unités ne sont pas fournies dans la base de données mais on peut estimer que l'âge est donné en jours et la circonférence en centimètres.

Question 2 :

On s'intéresse ici à la vitesse de croissance des arbres étudiés. Pour ce faire on trace sur un même graphe et pour chaque arbre le nuage de points (avec les points reliés) représentant l'évolution de la circonférence de l'arbre en fonction de son âge (figure 1). Il apparaît que la vitesse de croissance des arbres n'est pas homogène car l'arbre le plus large au 118-ème jour n'est pas le plus large au 1582-éme jour. On remarque également que les arbres ne grandissent pas à la même vitesse. En effet, la différence de taille entre deux arbres augmente en fonction du temps. En prenant comme exemple au 118-ème jour l'arbre le plus fin (en vert sur la figure 1) et le second arbre le plus large (en violet sur la figure 1), on trace l'évolution de la différence de taille (figure 2).
Cette différence croît strictement, démarre à 1 centimètre au 118-ème jour et finit à plus de 70 centimètres au 1582-ème jour.



(Figure 1) Représentation de la circonférence en fonction de l'âge pour différents arbres



(Figure 2) Différence de circonférence entre les arbres 3 et 4 en fonction de leur âge

Question 3 :

On va maintenant chercher un modèle pour expliquer ce jeu de données. A la vue des tracés graphiques (figure 1), il semble naturel d'appliquer un modèle linéaire à ces données. À savoir, en notant $Y_{i,j}$ la circonférence de l'arbre i à l'âge X_j, le modèle recherché serait : $Y_{i,j} = \beta_{0,i} + \beta_{1,i} X_j + \epsilon_j$ avec $\epsilon_j \sim \mathcal{N}(0, \sigma^2)$, $i \in [\![1,5]\!]$ et $j \in [\![1,7]\!]$.
Dans cette optique, à l'aide d'un code implémenté en R (fonction lm()), nous avons calculé les estimateurs des coefficients de régression notés $\hat{\beta}_{0,i}$ et $\hat{\beta}_{1,i}$ (figure 3).

Identifiant de l'arbre (i)	1	2	3	4	5
$\hat{\beta}_{0,i}$	24.44	19.96	19.2	14.64	8.76
$\hat{\beta}_{1,i}$	8.15×10^{-2}	1.25×10^{-1}	8.11×10^{-2}	1.35×10^{-1}	1.11×10^{-1}

(Figure 3) Récapitualitif des coefficients par arbre

Question 4 :

On se demande maintenant si les coefficients $\hat{\beta}_{0,i}$ et $\hat{\beta}_{1,i}$ de chaque arbre sont indépendants deux à deux. On va donc faire un test ANOVA (analyse sur la variance) de niveau $\alpha = 5\%$ avec comme hypothèses H_0 : "Les coefficients ne sont pas indépendants" contre H_1 : "Les coefficients sont indépendants".
Pour pouvoir utiliser la fonction ANOVA de R, on doit dans un premier temps vérifier les hypothèses :
-Les données étant des mesures pour différents types d'arbres et aucune mesure n'appartient à deux types d'arbres simultanéments. L'indépendance des obsesrvations est donc vérifiée.
-A la vue du nuage de points de la question 2 (figure 1), on constate qu'il n'y a aucune variable parasite ou aberrante dans nos données.
-Sachant que le modèle étudié est $Y_{i,j} = \beta_{0,i} + \beta_{1,i} X_j + \epsilon_j$ avec $\epsilon_j \sim \mathcal{N}(0, \sigma^2)$, on en déduit que $Y_{i,j} \sim \mathcal{N}(\beta_{0,i} + \beta_{1,i} X_j, \sigma^2)$ et donc que l'homogénéïté des variances des variables réponses est bien vérifiée.

Les hypothèses étant vérifiées, on applique le test ANOVA à 5% à notre jeu de données. On obtient comme résultat (figure 4) la p-valeur associée au terme d'interaction qui vaut dans ce cas : p-val = 9.402×10^{-5}. La p-valeur étant beaucoup plus faible que le niveau du test ($\frac{\text{p-val}}{\text{niveau}} = \frac{9.402 \times 10^{-5}}{0.05} \simeq 2.10^{-4}$). Nous sommes donc très confiants pour conclure H_1, les coefficients des différents modèles sont indépendants deux à deux.

```
Analysis of Variance Table

Response: circumference
          Df Sum Sq Mean Sq  F value    Pr(>F)
age        1  93772   93772 864.7348 < 2.2e-16 ***
Tree       4  11841    2960  27.2983 8.428e-09 ***
age:Tree   4   4043    1011   9.3206 9.402e-05 ***
Residuals 25   2711     108
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(Figure 4) Résultat du test ANOVA sur la base de données Orange

Étude sur la criminalité aux États-Unis

Partie 1 :

Question 1 :

On considère dans un premier temps le modèle suivant : $(\mathcal{M}_{0,1}) : Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_{i,1}, \sigma^2)$. Avec Y_i le nombre de meurtres commis par million d'habitants dans l'état i ($i \in [\![1, n = 50]\!]$), $x_{i,1}$ le pourcentage de la population vivant sous le seuil de pauvreté dans l'état i.
Ainsi, $(\mathcal{M}_{0,1})$ représente un modèle de régression linéaire pour le nombre de meurtres par million d'habitants en fonction du pourcentage de la population vivant sous le seuil de pauvreté, par état.
En posant $S^2_{n,X} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$, on a que β_0 peut être estimé par $\hat{\beta}_0 = \overline{Y}_n - \hat{\beta}_1\overline{X}_n$ avec $\hat{\beta}_0 \sim \mathcal{N}(\beta_0, \frac{\sigma^2}{n}(1 + \frac{\overline{X}^2_n}{S^2_{n,X}}))$ et que β_1 peut être estimé par $\hat{\beta}_1 = \frac{Cov_n(X,Y)}{S^2_{n,X}}$ avec $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \frac{\sigma^2}{nS^2_{n,X}})$

Question 2 :

Nous avons utilisé le critère d'information d'Akaike (AIC) pour se donner une idée de quel modèle était le plus qualitatif. Voici les résultats obtenus :

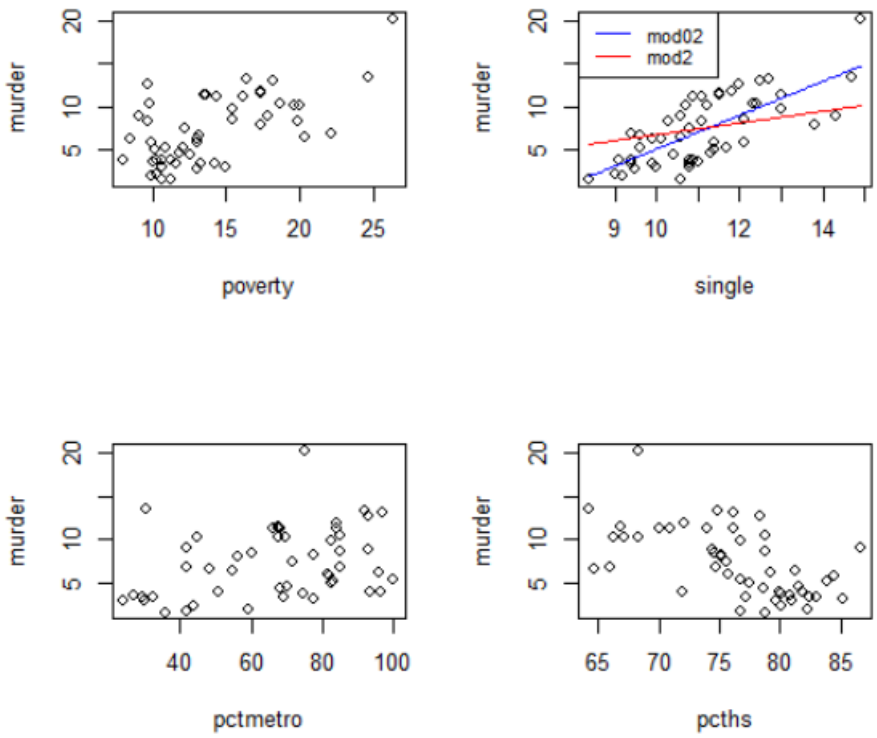
	df	AIC
mod01	3	259.9726
mod02	3	247.3522
mod03	3	279.3271
mod04	3	264.8143
mod1	2	258.2990
mod2	2	265.2735
mod3	2	280.8322
mod4	2	291.0776

(Figure 5) AIC des 8 modèles considérés

Les modèles dont le nom comporte un 0 sont ceux où l'intersect est pris en compte et ceux sans le 0 sont ceux où il n'est pas pris en compte. Le modèle le plus qualitatif est celui dont l'AIC est le plus petit, il s'agit donc du modèle $(\mathcal{M}_{0,2})$ (figure 5). Le modèle $(\mathcal{M}_{0,2})$ semble être celui qui explique au mieux les données.

Question 3 :

D'après la figure 6 (cf page 6), nous apercevons graphiquement que les données de "single" sont les plus adaptées à un modèle linéaire simple. Nous avons ensuite ajouté les droites de regréssion avec et sans intercept et on remarque que la droite avec intercept ajuste plus correctement le nuage de points. Graphiquement, on choisit également le modèle $(\mathcal{M}_{0,2})$.



(Figure 6) Visualisations des différents nuages de points

Question 4 :

Nous utilisons maintenant le modèle 02. On souhaite effectuer le test suivant : $H_0 : \beta_2 \leq 0$ contre $H_1 : \beta_2 > 0$ au risque $\alpha = 5\%$.

Nous utilisons la statistique de test suivante : $T_n = \sqrt{nS^2_{n,X}}\dfrac{\hat{\beta}_2 - \beta_2}{\sqrt{\hat{\sigma}^2_n}}$. Sous H_0, T_n suit une loi de Student à 48 degrés de liberté, c'est-à-dire $T_n \sim \mathcal{T}(48)$.
Théoriquement, la zone de rejet est $\mathcal{R} = \{T_n \geq t_{1-\alpha,48}\}$ avec $t_{1-\alpha,48}$ tel que $\mathbb{P}(T_n \leq t_{1-\alpha,48}) = 1-\alpha$.
Après calculs, $\mathcal{R} = \{T_n \geq 1.68\}$ et $T_n = 7.36 \in \mathcal{R}$, on rejette donc H_0 et on décide $\beta_2 > 0$.
Sur R, on trouve $\hat{\beta}_2 = 1.9664$ (figure 7).

```
Call:
lm(formula = uscrimes[, 1] ~ uscrimes[, 3])

Residuals:
    Min      1Q  Median      3Q     Max
-4.7292 -2.0126  0.1007  2.5683  5.5155

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -14.5143     2.9944  -4.847 1.35e-05 ***
uscrimes[, 3]   1.9664     0.2672   7.358 2.08e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.759 on 48 degrees of freedom
Multiple R-squared:  0.5301,    Adjusted R-squared:  0.5203
F-statistic: 54.15 on 1 and 48 DF,  p-value: 2.079e-09
```

(Figure 7) Résultats de la régression pour $(\mathcal{M}_{0,2})$

Partie 2 :

Question 1 :

Nous pouvons écrire $Y = X\beta + \epsilon$ avec Y le vecteur des variables à expliquer de taille (50,1), X le vecteur des variables explicatives de taille (50,3) avec que des 1 en première colonne, pour l'intercept, β le vecteur des coefficients de régression de taille (3,1) et ϵ le vecteur des erreurs de taille (50,1).

Question 2 :

L'estimateur des moindres carrés de β est $\hat{\beta} = \underset{\beta \in \mathbb{R}^3}{\operatorname{argmin}} ||Y - X\beta||^2$. La forme matricielle de $\hat{\beta}$ est $\hat{\beta} = (X^tX)^{-1}X^tY$. Après calculs, $\hat{\beta} = (-14.56, 0.36, 1.52)^t$.

Question 3 :

Nous allons exprimer SR en fonction de Y^tY, X^tY et de $(X^tX)^{-1}$.

$$\begin{aligned}
SR &= \sum_{i=1}^{50} (Y_i - \hat{Y}_i)^2 \\
&= Y^tY - 2Y^t\hat{Y} + \hat{Y}^t\hat{Y} \\
&= Y^tY - 2Y^tX\beta + \beta^tX^tX\beta \\
&= Y^tY - 2Y^tX(X^tX)^{-1}X^tY + Y^tX(X^tX)^{-1}X^tX(X^tX)^{-1}X^tY \\
&= Y^tY - 2Y^tX(X^tX)^{-1}X^tY + Y^tX(X^tX)^{-1}X^tY \\
&= Y^tY - (X^tY)^t(X^tX)^{-1}X^tY
\end{aligned}$$

Après calculs, SR = 270.5672.

Question 4 :

La somme du carré des résidus sur la variance du modèle suit une loi du Khi-deux à $n - p - 1$ degrés de liberté, avec $n = 50$ le nombre d'observations et $p = 2$ le nombre de variables explicatives. On en déduit que $SR \sim \sigma^2\chi^2(47)$.

Question 5 :

Nous avons que $\sigma_n^2 = \dfrac{SR}{47}$. En utilisant la formule de la question 3, nous obtenons après calculs que $\sigma_n^2 = 5.76$.

Question 6 :

Nous allons maintenant effectuer le test $H_0 : \beta_1 = \beta_2 = 0$ contre $H_1 : \exists j \in \{1,2\}, \beta_j \neq 0$ au risque $\alpha = 5\%$.

Nous allons utiliser la statistique $T_n = \dfrac{R^2/p}{(1-R^2)/(n-p-1)}$ avec p le nombre de variables

explicatives. Ici $n = 50$ et $p = 2$. Sous H_0, T_n suit une loi de Fisher de paramètres $(2,47)$,c'est-à-dire $T_n \sim \mathscr{F}(2,47)$. Théoriquement, la zone de rejet est $\mathscr{R} = \{T_n \geq f_{1-\alpha,2,47}\}$ avec $f_{1-\alpha,2,47}$ tel que $\mathbb{P}(T_n \leq f_{1-\alpha,2,47}) = 1-\alpha$.
D'après les données, on trouve $R^2 = 0.65$ et après les calculs, $\mathscr{R} = \{T_n \geq 3.2\}$ et $T_n = 44.05 \in \mathscr{R}$, on rejette H_0 et on décide donc que les variables expliquées Y_i suivent le modèle $(\mathscr{M}_{0,1,2})$.

Question 7 :

Nous allons maintenant effectuer le test $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$ au risque $\alpha = 5\%$.
Nous allons utiliser la statistique $T_n = \dfrac{\hat{\beta}_1}{\sqrt{\sigma_n^2 (X^t X)^{-1}_{2,2}}}$. Sous H_0, T_n suit une loi de Student à 47 degrés de liberté, c'est-à-dire $T_n \sim \mathscr{T}(47)$ (puisque les bruits sont gaussiens i.i.d. et d'après le théorème de Cochran, $\hat{\beta}_1$ et $\hat{\sigma}^2$ sont indépendants). La zone de rejet est $\mathscr{R} = \{|T_n| \geq t_{1-\frac{\alpha}{2},47}\}$.
Après les calculs, $\mathscr{R} = \{T_n \geq 2.01\}$ et $T_n = 4.06 \in \mathscr{R}$, on rejette H_0 et on décide que $\beta_1 \neq 0$.

Question 8 :

Un intervalle de confiance à $\alpha = 95\%$ pour β_1 est donné par :
$IC = \left[\hat{\beta}_1 - t_{1-\frac{\alpha}{2},47}\sqrt{\sigma_n^2 (X^t X)^{-1}_{2,2}},\ \hat{\beta}_1 + t_{1-\frac{\alpha}{2},47}\sqrt{\sigma_n^2 (X^t X)^{-1}_{2,2}}\right]$.
Après calculs, $IC = \left[0.1814596,\ 0.5378596\right]$.