TD9: Régression linéaire

Probabilités et Statistiques pour l'Ingénieur
M1 Info
Vittorio Perduca

2018-2019

Exercice 1. L'étude statistique ci-dessous porte sur les poids respectifs des pères et de leur fils aîné.

Père	65	63	67	64	68	62	70	66	68	67	69	71
Fils	68	66	68	65	69	66	68	65	71	67	68	70

Voici les résultats numériques que nous avons obtenus :

$$\sum_{i=1}^{12} p_i = 800, \quad \sum_{i=1}^{12} p_i^2 = 53418, \quad \sum_{i=1}^{12} p_i f_i = 54107, \quad \sum_{i=1}^{12} f_i = 811, \quad \sum_{i=1}^{12} f_i^2 = 54849.$$

avec p_i = poids père i-ème et f_i = poids fils i-ème.

1. Calculez la droite des moindres carrés $f = \hat{\alpha} + \hat{\beta} p$ du poids des fils en fonction du poids des pères.
2. Estimer les paramètres du modèle à l'aide de la fonction `lm()`.
3. Tester si les paramètres α et β sont nuls. Indication : on regardera la sortie de `summary(mod)` où `mod` est le modèle défini dans à la réponse précédente par `lm`.
4. Quel est le pourcentage de la variabilité totale qui est expliqué par le modèle ? Au vu de ce résultat, que pouvez-vous conclure sur la qualité du modèle ?
5. Représenter graphiquement les données et la droite des moindres carrés.
6. Calculez la droite des moindres carrés du poids des pères en fonction du poids des fils.

Exercice 2. Nous souhaitons exprimer la hauteur y (en pieds) d'un arbre d'une essence donnée en fonction de son diamètre x (en pouces) à 1m30 du sol. Pour cela, nous avons mesuré 20 couples (diamètre,hauteur) et effectué les calculs suivants : $\bar{x} = 4.53$, $\bar{y} = 8.65$ et

$$\sum_{i=1}^{20} (x_i - \bar{x})^2 = 219.4, \quad \sum_{i=1}^{20} (y_i - \bar{y})^2 = 44.8, \quad \sum_{i=1}^{20} (x_i - \bar{x})(y_i - \bar{y}) = 75.4.$$

1. On note $y = \hat{\beta}_0 + \hat{\beta}_1 x$ la droite de régression. Calculer $\hat{\beta}_0$ et $\hat{\beta}_1$.
2. Donner et commenter une mesure de la qualité de l'ajustement des données du modèle.

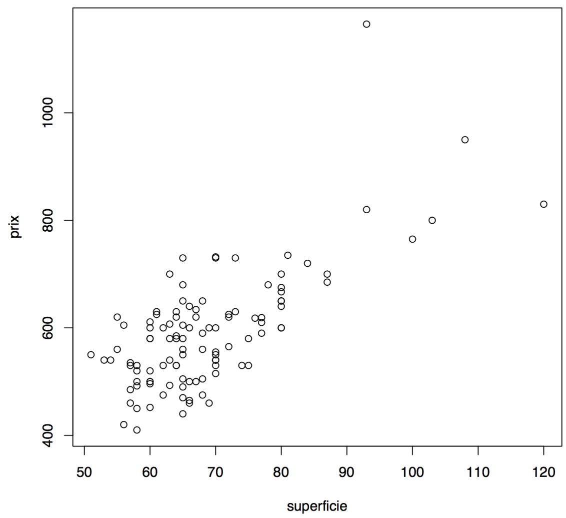

Figure 1 – Prix de location des appartements en fonction de leur superficie.

Exercice 3. En juin 2005, on a relevé dans les petites annonces les superficies (en m2) et les prix (en milliers d'euros) de 108 appartements de type T3 à louer sur l'agglomération de Rennes : voir Fig. 1.

1. Proposer un modèle permettant d'étudier la relation entre le prix des appartements et leur superficie. Préciser les hypothèses de ce modèle.
2. On estime le modèle de régression linéaire avec `R`. D'après le listing du tableau de la Fig. 2, donner les estimations des paramètres.
3. Tester si les paramètres sont nuls ou pas.
4. Ecrire l'équation de la droite des moindres carrés.
5. Est-ce que la superficie joue un rôle sur le prix des appartements de type 3 ? Considérez-vous ce rôle comme important ?
6. Donner une mesure de la qualité du modèle. Donner une estimation du coefficient de corrélation entre le prix et la superficie d'un appartement T3.
7. Dans l'échantillon dont on dispose, comment savoir quels sont les appartements "bon marché" du seul point de vue de la surface ?

D'après *A. Guyader, Régression linéaire.*

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 134.3450    45.4737   2.954  0.00386
Superficie    6.6570     0.6525  10.203  < 2e-16

Residual standard error: 77.93 on 106 degrees of freedom
Multiple R-Squared: 0.4955,     Adjusted R-squared: 0.4907
F-statistic: 104.1 on 1 and 106 DF,  p-value: < 2.2e-16
```

FIGURE 2 – Prix en fonction de la superficie : résultats de la régression linéaire simple (sortie R).

Rappels modèle de régression linéaire. On considère des couple de points (x_i, y_i), $i = 1, \dots, n$ et on souhaite modéliser leur relation par le modèle de régression linéaire simple

$$y_i = \alpha + \beta x_i + \epsilon_i, \text{ pour tout } i$$

avec $(\epsilon_i)_{=1,\dots,n}$ erreurs i.i.d. avec $E[\epsilon_i] = 0$ et $V(\epsilon_i) = \sigma^2$ pour tout i. α, β sont deux paramètres à estimer. On a $E[Y|X = x] = \alpha + \beta x$ et $V(Y) = \sigma^2$, avec Y et X variables aléatoires de réalisations $y_1, \dots, y_n$ et $x_1, \dots, x_n$ respectivement.

Pour α, β on a les estimateurs des moindres carrés ordinaires

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i y_i) - n\bar{x}\bar{y}}{\sum_{i=1}^n (x_i^2) - n\bar{x}^2}$$

d'où l'équation de la droite des moindres carrés ordinaires

$$y = \hat{\alpha} + \hat{\beta}x.$$

L'estimateur non biaisé de σ^2 est

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon}_i^2$$

avec $\hat{\epsilon}_i = y_i - \hat{y}_i$ (résidus).

Le coefficient de détermination est

$$R^2 = \rho_{X,Y}^2 = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \right)^2.$$

R^2 est la fraction de la variation totale expliquée par le modèle :

$$R^2 = \frac{SCE}{SCT}$$

avec

$$SCT = SCE + SCR$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

Si $R^2 = 1$, le modèle explique tout, si $R^2 \simeq 0$ le modèle est inadapté.

Solutions.

Exercice 1.

```
> p=c(65, 63, 67, 64, 68, 62, 70, 66, 68, 67, 69, 71)
> f=c(68, 66, 68, 65, 69, 66, 68, 65, 71, 67, 68, 70 )
> #Q1
> (beta<-(sum(p*f)-12*mean(p)*mean(f))/(sum(p^2)-12*mean(p)^2))
> (54107-12*(800/12)*(811/12))/(53418-12*(800/12)^2)
> (alpha<-mean(f)-beta*mean(p))
> (811/12)-0.48*(800/12)
> #Q2
> mod=lm(f~p)
> mod
> #Q3-Q4
> summary(mod)
> cor(f,p)^2
> #Q5
> plot(f,p)
> abline(mod)
> #Q6
> b=(sum(p*f)-12*mean(p)*mean(f))/(sum(f^2)-12*mean(f)^2); b
> (54107-12*(800/12)*(811/12))/(54849-12*(811/12)^2)
> a=mean(p)-b*mean(f); a
> (800/12)-1.04*(811/12)
> mod2=lm(p~f)
> summary(mod2)
> #alpha/beta; 1/beta
> #
> par(mfrow=c(2,1))
> plot(f,p,xlim=c(60,75),ylim=c(60,75))
> abline(mod)
> plot(p,f,xlim=c(60,75),ylim=c(60,75))
> abline(mod2,xlim=c(60,75),ylim=c(60,75))
```

Exercice 2.

```
> beta1=75.4/219.4; beta1
> beta0=8.65-beta1*4.53; beta0
> R2=(75.4)^2/(219.4*44.8); R2
```