HEG Genève
Prof. S. Varone

Corrigé série 12

Problème 1 Attitude envers la ville

Un chargé d'études veut expliquer l'attitude d'individus envers leur ville de résidence à partir du critère de durée de résidence dans cette ville. L'attitude est mesurée sur une échelle de onze points (avec 1 = n'aime pas la ville, 11 = adore la ville), et la durée de résidence est mesurée par le nombre d'années que l'individu a passé dans la ville. Lors d'un pré-test sur douze individus, les données suivantes ont été recueillies :

individu	1	2	3	4	5	6	7	8	9	10	11	12
attitude	6	9	8	3	10	4	5	2	11	9	10	2
durée	10	12	12	4	12	6	8	2	18	9	17	2

a) Définir la variable dépendante y et la variable indépendante x.

b) Dessiner le diagramme de dispersion.

c) Calculer r_{xy}, le coefficient de corrélation. Commenter.

d) Calculer R^2, le coefficient de détermination. Commenter.

e) Calculer les coefficients de la droite de régression.

f) Calculez la statistique de test sur le coefficient de corrélation, en supposant qu'il existe une corrélation linéaire positive entre l'attitude et la durée.

g) On donne la p-valeur $= 3.773 \cdot 10^{-6}$. Quel sera la conclusion du test ?

Réponse:

a) y = attitude, x = durée

b) Figure : voir plus loin.

c) Le coefficient de corrélation est calculé comme suit :

$$\begin{aligned}
\bar{x} &= \tfrac{10+12+12+4+12+6+8+2+18+9+17+2}{12} = \tfrac{112}{12} = \tfrac{28}{3} = 9.\bar{3} \\
\bar{y} &= \tfrac{6+9+8+3+10+4+5+2+11+9+10+2}{12} = \tfrac{79}{12} = 6.58\bar{3}
\end{aligned}$$

$$\begin{aligned}
\sum_{i=1}^{n} x_i y_i &= 10 \cdot 6 + 12 \cdot 9 + 12 \cdot 8 + 4 \cdot 3 + 12 \cdot 10 + 6 \cdot 4 \\
&\quad + 8 \cdot 5 + 2 \cdot 2 + 18 \cdot 11 + 9 \cdot 9 + 17 \cdot 10 + 2 \cdot 2 \\
&= 917 \\
\sum_{i=1}^{n} x_i^2 &= 10^2 + 12^2 + 12^2 + 4^2 + 12^2 + 6^2 \\
&\quad + 8^2 + 2^2 + 18^2 + 9^2 + 17^2 + 2^2 \;=\; 1'350 \\
\sum_{i=1}^{n} y_i^2 &= 6^2 + 9^2 + 8^2 + 3^2 + 10^2 + 4^2 \\
&\quad + 5^2 + 2^2 + 11^2 + 9^2 + 10^2 + 2^2 \;=\; 641
\end{aligned}$$

Par conséquent, on a

$$n \cdot \mathrm{Cov(x,y)} = \sum_i \mathrm{x_i y_i} - \mathrm{n} \cdot \bar{\mathrm{x}}\bar{\mathrm{y}} = 917 - 12 \cdot \tfrac{28}{3} \cdot \tfrac{79}{12} = \tfrac{539}{3} = 179.\bar{6}$$

$$n \cdot \mathrm{Var(x)} = \sum_i \mathrm{x_i^2} - \mathrm{n} \cdot \bar{\mathrm{x}}^2 = 1'350 - 12 \cdot (\tfrac{28}{3})^2 = \tfrac{914}{3} = 304.\bar{6}$$

$$n \cdot \mathrm{Var(y)} = \sum_i \mathrm{y_i^2} - \mathrm{n} \cdot \bar{\mathrm{y}}^2 = 641 - 12 \cdot (\tfrac{79}{12})^2 = \tfrac{1'451}{12} = 120.91\bar{6}$$

D'où finalement

$$r_{xy} = \frac{n \cdot \mathrm{Cov(x,y)}}{\sqrt{n \cdot \mathrm{Var(x)} \cdot \mathrm{n} \cdot \mathrm{Var(y)}}} = \frac{179.\bar{6}}{\sqrt{304.\bar{6} \cdot 120.91\bar{6}}} \simeq 0.9361$$

Cette valeur est proche de 1, ce qui signifie que la durée de résidence est fortement liée à l'attitude des individus envers la ville. En outre, le fait que r_{xy} soit positif implique une relation positive : plus la durée de résidence est longue, plus l'attitude envers la ville est favorable et inversement, ce que nous pouvons constater au moyen d'une représentation graphique des points (voir plus loin).

d) Puisque $R^2 = r_{xy}^2 \simeq 0.8762$, l'ajustement par une droite de régression sera considéré comme un bon modèle.

e) La droite de régression s'écrit

$$Y = a + bX$$

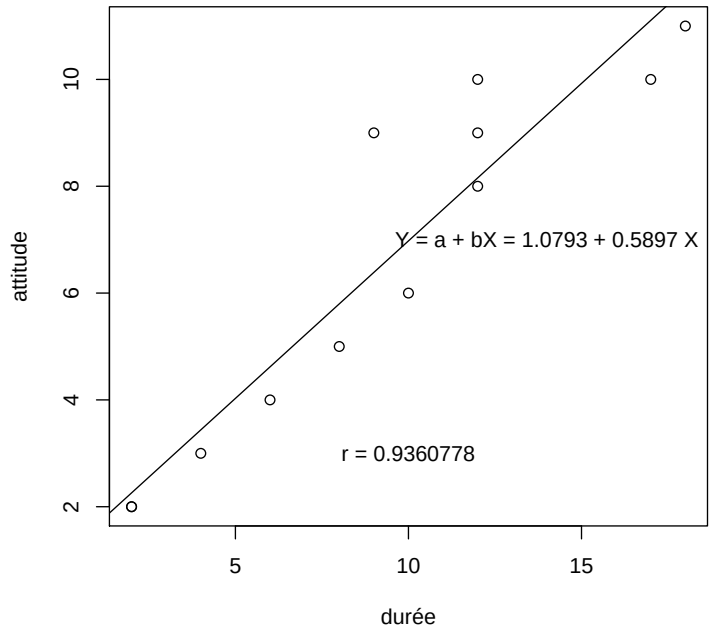
avec

$$b = \frac{n \cdot \mathrm{Cov(x,y)}}{n \cdot \mathrm{Var(x)}} = \frac{179.\bar{6}}{304.\bar{6}} \simeq 0.5897$$

$$a = \bar{y} - b\bar{x} = 6.58\bar{3} - 0.5897 \cdot 9.\bar{3} \simeq 1.0793$$

La meilleure droite ajustant le modèle, au sens des moindres carré, est donc

$$Y = 1.0793 + 0.5897 \cdot X$$




```
duree <- c(10,12,12,4,12,6,8,2,18,9,17,2)
attitude <- c(6,9,8,3,10,4,5,2,11,9,10,2)
plot(duree,attitude, xlab="durée", ylab="attitude")
regression <- lm(attitude~duree)
abline(regression)
cor(attitude,duree)
text(10,3,"r = 0.9360778")
text(14,7, "Y = a + bX = 1.0793 + 0.5897 X")
summary(regression)
cor.test(duree,attitude,method="pearson", alternative="greater")
```

f) On est en train de tester $\begin{array}{ll} H_0 : & \rho \leq 0 \\ H_1 : & \rho > 0 \end{array}$

Avec $r = 0.9361$ et $n = 12$, la statistique du test est

$$t_0 = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.9361}{\sqrt{\frac{1-(0.9361)^2}{10}}} \simeq 8.416$$

g) La p-valeur étant inférieure à n'importe quel choix raisonnable de α, on va rejeter H_0. La relation linéaire entre les deux variables est significative.

Remarque : on a la même conclusion en voyant que dans la table $P(T_{10} > t_\alpha)$ est inférieure à 3.17. On peut donc dire que dans un cas extrême, la région de rejet est $[3.17, \infty[$. Comme $t_0 = 8.416$, on rejette H_0

Problème 2 Attitude envers la ville, le retour

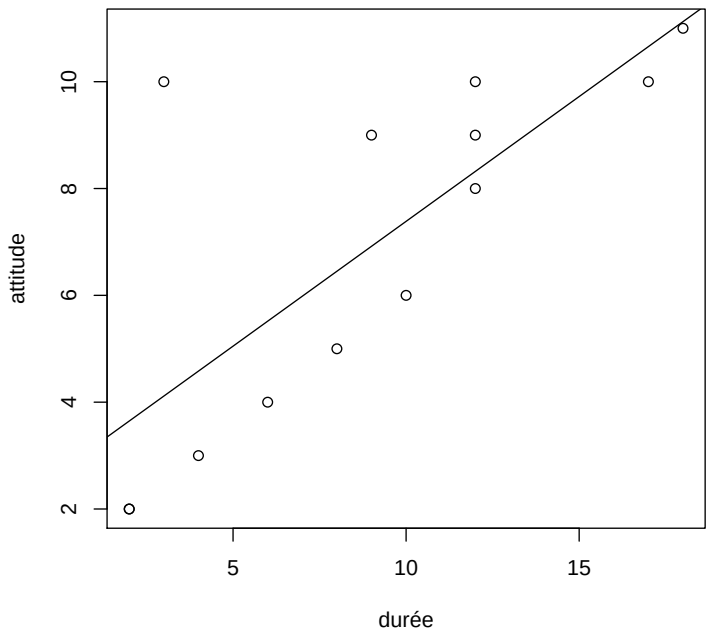
Supposons que dans notre échantillon on ajoute l'opinion d'un treizième individu, qui a passé 3 ans dans la ville et a eu le coup de foudre pour celle-ci : il lui attribue la note 10.

individu	1	2	3	4	5	6	7	8	9	10	11	12	13
attitude	6	9	8	3	10	4	5	2	11	9	10	2	10
durée	10	12	12	4	12	6	8	2	18	9	17	2	3

a) Rajoutez le point dans votre graphique précédent. Quel va être l'impact de cette nouvelle valeur sur la pente de la droite de régression ?

b) On donne $r = 0.7528$. Commenter.

Réponse:




```
duree <- c(duree,3)
attitude <- c(attitude,10)
plot(duree,attitude, xlab="durée", ylab="attitude")
regression <- lm(attitude~duree)
abline(regression)
cor(attitude,duree)
summary(regression)
cor.test(duree,attitude,method="pearson", alternative="greater")
```

Nous voyons que l'introduction de cette nouvelle donnée va faire pencher la droite de son côté de façon non négligeable, et le modèle ne permettra plus d'ajuster les autres valeurs aussi bien qu'avant.

De fait, le calcul du coefficient donne maintenant $r = 0.7528$ et donc un coefficient de détermination égal à $R^2 = 0.5667$, ce qui est mauvais : le modèle ne pourra plus être utilisé pour ajuster les données de façon fiable.