



## TP2 : Analyse bivariée

### Croisement Quantitatif-Quantitatif

*Durée : 2h30*

*L'objectif de ce TP est d'étudier un lien éventuel entre deux variables quantitatives et de construire un modèle prédictif le cas échéant.*

#### Exercice 1

#### *Un peu de géométrie*

Dans une population  $\Omega$  de taille  $n$ , on observe deux variables quantitatives continues,  $x=\{x_k\}_{k=1,\dots,n}$ , et  $y=\{y_k\}_{k=1,\dots,n}$ , de moyennes  $\bar{x}$  et  $\bar{y}$  et de variances  $s_x^2$  et  $s_y^2$ .

On définit le produit scalaire,

$$\langle x, y \rangle = \frac{1}{n} \sum_{k=1}^n x_k y_k .$$

- 1) Exprimez la covariance en fonction du produit scalaire.
- 2) Déterminez la norme du vecteur centré  $x - \bar{x}$  ?
- 3) Comment peut-on écrire la moyenne  $\bar{x}$  à l'aide du produit scalaire ?
- 4) D'un point de vue géométrique à quoi correspond le coefficient de corrélation linéaire ?
- 5) On dit que les deux variables  $x$  et  $y$  sont non corrélées si  $r_{xy}=0$  et entièrement corrélées si  $r_{xy}=\pm 1$ . Qu'est-ce que cela signifie géométriquement ?
- 6) A l'aide du produit scalaire, montrez que
  - a) la moyenne des valeurs prédites est égale à la moyenne de la série observée  $y$ ,
  - b) les résidus sont de moyenne nulle,
  - c) les résidus sont non corrélés avec la série observée  $x$ .

#### Exercice 2

#### *Chômage en 1982*

On donne pour les six premiers mois de l'année 1982 les nombres d'offres d'emploi (concernant des emplois durables à temps plein) et de demandes d'emploi (déposées par des personnes sans emploi, immédiatement disponibles, à la recherche d'un emploi durable à plein temps). Les nombres sont exprimés en milliers.

Offres ( $x_i$ )	61	66,7	75,8	78,6	82,8	87,2
Demandes ( $y_i$ )	2034	2003,8	1964,5	1928,2	1885,3	1867,1

On a les résultats suivants

$$\bar{x} = 75,35 \quad \bar{y} = 1947,15 \quad s_x^2 = 97,49 \quad s_y^2 = 4329,14 \quad c_{xy} = -639,90$$

- 1) Calculer le coefficient de corrélation linéaire. Conclusion
- 2) Déterminer la droite de régression.
- 3) Calculer la prévision de la demande d'emploi s'il y a 61 milliers d'offres. Comparer avec la demande réelle.
- 4) Vérifier la formule de la décomposition de la variance. En déduire le coefficient de détermination.

### Exercice 3

*Données : DepensesEduData.xls*

Le fichier DepensesEduData.csv recense les dépenses publiques de certains états pour l'éducation ainsi que le nombre d'élèves (donnée Eurostat 2008).

- 1) Tracer le nuage de points des dépenses en fonction du nombre d'élèves.

```
tab <- read.table("DepensesEduData.csv",header=T,sep=";",dec=",")
summary(tab)
x <- as.vector(tab$nbEleves)
y <- as.vector(tab$Depenses)
boxplot(x,y)
### nuage de points
plot(x,y,main="Budget en fonction du nombre d'élèves en Europe", xlab="nombre d'étudiants (en
milliers)",ylab="Budget (K€)")
text(x,y,row.names(tab),cex=0.8) # cex=taille de la police
```

- 2) Calculer le coefficient de corrélation linéaire. Conclusion

```
cor(tab) # calcule la corrélation entre les variables
```

- 3) Déterminer la droite de régression. Tracer la droite sur le graphique.

```
RegLin <- lm(y~x) # construit le modèle de régression linéaire / lm = linear model
summary(RegLin) # résume toutes les caractéristiques du modèle
attributes(RegLin) # donne tous les attributs de l'objet « lm »
RegLin$coef # donne les coefficients de la droite
abline(RegLin$coef[1],RegLin$coef[2],col="red",lwd=2) # trace la droite
```

- 4) Vérifier les hypothèses sur les résidus. Quel pays semble atypique par rapport au modèle ?

```

RegLin$fitted      # affiche les prévisions données par le modèle aux points du tableau
RegLin$residuals   # affiche les résidus
restd <- rstandard(RegLin) # affiche les résidus standardisés
X11()              # ouvre une nouvelle fenêtre graphique
plot(RegLin$fitted, restd ,ylim=range(-2,2),main="Résidus standardisés")
                                     # range donne le min et le max d'une série de nombres
abline(h=2,col="red",lwd=2) # ajoute les lignes pour détecter les observations atypiques
abline(h=-2,col="red",lwd=2)
text(RegLin$fitted, restd ,row.names(tab)) # ajoute le nom des pays

```

- 5) Supprimer le pays atypique et refaire la même chose.
- 6) Quel pourcentage de variabilité des dépenses est expliqué par la droite de régression ? Est-ce que vous validez le modèle ?
- 7) Calculer les budgets prédits par le modèle pour 1000, 6000 et 9500 milliers d'étudiants. Placer les sur le graphique.

```

newx <- data.frame(x = c(1000,6500,9000)) # nouveaux points
prev <- predict(RegLin,newdata=newx)      # calcul les prévisions en de nouveaux points
plot(x,y,main="Budget en fonction du nombre d'élèves en Europe", xlab="nombre d'étudiants
      (en milliers)",ylab="Budget (K€)")
points(t(newx),prev,col="green",lwd=2)    # t pour transposer le vecteur newx

```

## Exercice 4

## Ventes

(PY Bernard, exercices corrigés de statistique descriptive, ed. economica)

Une étude a été menée auprès d'entreprises afin d'établir le lien entre les quantités commandés d'un bien, Y, et son prix, X et on obtient les observations suivantes (Commandes.csv).

Prix de vente (€)	Quantités commandées
95	104
130	58
148	42
210	12
250	8
330	5

- 1) Tracer le nuage de points.
- 2) Calculer le coefficient de corrélation linéaire entre X et Y. Conclusion
- 3) Déterminer la droite de régression de Y en fonction de X.

- 4) Quel est le pourcentage de variation des quantités de commande expliquée par la droite de régression ?
- 5) Calculer les résidus et vérifier les hypothèses sur les résidus. Conclusion.
- 6) On pose  $u=\log(x)$  et  $v=\log(y)$ . Quelle est la relation entre  $u$  et  $v$  ?
- 7) Calculer le coefficient de corrélation linéaire entre  $u$  et  $v$ .
- 8) Trouver la droite de régression de  $v$  sur  $u$ .
- 9) Quel est le pourcentage de variation des quantités de commande expliquée par la droite de régression ?
- 10) Valider le modèle.
- 11) En déduire la quantité qui serait commandée si le prix était fixé à 75€.

**Exercice 1***(Facultatif - Suite)*

- 1) Montrer que les résidus sont non corrélés avec la série X. Qu'est-ce que cela signifie ?
- 2) Montrer la formule de décomposition de la variance

$$s_Y^2 = s_E^2 + s_R^2$$

où  $s_E^2$  est la *variance expliquée* par la droite de régression, et  $s_R^2$  est la *variance résiduelle*.

On peut alors montrer que le *coefficient de détermination*

$$R^2 = \frac{s_E^2}{s_Y^2},$$

qui donne le taux de variance expliquée par la droite de régression, est égale au coefficient de corrélation linéaire au carré,  $R^2 = r_{xy}^2$ .