



UE : Outils d'analyse , EC : Analyse de données, Semestre 1

Travaux Dirigés (Régression linéaire)

Responsable: Dr. Mor NDONGO

Niveau: Licence 3 Informatique

Exercice 1

Commenter les nuages de points de la figure suivante. Les variables vous semblent-elles liées ? Sous quelle forme ?



Exercice 2

Chaque semaine de l'année comptant six jours ouvrables, on a relevé la recette en millions de centimes d'un supermarché le lundi et le samedi.

Un échantillon de dix semaines a donné les résultats suivants :

Semaine	1	2	3	4	5	6	7	8	9	10
Recette du lundi (X)	57	60	52	49	56	46	51	63	49	57
Recette du samedi (Y)	86	93	77	67	81	70	71	91	67	82

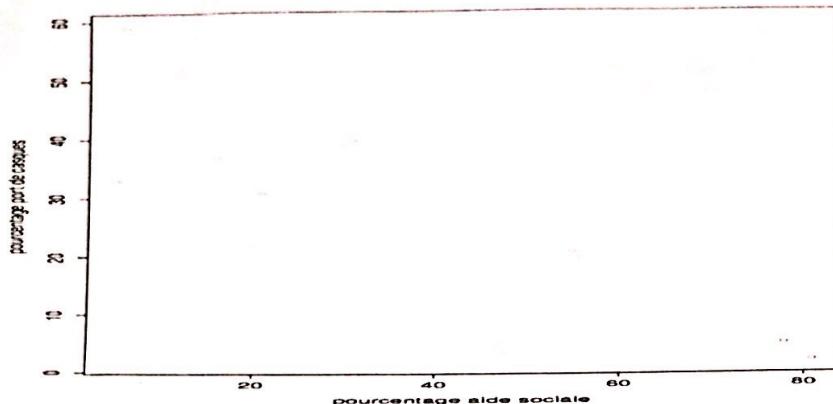
On donne :

$$\sum_{i=1}^{10} x_i = 540 \quad \sum_{i=1}^{10} x_i^2 = 29462 \quad \sum_{i=1}^{10} y_i = 785 \quad \sum_{i=1}^{10} y_i^2 = 62459 \quad \sum_{i=1}^{10} x_i y_i = 42836$$

- 1) Calculer les moyennes et écarts type de X et Y
- 2) Donner la valeur d'un paramètre qui permet de mesurer l'intensité de la relation linéaire entre X et Y. Interpréter le résultat.
- 3) Déterminer l'équation de la droite de régression des recettes du samedi en fonction des recettes du lundi.
- 4) Quelle peut être la recette du samedi d'une semaine où la recette du lundi est de 55 ?

Exercice 3

Le graphique ci-dessous représente la relation entre le pourcentage de familles recevant une aide sociale dans un quartier et le pourcentage de cyclistes portant des casques dans le même quartier. Les données ont été collectées pour les 12 quartiers de la ville d'Elithan.



Le pourcentage moyen de famille recevant une aide sociale à Elithan est de $m_x = 30,8\%$ avec un écart type de $s_x = 26,7\%$. Le pourcentage moyen de port de casque à Elithan est de $m_y = 38,8\%$ avec un écart type de $s_y = 16,9\%$. On postule une relation linéaire entre fraction en difficultés sociales et fraction portant le casque à vélo.

- 1) Que peut-on dire de l'augmentation du pourcentage de l'aide sociale par rapport au pourcentage de port de casques ?
- 2) Si le R^2 de la régression linéaire par la méthode des moindres carrés est de 72 %, quelle est le coefficient de corrélation r entre aide sociale et port de casques ? (On rappelle que $r^2 = R^2$). Interpréter ce coefficient de corrélation.
- 3) Ecrire le modèle linéaire correspondant.
- 4) Calculez la pente et l'intercept de la droite de régression par les moindres carrés.
On vous rappelle qu'en régression linéaire simple $r = \frac{s_x}{s_y} b$ et $a = m_y - b \times m_x$ (avec b est la pente de la droite de régression de y en x et a est l'intercept).
- 5) Quelle interprétation donnez-vous à la valeur de l'intercept ?
- 6) Quelle interprétation donnez-vous à la valeur de la pente ?
- 7) Un quartier d'Elithan n'avait pas été inclus dans l'étude pour des raisons organisationnelles. Dans ce quartier 40 % des familles reçoivent une aide sociale et 40 % des cyclistes portent le casque. Calculez la valeur du résidu par rapport à la régression. Comment peut-on interpréter ce résidu ?



UE : Ingénierie du calcul , EC : Analyse de données, Semestre 5

Travaux Pratique (Régression linéaire avec le logiciel R)

Responsable: Dr. Mor NDONGO

Niveau: Licence 3 Informatique

Exercice 1 *

Pour une station de ski donnée, on s'intéresse aux nombre d'accidents observés sur les pistes Y en fonction du nombre de skieurs X au cours des années a_1 à a_{10} . On a les observations suivantes :

Années	x_i	y_i
a_1	3700	150
a_2	4100	180
a_3	4300	200
a_4	4500	200
a_5	4900	210
a_6	5200	220
a_7	5700	240
a_8	6400	250
a_9	6800	270
a_{10}	7400	280

1. Saisir les données dans R sous forme d'un data.frame nommé `ski`.
2. Tracer dans R le nuage de points correspondant aux données.
3. Que peut-on dire sur l'augmentation du nombre d'accidents par rapport au nombre de skieurs ?
4. En utilisant le data.frame `ski`, calculer $\sum_{i=1}^{10} x_i$ $\sum_{i=1}^{10} y_i$ $\sum_{i=1}^{10} x_i y_i$ $\sum_{i=1}^{10} y_i^2$ et $\sum_{i=1}^{10} x_i^2$.
5. Grâce aux valeurs précédentes, calculer dans R les estimations de la pente et de l'intercept du modèle (arrondir à 4 chiffres après la virgule). On les note b et a , respectivement. On rappelle que :
$$b = \frac{\text{cov}(X, Y)}{\text{Var}(X)}, \quad a = m_X - b m_Y$$
6. On veut retrouver ces résultats avec la commande `lm` de R. Quelles lignes de code doit-on écrire pour vérifier les résultats calculés à la question 5) ?
7. A l'aide de l'objet renvoyé par la fonction `lm` de R, conclure sur la significativité des paramètres (pente et intercept) et du modèle global.
8. Que peut on dire sur l'adéquation du modèle ?
9. Tracer la droite de régression sur le nuage de points.
10. Au cours de la onzième année, on observe 270 accidents pour 8000 skieurs. Calculer dans R l'intervalle de prédiction au niveau de confiance 95 % pour $x_{11} = 8000$?

Exercice 2

Le tableau ci-dessous donne le produit national brut et la consommation privée pour les années 1960 à 19698 en France (exprimés en francs constants de 1963).

Années	1960	1961	1962	1963	1964	1965	1966	1967	1968
PNB (x)	346	365	392	412	439	460	486	508	533
Conso. privée (y)	209	222	238	255	269	281	294	309	326

1. Saisir les données dans R sous forme d'un data.frame nommé **pnbdata**.
2. Tracer dans R le nuage de points correspondant aux données.
3. Que peut-on dire sur l'augmentation de la consommation privée par rapport au produit national bruit ?
4. En utilisant le data.frame **pnbdata**, calculer $\sum_{i=1}^{10} x_i$ $\sum_{i=1}^{10} y_i$ $\sum_{i=1}^{10} x_i y_i$ $\sum_{i=1}^{10} y_i^2$ et $\sum_{i=1}^{10} x_i^2$.
5. Grâce aux valeurs précédentes, calculer dans R les estimations de la pente et de l'intercept du modèle (arrondir à 4 chiffres après la virgule). On les note b et a , respectivement. On rappelle que :
$$b = \frac{\text{cov}(X, Y)}{\text{Var}(X)}, \quad a = m_X - b m_Y$$
6. On veut retrouver ces résultats avec la commande **lm** de R. Quelles lignes de code doit-on écrire pour vérifier les résultats calculés à la question 5) ?
7. A l'aide de l'objet renvoyé par la fonction **lm** de R, conclure sur la significativité des paramètres (pente et intercept) et du modèle global.
8. Que peut on dire sur l'adéquation du modèle ?
9. Tracer la droite de régression sur le nuage de points.
10. Au cours de l'année 1969, on observe une consommation privée de 350 pour un PNB de 575. Calculer dans R l'intervalle de prédiction au niveau de confiance 95 % pour $x_{10} = 575$?



UE : Outils d'analyse , EC : Analyse de données, Semestre 1

Examen première session (Durée : 2 heures)

Responsable: Dr. Mor NDONGO

Niveau: Licence 3 Informatique

Exercice 1 (5 points)

Expliquez brièvement les méthodes statistiques pour étudier la liaison entre :

1. Deux variables quantitatives
2. Deux variables qualitatives
3. une variable quantitative et une variable qualitative

Exercice 2 (6 points)

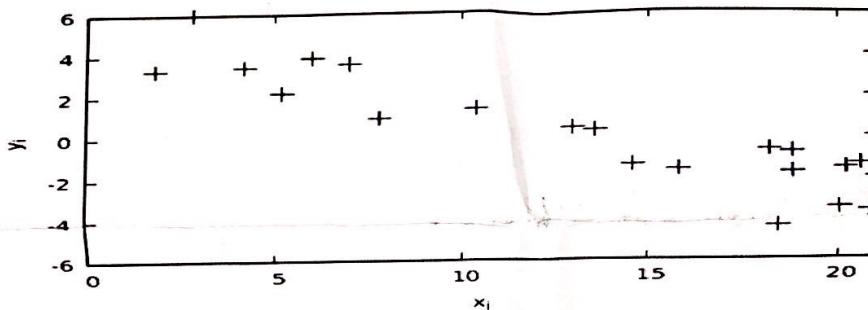


Figura 1: Représentation d'une série $(x_i; y_i)$ en nuage de points.

Question 1 (QCM) : On considère le nuage de points de la figure 1; on analyse ces données par la méthode des moindres carrés : on calcule b_1 , estimation du coefficient directeur de la droite de régression, et r^2 , le coefficient de détermination. Quel est le résultat correct?

NB : il n'y a pas besoin de faire de calcul pour répondre à cette question.

- A) $b_1 = 0,39$ et $r^2 = 0,83$ B) $b_1 = 0,39$ et $r^2 = 0,17$ C) $b_1 = 2,90$ et $r^2 = 0,83$
 D) $b_1 = -0,39$ et $r^2 = 0,17$ E) $b_1 = -0,39$ et $r^2 = 0,83$ F) $b_1 = -2,90$ et $r^2 = 0,83$

Question 2 (QCM) : On considère les données suivantes

x_i	15	10	14	14	23	14	13
y_i	-136	-114	-155	-132	-237	-121	-143

On suppose que les hypothèses du modèle de régression linéaire sont vérifiées. On note $y = b_0 + b_1x$ l'équation de la droite de régression estimée.

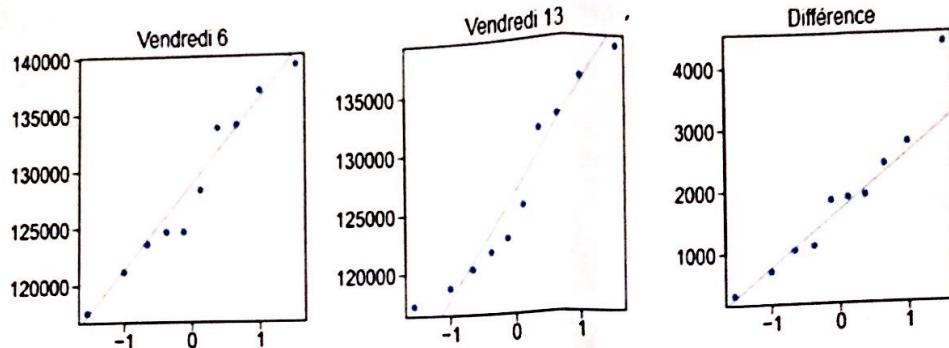
Parmi les valeurs suivantes quelle est la plus proche de la valeur de b_1 ?

- A) $b_1 = -9,721$ B) $b_1 = 5,537$ C) $b_1 = 9,721$ D) $b_1 = -5,537$ E) $b_1 = 8,721$

Question 3 (QCM) : On utilise pour cette question, les données de la question précédente. Parmi les valeurs suivantes, quelle est la plus proche du coefficient de détermination r^2 ?

- A) $r^2 = 0,910$ B) $r^2 = -0,881$ C) $r^2 = 0,777$ D) $r^2 = -0,939$ E) $r^2 = 0,881$

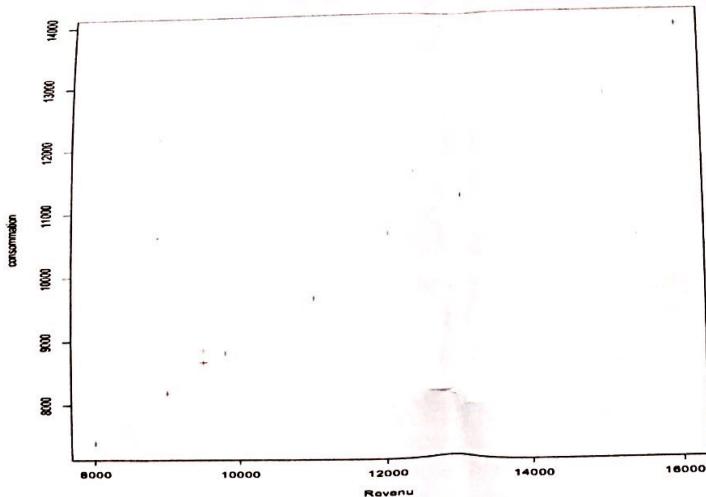
Question 4 : Une étude a collecté des données de trafic routier et accidents de la route ayant lieu les vendredis 13 et les vendredis précédents, vendredi 6. Les graphiques ci-dessous traduisent les trafics routiers pour 10 dates de vendredi 13 et 10 dates de vendredi 6 (nombre de véhicules ayant traversé une certaine intersection).



Comment appelle-t-on ce type de graphique ? Quelle information peut-on tirer par rapport à la distribution de l'intensité du trafic routier en général ?

Exercice 3 (9 points)

Pour un pays donné, on s'intéresse à la consommation des ménages en euros (variable Y) en fonction du revenu disponible (variable X) sur la période 1992-2001. Le nuage de point correspondant aux données est représenté sur la figure suivante :



- 1) Que peut-on dire sur l'augmentation de la consommation des ménages par rapport au revenu disponible ?
- 2) Quel test statistique permet de vérifier la question 1) ? Formuler les hypothèses testées.
- 3) Quelles lignes de commande sous R permettent de réaliser ce test ?
- 4) Avec les lignes de commande écrites en 3), on obtient les résultats suivants :

```
t = 43.535, df = 8, p-value = 8.549e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.9907763 0.9995215
sample estimates:
cor
0.9978962
```

A l'aide des résultats obtenus :

- 4-a) Donner une estimation ponctuelle et une estimation par intervalle de confiance à 95% du coefficient de corrélation. Interpréter le coefficient de corrélation estimé.
- 4-b) peut on accepter que les deux variables X et Y sont linéairement liées ? **Justifier votre réponse.**
- 5) Ecrire le modèle linéaire correspondant.
- 6) Donner les hypothèses de ce modèle.
- 7) On veut estimer les paramètres de ce modèle avec la commande **lm** de R. Quelles lignes de commande doit on écrire ?
- 8) Avec les lignes de commande écrites en 7), on obtient les résultats suivants :

```
Residuals:
    Min      1Q Median      3Q     Max
-150.69 -109.97 -34.62   54.45  236.28

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.176e+03 2.074e+02  5.671 0.00047 ***
X           7.810e-01 1.794e-02 43.535 8.55e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 143.7 on 8 degrees of freedom
Multiple R-squared:  0.9958, Adjusted R-squared:  0.9953
F-statistic: 1895 on 1 and 8 DF,  p-value: 8.549e-11
```

- 8-a) Donner les estimations de l'intercept et de la pente. **Interpréter**
- 8-b) Après avoir écrit les hypothèses du test, tester la significativité de la pente à l'aide des résultats obtenus par la commande **lm**.
- 8-c) Tester la significativité globale du modèle.
- 8-d) Que peut on conclure sur la qualité de l'ajustement du modèle? **justifiez**
- 9) En 2002 et 2003, on prévoit respectivement 16800 euros et 17000 euros pour la valeur du revenu. Déterminer les valeurs prévues de la consommation pour ces deux années, ainsi que les valeurs des résidus pour ces deux années. Comment peut-on interpréter ces résidus ?

UE : Ingénierie du calcul, EC : Analyse de données, Semestre 1

Examen première session (Durée : 2 heures)

Responsable: Dr. Mor NDONGO

Niveau: L3 Informatique

Documents interdits

LES TELEPHONES PORTABLES DOIVENT ÊTRE ETEINTS ET RANGÉS.

Exercice 1 (8 points)

Question 1 (QCM) : On considère les données suivantes

x_i	15	10	14	14	23	14	13
y_i	-136	-114	-155	-132	-237	-121	-143

On suppose que les hypothèses du modèle de régression linéaire sont vérifiées. On note $y = b_0 + b_1x$ l'équation de la droite de régression estimée.

Parmi les valeurs suivantes quelle est la plus proche de la valeur de b_1 ?

- A) $b_1 = -9,721$ B) $b_1 = 5,537$ C) $b_1 = 9,721$ D) $b_1 = -5,537$ E) $b_1 = 8,721$

Question 2 (QCM) : On utilise pour cette question, les données de la question précédente. Parmi les valeurs suivantes, quelle est la plus proche du coefficient de détermination r^2 ?

- A) $r^2 = 0,910$ B) $r^2 = -0,881$ C) $r^2 = 0,777$ D) $r^2 = -0,939$ E) $r^2 = 0,881$

Question 3 (QCM) : La sortie d'une régression multiple du type $Y = a_0 + a_1X_1 + a_2X_2$ donne les résultats suivants :

- $R^2 = 0,81$ $(Pr > F) = 0,049$ $(Pr > |t|) = 0,050$ pour a_0 $(Pr > |t|) = 0,005$ pour a_1
- $(Pr > |t|) = 0,19$ pour a_2

Que concluez vous? Veuillez choisir une réponse.

- a) Le modèle $Y = a_0 + a_1X_1$ est significatif et explicatif
- b) Le modèle $Y = a_0 + a_2X_2$ est non significatif et explicatif
- c) Le modèle $Y = a_0 + a_1X_1 + a_2X_2$ est significatif et explicatif

Question 4 (QCM) : La sortie d'une régression simple donne les résultats suivants :

- $R^2 = 0,77$ $(Pr > F) = 0,50$ $(Pr > |t|) = 0,045$ pour a_0 $(Pr > |t|) = 0,005$ pour a_1

Que concluez vous? Veuillez choisir une réponse.

- a) Le modèle n'est ni explicatif ni significatif
- b) Le modèle est significatif mais non explicatif
- c) Le modèle est explicatif mais aucun des paramètres n'est significatif
- d) Le modèle est explicatif mais non significatif

L3. Informatique - Examen première session

Question 5 (QCM) : Vous voulez savoir si l'indice de satisfaction des clients d'une station de ski des Alpes dépend de leur origine nationale, vous préconisez
Veuillez choisir une réponse.

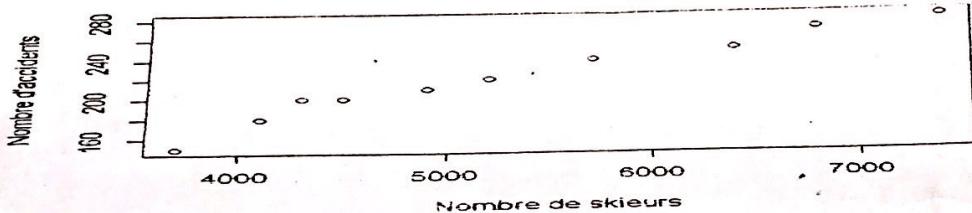
- a) Analyse de la variance b) Régression logistique
 c) test du khi-deux d'ajustement d) test du khi-deux de contingence

Question 6 (QCM) : Pour répondre à la question suivante quelle procédure statistique mettez vous en oeuvre : La rentabilité d'un actif exprimée en % dépend-elle de la durée de placement en jours et du risque exprimé par la variance du cours observée pendant la durée du placement ?
Veuillez choisir une réponse.

- a) Analyse de la variance b) Régression simple
 c) Régression multiple d) Analyse de la covariance

Exercice 2 (12 points)

Pour une station de ski donnée, on s'intéresse aux nombres d'accidents observés sur les pistes Y en fonction du nombre de skieurs X au cours des années a_1 à a_{10} . Le nuage de point correspondant aux données est représenté sur la figure suivante :



- 1) Que peut-on dire sur l'augmentation du nombre d'accidents par rapport au nombre de skieurs ?
- 2) Quel test statistique permet de vérifier la question 1) ? Formuler les hypothèses testées.
- 3) Quelles lignes de commande sous R permettent de réaliser ce test ?
- 4) Avec les lignes de commande écrites en 3), on obtient les résultats suivants :

```
t = 12.848, df = 8, p-value = 1.272e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9010409 0.9946367
sample estimates:
cor
0.9766149
```

A l'aide des résultats obtenus :

- 4-a) Donner une estimation ponctuelle et une estimation par intervalle de confiance à 95% du coefficient de corrélation. Interpréter le coefficient de corrélation estimé.
- 4-b) peut-on accepter que les deux variables X et Y sont linéairement liées ? Justifier votre réponse.
- 5) Ecrire le modèle linéaire correspondant.