RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE UNIVERSITÉ BADJI MOKHTAR ANNABA

FACULTÉ DES SCIENCES
DÉPARTEMENT DES MATHÉMATIQUES

BIOSTATISTIQUE : RÉGRESSION LINÉAIRE SIMPLE AVEC R

Présentée par : Dr. Hafida Goual
Maître de conférence classe B en mathématique

Polycopié de cours destiné aux étudiants de Master 1 LMD (M1) des spécialités Biochimie et Biotechnologie végétale

Année universitaire : 2018/2019

Chapitre 1

Régression linéaire simple

La régression linéaire simple est une méthode statistique qui permet de résumer et d'étudier les relations entre deux variables continues (quantitatives). Cette section présente le concept et les procédures de base de la régression linéaire simple. Nous allons également apprendre deux mesures décrivant la force de l'association linéaire trouvée entre les données.

Une variable, notée x, est considérée comme la variable prédictive, explicative ou indépendante. L'autre variable, notée y, est considérée comme la réponse, le résultat ou la variable dépendante. Comme les autres termes sont utilisés moins fréquemment aujourd'hui, nous utiliserons les termes "prédicteur" et "réponse" pour faire référence aux variables rencontrées dans ce cours. Les autres termes mentionnés, vous les rencontrez dans d'autres domaines. La régression linéaire simple prend son adjectif "simple", car elle ne concerne que l'étude d'une seule variable prédictive.

1.1 Définition d'un modèle de régression linéaire simple

Le but de la régression linéaire est de modéliser une variable continue Y en tant que fonction mathématique d'une ou de plusieurs variables X , de sorte que nous puissions utiliser ce modèle de régression pour prédire le Y lorsque seul X est connu. Cette équation mathématique peut être généralisée comme suit :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1, ..., n.$$

avec

Y_i : la variable endogène (dépendante, à expliquer) à l'indice i.

X_i : la variable exogène (indépendante, explicative) à l'indice i.

β_0, β_1 : les paramètres inconnus du modèles où
β_0 : représente le point x de la droite de régression avec l'ordonnée à l'origine.
β_1 : est la pente de la droite de régression.
ε_i : l'erreur aléatoire du modèle.

n : nombre d'observation.

1.2 Hypothèses du modèle

Le modèle repose sur les hypothèses suivantes

1. $E(\varepsilon_i) = 0$, l'erreur est centrée.
2. $E(\varepsilon_i^2) = \sigma_\varepsilon^2$, la variance de l'erreur est constante.
3. $Cov(\varepsilon_i, \varepsilon_{i'}) = 0, i \neq i^{'}$ les erreurs ne sont pas autocorrélées.
4. $Cov(x_i, \varepsilon_{i'}) = 0$, l'erreur n'est pas corrélée avec la variable exogène.
5. La variable exogène X_i n'est pas aléatoire.
6. Le modèle est linéaire en X par rapport aux paramètres.

Avant de commencer à construire le modèle de régression, il est recommandé d'analyser et de comprendre les variables. L'analyse graphique et l'étude de corrélation ci-dessous aideront à résoudre ce problème.

1.3 Représentation graphique

Le but de cette étude est de construire un modèle de régression simple que nous pouvons utiliser pour prédire la mortalité (Mort) en établissant une relation linéaire statistiquement significative avec la latitude (Lat). Mais avant de passer à la syntaxe, essayons de comprendre ces variables graphiquement. En règle générale, pour chacune des variables indépendantes (variables prédites), les tracés suivants sont dessinés afin de visualiser le comportement suivant :

1. **Nuage de points :** Visualisez la relation linéaire entre le prédicteur et la réponse.

2. **Boîte à moustaches :** Pour repérer les observations aberrantes de la variable. Avoir des valeurs aberrantes de votre prédicteur peut affecter considérablement les prévisions, car elles peuvent également influer sur la direction / la pente de la droite du meilleur ajustement.
3. **Diagramme de densité :** pour voir la distribution de la variable prédictive. Idéalement, une distribution proche de la normale (une courbe en forme de cloche), sans être oblique vers la gauche ou la droite, est préférable. Voyons comment faire chacune d'elles.

1.3.1 Nuage de points

Les diagrammes de dispersion peuvent aider à visualiser toute relation linéaire entre la variable dépendante (réponse) et les variables indépendantes (prédicteurs). Idéalement, si vous avez plusieurs variables explicatives, un diagramme de dispersion est tracé pour chacune d'elles en fonction de la réponse, ainsi que la droite des meilleurs ajustement, comme indiqué ci-dessous.

Exemple

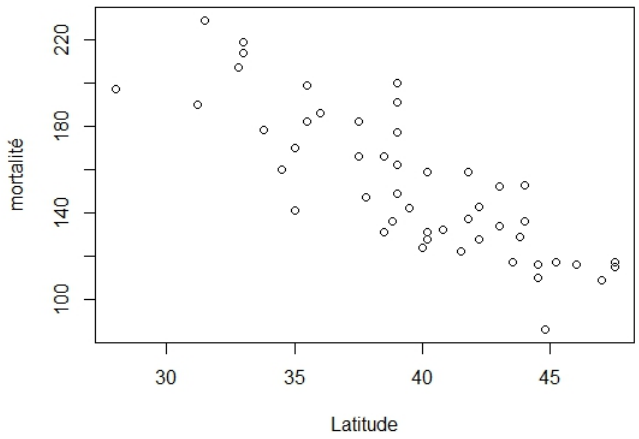

Fig. 1. Mortalité par cancer de la peau par rapport à la latitude de l'État

Le diagramme de dispersion ci-dessus suggère une relation à croissance linéaire entre les variables "Lat" et "Mort". C'est une bonne chose car l'une des hypothèses sous-jacentes

à la régression linéaire est que la relation entre la réponse et les variables prédictives est linéaire et additive.

1.3.2 Boîte à moustache

Le but de traçage d'une Boîte à moustache est de vérifier les valeurs aberrantes.

En règle générale, tout point de données situé en dehors de l'intervalle interquartile est considéré comme une valeur aberrante, où IQR est calculé comme étant la distance entre les valeurs du 25ème centile et du 75ème centile de cette variable.

Exemple

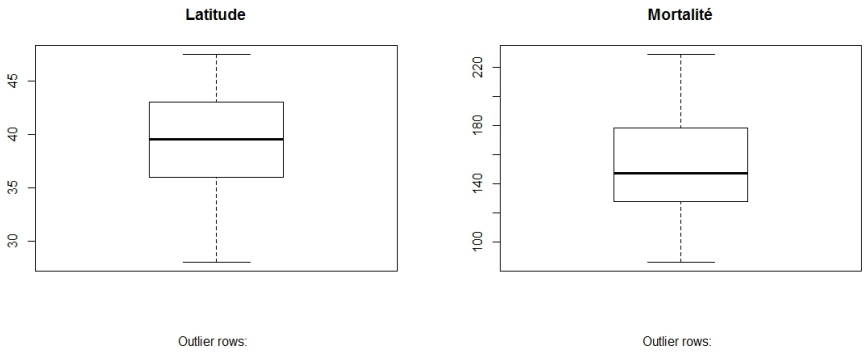

Fig. 2. Boites à moustache de Latitude et Mortalité

1.3.3 Diagramme de densité

Le but est de vérifier si la variable de réponse est proche de la normalité.

Exemple

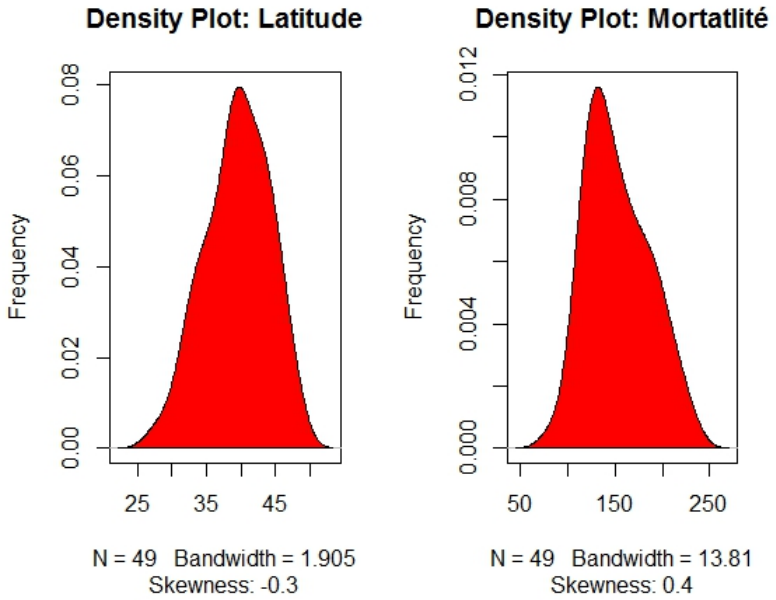

Fig. 3. Diagrammes de densité de latitude et mortalité

1.4 Corrélation

La corrélation est une mesure statistique qui suggère le niveau de dépendance linéaire entre deux variables, qui se produisent par paire - tout comme ce que nous avons ici en mortalité et en latitude. La corrélation peut prendre des valeurs comprises entre -1 et $+1$.

Si nous observons pour chaque cas où la mortalité augmente, la latitude augmente également avec elle, alors il existe une forte corrélation positive entre eux et donc la corrélation entre eux sera plus proche de 1. L'inverse est vrai pour une relation inverse, dans laquelle dans ce cas, la corrélation entre les variables sera proche de -1.

Une valeur plus proche de 0 suggère une faible relation entre les variables. Une faible corrélation suggère probablement qu'une grande partie de la variation la variable réponse (Y) n'est pas expliquée par le prédicteur (X). Dans ce cas, nous devrions probablement rechercher de meilleures variables explicatives.

Exemple avec R

cor(skincancer$Lat, skincancer$Mort) # calculer la corrélation entre latitude et la mortalité

[1] -0.8245178

1.5 Estimation des paramètres par la méthode des moindres carrés ordinaires (M.C.O)

Dans cette section, nous allons présenter la méthode des moindres carrés pour trouver une droite du meilleur ajustement, elle décrit l'association linéaire entre une variable réponse Y et un prédicteur X, basée sur un échantillon aléatoire de taille n. Cette méthode consiste à ajuster le nuage de points à l'aide d'une droite en minimisant la distance au carré entre chaque valeur observée et la droite d'estimation, cette distance mesure le résidu $\widehat{e_i} = \varepsilon_i = y_i - \widehat{y_i}$.

L'estimation des paramètres β_0, β_1 du modèle de régression linéaire simple est obtenue en minimisant la somme des carrés des erreurs

$$\min \sum_{i=1}^{n} \varepsilon_i^2 = \min \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = \min \sum_{i=1}^{n} O^2$$

Pour que cette fonction ait un minimum, il faut que les dérivées par rapport à β_0 et β_1 soient nuls.

Ainsi, ces dérivées sont données comme suit

$$\frac{\partial O}{\partial \beta_0} = 0 \Longleftrightarrow 2\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)(-1) = 0 \Longrightarrow \sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i \qquad (1)$$

$$\frac{\partial O}{\partial \beta_1} = 0 \Longleftrightarrow 2\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0 \Longrightarrow \sum_{i=1}^{n} Y_i X_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2 \qquad (2)$$

On notent $\widehat{\beta_0}$ et $\widehat{\beta_1}$ les solutions des équations (1) et (2) respectivement.
D'après l'équation (1), on obtient

$$\widehat{\beta_0} = \frac{\sum_{i=1}^{n} Y_i}{n} - \widehat{\beta_1} \frac{\sum_{i=1}^{n} X_i}{n}$$

ou bien

$\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{X}$ puisque $\left[\frac{\sum_{i=1}^{n} Y_i}{n} = \overline{Y}\right]$ et $\left[\frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}\right]$.

En remplaçant la valeur de $\widehat{\beta_0}$ dans l'équation (2), on obtient

$$\sum_{i=1}^{n} Y_i X_i - \overline{Y} \sum_{i=1}^{n} X_i = \widehat{\beta}_1 \left(\sum_{i=1}^{n} X_i^2 - \overline{X} \sum_{i=1}^{n} X_i \right)$$

d'où

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \overline{Y} \sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2 - \overline{X} \sum_{i=1}^{n} X_i} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{YX}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} = \frac{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$$

On peut maintenant écrire le modèle théorique (modèle non ajusté) comme suit

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

La forme d'un modèle estimé (ajusté) est alors

$$Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + e_i$$

avec

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

et

$$e_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$$

Où e_i est le résidu du modèle.