Statistique Numérique et Analyse de Données
Ecole des Ponts ParisTech, 2ème année

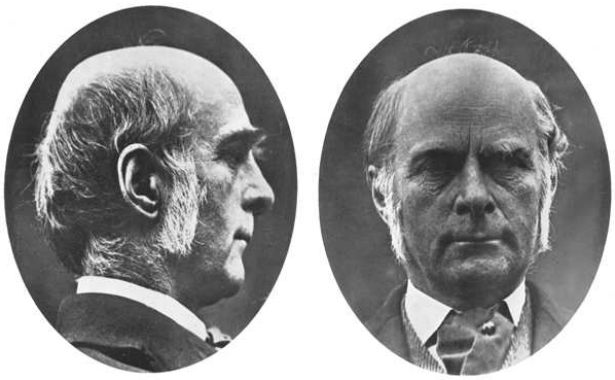
TP4 : RÉGRESSION LINÉAIRE

Note historique

Le terme ***regression*** a été introduit par Francis Galton, chercheur britannique du 19e siècle et cousin de Charles Darwin, dans le célèbre article :

> Regression towards mediocrity in hereditary stature
> *Journal of the Anthropological Institute* 15 : 246-63 (1886)
> `galton.org/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf`

pour décrire un phénomène biologique. Le phénomène est que la taille des enfants nés des parents inhabituellement grands (ou petits) se rapproche de la taille moyenne de la population.



Sir Francis Galton 1822-1911

1. Régression linéaire simple : les données de Galton

On se propose maintenant d'utiliser le modèle de régression simple pour analyser les données des tailles utilisées par Galton.

1. Charger les données :
   ```
   library(UsingR)
   data(galton)
   attach(galton)
   ```
2. Afficher les histogrammes des variables `parent` et `child` pour avoir une idée de la façon dont elles sont réparties.
3. Déterminer les moyennes et les écart-types des variables `parent` et `child`.

La Commande `lm`

La commande `lm` permet d'effectuer une régression linéaire multiple. La syntaxe générale est

`fit=lm(` *formule , jeu de données , options* `)`

- Les arguments
  - L'argument `formule` est de forme `VAR1` $\sim$ `VAR2 + VAR3 + VAR4` où `VAR1` désigne la variable réponse, `VAR2 - VAR4` désignent les variables explicatives (il peut y en avoir autant qu'on veut). Il faut noter que la variable explicative constante est inclue par défaut dans la régression. Si l'on souhaite qu'elle soit exclue, il faut saisir `VAR1` $\sim$ `0 + VAR2 + VAR3 + VAR4`.
  - L'argument `jeu de données` est optionnel ; il sers à spécifier le jeu de données dans lequel se trouvent les variables de la régression.
  - L'argument `options` n'est utilisée que pour une analyse très avancée.
- Le résultat `fit` est un objet (de classe `lm`) ayant pour attributs principaux :
  - `coefficients` les valeurs estimées des coefficients, $\hat{\beta}_j$,
  - `fitted` les valeurs ajustées, $\hat{y}_i$
  - `residuals` les résidus, $\hat{\varepsilon}_i$,
  - `df` le degré de liberté, $d = n - p - 1$ (et $d = n - p$ si la variable explicative constante a été exclue).
- Les fonctions `plot`, `summary` mais aussi `anova` peuvent prendre comme argument un objet de classe `lm`.

4. Effectuer une régression linéaire :

```
LinReg=lm(child ~ parent)
plot(parent,child,bg="red")
abline(LinReg, lwd=3, col="blue")
summary(LinReg)
```

   En déduire les estimateurs des valeurs de β_0 et de β_1 tels que

$$\texttt{child} = \beta_0 + \beta_1 \cdot \texttt{parent} + \varepsilon. \tag{1}$$

   Quelle est la valeur estimée de la variance des erreurs ε ?

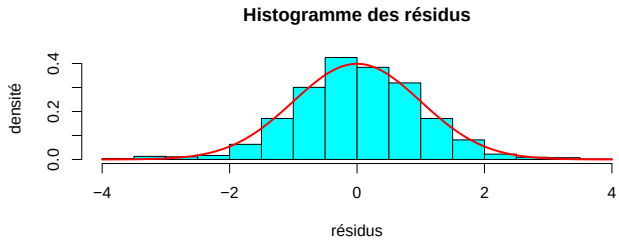
5. La valeur estimée de β_1 confirme-t-elle la loi héréditaire proposée et défendue par Galton ?

6. Au vu de la valeur du coefficient de détermination $\texttt{R}^2$, discuter de la qualité prédictive du modèle linéaire (1).

7. La théorie de Galton a été étudiée de façon plus détaillée par Karl Pearson (1857 – 1936), l'un des fondateurs de la statistique mathématique. Il a fait des statistiques

sur un échantillon plus grand. Ces données se trouvent dans le fichier `father.son` :

```
library(UsingR)
data(father.son)
names(father.son)
par(bg="cornsilk",pch=21)
plot(father.son,bg="red")
```

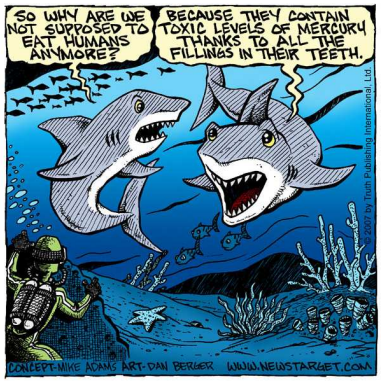


Tracer l'histogramme des résidus standardisés et le superposer avec la courbe de la densité de la loi gaussienne centrée réduite. Est-il raisonnable d'affirmer que les résidus suivent la loi $\mathcal{N}(0,1)$?

2. La tension de vapeur du mercure

On cherche à déterminer une formule mathématique permettant le calcul de la tension de vapeur du mercure, mesurée en mm de mercure, comme une fonction de la température (mesurée en degré Celsius).

La tension de vapeur, également appelée pression de vapeur saturante, est la pression à laquelle la phase gazeuse de cette substance (le mercure) est en équilibre avec sa phase liquide ou solide. En d'autres termes, la tension de vapeur est la pression à laquelle un fluide passe de l'état gazeux à l'état liquide (ou de l'état liquide à gazeux) pour une température donnée. La tension de vapeur dépend de la température.



1. Charger les données et afficher les :

   ```
   library(datasets)
   data(pressure)
   pp=pressure$pressure
   pt=pressure$temperature
   par(bg="cornsilk",pch=21)
   plot(pt,pp,bg="red")
   ```

   Quelle fonction élémentaire pourrait fournir un bon ajustement à ce nuage des points ?

2. Effectuer une régression linéaire simple pour trouver les valeurs estimées de β_0 et β_1 tels que

   $$\log(\texttt{tension de vapeur}) = \beta_0 + \beta_1 \cdot \texttt{temperature} + \varepsilon.$$

   Le modèle obtenu est-il fiable ?

3. En utilisant le modèle ci-dessus, donner une prédiction de la tension de vapeur du mercure correspondant à la température $T = 90; 230; 400$.

```
new=data.frame(pt=c(90,230,400))
Pred=predict(lm(log(pp) ~ pt),new,interval="confidence")
Pred=exp(Pred)
```

4. **Formule de Dupré :** les physiciens affirment qu'une meilleure façon de calculer la tension de vapeur en fonction de la température serait d'utiliser la formule

$$P = \alpha_1 T^{\alpha_2} e^{\alpha_3 / T} \tag{2}$$

où P désigne la tension de vapeur, T désigne la température et $\alpha_1, \alpha_2, \alpha_3$ sont des constantes (propres à chaque substance).

   (a) En effectuant des transformations de variables, montrer que la recherche des constantes α_j dans la formule (2) se ramène à une régression linéaire multiple de la variable $\log(P)$ sur les variables explicatives $1/T$ et $\log(T)$.

   (b) Déterminer les valeurs estimées de α_j, $j = 1, 2, 3$. Comme la fonction log n'est pas définie en 0, on modifiera légèrement les données en remplaçant la première valeur de la variable `temperature` par 0.1. Cela peut être fait, par exemple, en utilisant la commande `edit`.

   (c) Le modèle

$$\log(P) = \beta_0 + \beta_1 \log(T) + \beta_2 / T + \varepsilon \tag{3}$$

   est-il plus fiable que celui de la question 2 ?

   (d) Déterminer la prédiction de la tension de vapeur du mercure correspondant à la température $T = 90; 230; 400$ selon ce nouveau modèle.

5. Effectuer une régression linéaire multiple de la variable $\log(P)$ sur les variables explicatives $\log(T)$, $1/T$, $1/T^2$ et T. Peut-on affirmer que la variable T est inutile (au seuil de 5%) ?

3. Exercice

On observe un échantillon i.i.d. bivarié $Z_1, \ldots, Z_n$ avec $Z_i = (X_i, Y_i)$ tel que

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

1. Montrer que les variables $\varepsilon_1, \ldots, \varepsilon_n$ sont i.i.d.
2. Supposons que $\varepsilon_1 \sim \mathcal{N}(0, \sigma^2)$, $X_1 \perp\!\!\!\perp \varepsilon_1$ et que X_1 admet une densité, notée p_X, indépendante de (β_0, β_1). Prouver que l'estimateur du maximum de vraisemblance de (β_0, β_1) coïncide avec l'estimateur des moindres carrés .