La régression linéaire: principes et méthodes

Cédric BATIONO

Aix Marseille Univ, INSERM, IRD, SESSTIM

MASTER INFORMATIQUE MEDICALE ET SCIENCE DES DONNEES
STATISTIQUES DESCRIPTIVES ET INFERENTIELLES (IFRISSE - Ouagadougou)
07 Mars 2022

Définition de la régression de Y en X

- Méthode qui permet de décrire comment Y varie en fonction de X.
- La distribution de Y quand X est fixé s'appelle : distribution conditionnelle de Y par rapport à X
- Autant de distribution conditionnelle que de valeurs de X
- Rend difficile la mesure de l'association entre X et Y
- Solutions :
  - caractériser les ditributions conditionnnelles par la moyenne $E(Y|x) = \mu_{Y|x}$ et la variance $V(Y|x) = \sigma^2_{Y|x}$
  - Etudier l'association entre X et $\mu_{Y|x}$ au lieu de celle entre Y et X.

$f(x) = E(Y|x)$

La fonction de régression de Y en X est celle qui décrit la variation de la moyenne conditionnelle de Y en fonction de X

Définition de la régression linéaire

- En pratique on ne s'interesse pas à a forme exacte (droite, exponentielle, parabole...) de la fonction f
- La droite est la forme la plus simple : pas toujours la plus adéquate

$f(x) = E(Y|x) = \alpha + \beta X$

De manière simplifiée : $\hat{y} = \alpha + \beta X$ où $\hat{y}$ est la valeur moyenne de pour un echantillon de sujets tel que $X = x$
De manière individuelle : $y = \alpha + \beta X + \epsilon$ où ϵ est appelé terme d'erreur mesurant l'écart entre la valeur individuelle y_i et la valeur moyenne $\hat{y}$.

Jeu de données sur la taille et le poids

Nous disposons d'un ensemble de données sur les tailles et les poids moyens de les femmes âgées de 30 à 39 ans.

```
# Le resumé des données montre:
data("women")
head(women)
```

```
##   height weight
## 1     58    115
## 2     59    117
## 3     60    120
## 4     61    123
## 5     62    126
## 6     63    129
```

```
summary(women)
```

```
##      height         weight
##  Min.   :58.0   Min.   :115.0
##  1st Qu.:61.5   1st Qu.:124.5
##  Median :65.0   Median :135.0
##  Mean   :65.0   Mean   :136.7
##  3rd Qu.:68.5   3rd Qu.:148.0
##  Max.   :72.0   Max.   :164.0
```

Jeu de données sur la taille et le poids

Nous disposons d'un ensemble de données sur les tailles et les poids moyens de les femmes âgées de 30 à 39 ans.

```
plot(women$height ~ women$weight)
abline(lm(women$height ~ women$weight, women), col="blue",lwd=2)
```

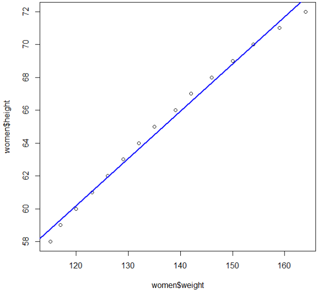


FIGURE – Droite de regression et nuage de points

Jeu de données sur la taille et le poids

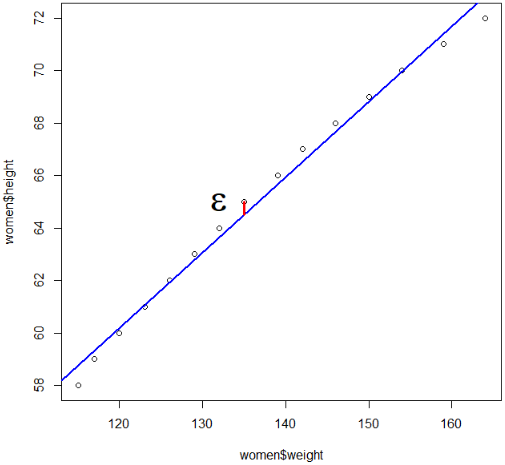


FIGURE – Droite de regression, nuage de points et erreur individuelle

Principe de l'estimation : la méthode des moindres carrés

- Première étape : calculer la Somme des Carrés des Ecarts

$$SCE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(\epsilon_i)^2 \qquad (1)$$

- Estimer α et β tel que : Somme des Carrés des Écarts prenne la valeur minimale
- En d'autres termes : trouver les paramètres de la droite de regression qui résume au mieux le n

Principe de l'estimation : Estimation de la pente β

- calculer β

$$b = \frac{cov(XY)}{var(X)} \quad (2)$$

- Estimation de la variance de X

$$S_x^2 = \frac{\sum_{i=1}^{n}(x_i^2) - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2)}{n-1} \quad (3)$$

- Estimation de la covariance de XY

$$cov(\hat{X}Y) = \frac{\sum_{i=1}^{n}(x_i y_i) - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n-1} \quad (4)$$

Exercice

Covariance de la taille et de l'âge :

```
cov(women$weight,women$height)

## [1] 69
```

Variance de l'âge

```
var(women$weight)

## [1] 240.2095
```

Estimation de β

```
b <- cov(women$weight,women$height)/var(women$weight)
b

## [1] 0.2872492
```

Estimation de α

- La droite passe par m_Y et m_X
- $m_Y = a + bm_X$
- $a = m_Y - bm_X$

Exercice

Estimation de α

```
a <- mean(women$height)-b*mean(women$weight)
a

## [1] 25.72346
```

l'équation s'écrit donc :

$$Taille = 25.72346 + 0.2872492 Poids + \epsilon$$

ou

$$E(Taille/Poids) = 25.72346 + 0.2872492 Poids$$

Interprétation

- Pente β :
  - ☞ $\beta = 0$: Il n' y a pas de lien en X et Y ou il y a une indépendance entre X et Y.
  - ☞ $\beta < 0$: X et Y évoluent dans le sens contraire. Par ex lorsque X augmente Y diminue.
  - ☞ $\beta > 0$: X et Y évoluent dans le même sens. Par ex lorsque X augmente Y augmente

Ordonnée à l'origine α

$$E(Y/X=0)=\alpha$$

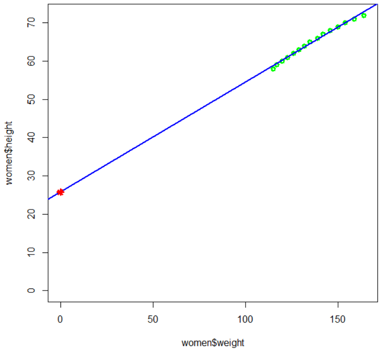


FIGURE – Valeur de Y quand $X = 0$

Test de la pente

Hypothèse

- H0 : $\beta = 0$,il n'y a pas de lien entre X et Y
- H1 : $\beta \neq 0$, il y a un lien entre X et Y

Prédiction : Sous H0

$$t_0 = \frac{b - \beta}{\sqrt{s_b^2}} \tag{5}$$

à $n - 2$ ddl et

$$\sqrt{s_b^2 = \frac{\frac{s_Y^2}{s_X^2} - b^2}{n - 2} (6)}$$

Conditions d'applications

Conditions :

- Relation linéaire entre X et Y (un écart à la linéarité entraine une perte de puissance)
- Une des deux distribution conditionnelle suit une loi Normale $L(Y/X) \sim N$
  - vérifié avec le Test de Kolmogorov-Smirnov par ex.
- Variance conditionnelle constante $var(Y/X)$
  - vérifié avec le test de Bartlett ou le test de Levene ou celui de Fisher en fonction des conditions d'app.
- Indépendance des individus

Exercice

- Application de la régression linéaire avec le logiciel R
- fonction *lm* sur R

```
mod1 <- lm(women$height~1+women$weight)
mod1

##
## Call:
## lm(formula = women$height ~ 1 + women$weight)
##
## Coefficients:
##  (Intercept)  women$weight
##      25.7235        0.2872
```

- $intercept = a$ et $weight = b$

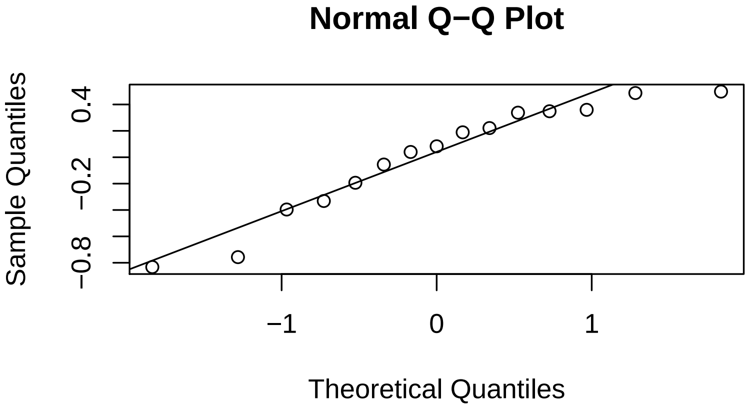
Exercice

```
summary(mod1)
```

```
##
## Call:
## lm(formula = women$height ~ 1 + women$weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.83233 -0.26249  0.08314  0.34353  0.49790
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.723456   1.043746   24.64 2.68e-12 ***
## women$weight  0.287249   0.007588   37.85 1.09e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.44 on 13 degrees of freedom
## Multiple R-squared:  0.991,Adjusted R-squared:  0.9903
## F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

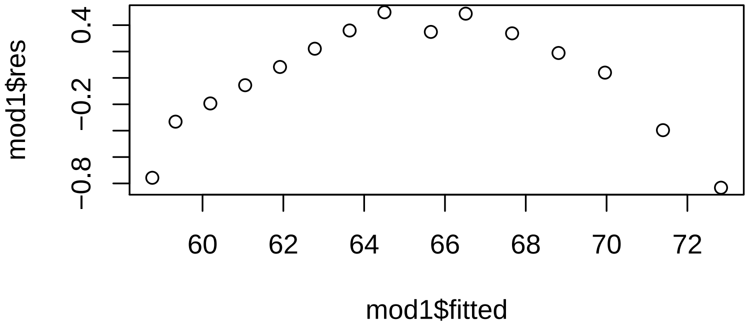
Verification des conditions d'application : Normalité

```r
qqnorm(mod1$res)
qqline(mod1$res)
```



Verification des conditions : relation linéaire entre X et Y

```r
plot(mod1$fitted, mod1$res)
```



Intervalle de confiance de la pente

- Variation aléatoire de b :
  $t_0 \pm t_{n-2,\frac{\alpha}{2}} * \sqrt{s_b^2}$
- Conditions d'application : idem (que pour la regression linéaire)

Exercice

- Intervalle de confiance des paramétres avec le logiciel R

```
confint(mod1)

##                    2.5 %     97.5 %
## (Intercept)  23.4685789 27.9783326
## women$weight  0.2708562  0.3036423
```

Définition de l'adéquation du modèle aux données

- Les observations sont il vraiment epliquées par les données ?
- Le pourcentage de variance expliqué :

$$R^2 = \frac{Pourcentage\ de\ variance\ expliqu\acute{e}\ par\ laregression}{variance\ totale}$$

$$\mathrm{R}^2 = \frac{\sum(m_{Y|x} - m_Y)^2}{\sum(y_i - m_Y)^2} (7)$$

- R correspond à l'estimation du coefficient de corrélation entre X et Y

Exercice

- Estimer l'adequation du modèle 1

```
r <- cor(women$weight,women$height)
r

## [1] 0.9954948

r*r

## [1] 0.9910098

var(mod1$fitted.value)/var(women$height)

## [1] 0.9910098
```

Exercice

- Estimer l'adequation du modèle 1 (recall)

```
summary(mod1)

##
## Call:
## lm(formula = women$height ~ 1 + women$weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.83233 -0.26249  0.08314  0.34353  0.49790
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.723456   1.043746   24.64 2.68e-12 ***
## women$weight  0.287249   0.007588   37.85 1.09e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.44 on 13 degrees of freedom
## Multiple R-squared:  0.991,Adjusted R-squared:  0.9903
## F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

Intérêts

- Déterminer plusieurs causes liés à Y
- ex : la Taille est lié à l'environnement, au niveau socio-economique, aux facteurs génétiques
- $E(Y|X_1, X_2, X_3) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
- Ajustement : estimation des paramètres en prenant en compte les variables (3)
- Permet de prendre en compte des interactions $E(Y|X_1, X_2, X_3) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 X_3$

Démarche

- Tests des β_1, β_2, β_3 à 0
- Interprétation identique
- Adéquation identique
- Approche pas à pas
- Choix des variables : notion de modèle
- Variables très corrélées

Exemple d'application

- Prédire la TAS en fonction de 5 variables
  - AGE : l'âge
  - DOSAGE : concentration biologique donnée
  - poids
  - taille
  - nbenfant
- Echantillon de 32 personnes

Exemple d'application

- En moyenne :

$$TAS = \alpha + \beta_1 x AGE + \beta_2 x DOSAGE + \beta_3 x poids + \beta_4 x taille + \beta_5 x nbenfant \quad (8)$$

- Description prealable des données : moyennes, variances, graphiques des distribution de chaque variable

Exemple d'application

- Estimation :

```
DATAtp <- read.csv2("DATAtp.csv",head=T)
dim(DATAtp)#nbre de d'indicidus et de variables

## [1] 32 16

tail(DATAtp,2)#2 dernieres observations

##     X Num Tabac TAS K IDM AGE SEXE PASSIF   DOSAGE GRAV DIG ATCD    poids
## 31 31  11     1 146 1   1  46    1      2 28.26142    2   1    2 48.63365
## 32 32  10     1 145 0   0  25    1      2 27.02906    2   1    2 50.48218
##      taille nbenfant
## 31 169.3168        2
## 32 170.2411        3

reg1 <-lm(TAS~1+ AGE+DOSAGE+poids+taille+nbenfant, data=DATAtp)
```

Exemple d'application

```
summary(reg1)


##
## Call:
## lm(formula = TAS ~ 1 + AGE + DOSAGE + poids + taille + nbenfant,
##     data = DATAtp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.131  -5.376   1.230   5.425  17.476
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45.2042   214.7751   0.210    0.835
## AGE           0.0595     0.1072   0.555    0.584
## DOSAGE        2.2880     0.2876   7.955  1.5e-08 ***
## poids         0.7254     4.6735   0.155    0.878
## taille            NA         NA      NA       NA
## nbenfant     -2.2858     6.3279  -0.361    0.721
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.268 on 27 degrees of freedom
## Multiple R-squared:  0.7043,Adjusted R-squared:  0.6604
## F-statistic: 16.07 on 4 and 27 DF,  p-value: 7.568e-07
```

Exemple d'application

modèle estimé

$$TAS = 45.20 + 0.05 x AGE + 2.28 x DOSAGE + 0.72 x poids - 2.28 x nbenfant \quad (9)$$

- Que faire ensuite ?

  conditions d'application

  intervalles de confiance des paramètres

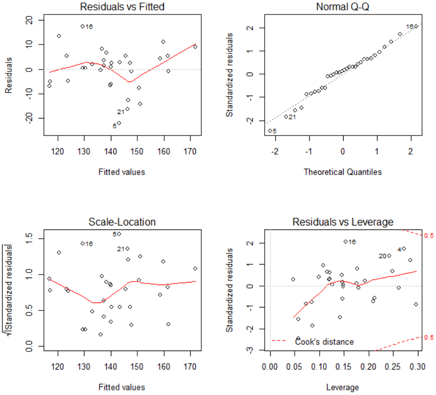
  adéquation : R^2

Conditions d'applications

- $L(Y/X) \sim N$
- V(Y/X) constantes pour tout X ; homoscédasticité
- indépendance des individus
- La régression est linéaire

exercice : verification des conditions

```
 par(mfrow=c(2,2))
plot(reg1)
par(mfrow=c(1,1))
```



exercice : verification des conditions

Graphe1 : doit être sans structure réparti de part et d'autre de l'axe des x

Graphe 2 : doit suivre la bissectrice

Graphe 3 : doit être sans structure

Graphe 4 : distances de Cook ou courbe de niveaux de leverage de distances de Cook's égales (influence des points à la stabilité du modèle)

exercice : intervalles de confiance et R^2

```
confint(reg1)
```

```
##                     2.5 %     97.5 %
## (Intercept) -395.4777866 485.886273
## AGE           -0.1605472   0.279548
## DOSAGE         1.6978730   2.878073
## poids         -8.8639060  10.314711
## taille                NA         NA
## nbenfant     -15.2695383  10.697939
```

```
#Adéquation: R2
var(reg1$fitted.value)/var(DATAtp$TAS)
```

```
## [1] 0.70425
```

Critére de sélection

- Guillaume d'Ockham, 1285-1349
  - « Les multiples ne doivent pas êtres utilisés sans nécessité »
  - Principe de parcimonie : pas d'ajout de nouvelles variables tant que celles présentes suffisent
  - Rique : overfitting $\sim$ hyperadéquation

Critére de sélection : AIC

- Akaike Information Criterion AIC

  $AIC = 2p - 2ln(L)$ où p correspond au nombre de parametres et L la vraisemblance.

  Par principe on cherche modèle avec le plus petit AIC
- Méthode de selection : pas à pas (forward ou ascendante, backward ou descendante, les deux sens)

exercice : selection

```
regsimple<-lm(TAS~1+DOSAGE, data=DATAtp)# modèle le plus simple
#Selection pas a pas descendant (choix)
aicmod <- MASS::stepAIC(reg1, scope=list(upper=reg1,lower=regsimple), direction=c("backward"))


## Start:  AIC=147.06
## TAS ~ 1 + AGE + DOSAGE + poids + taille + nbenfant
##
##
## Step:  AIC=147.06
## TAS ~ AGE + DOSAGE + poids + nbenfant
##
##            Df Sum of Sq    RSS    AIC
## - poids     1    2.0692 2321.0 145.09
## - nbenfant  1   11.2069 2330.2 145.21
## - AGE       1   26.4372 2345.4 145.42
## <none>                  2318.9 147.06
##
## Step:  AIC=145.09
## TAS ~ AGE + DOSAGE + nbenfant
##
##            Df Sum of Sq    RSS    AIC
## - AGE       1    31.968 2353.0 143.53
## - nbenfant  1   118.564 2439.6 144.68
## <none>                  2321.0 145.09
##
## Step:  AIC=143.53
## TAS ~ DOSAGE + nbenfant
##
##            Df Sum of Sq  RSS    AIC
## - nbenfant  1    110.02 2463 142.99
```

exercice : Modèle finale

```
summary(aicmod)

##
## Call:
## lm(formula = TAS ~ DOSAGE, data = DATAtp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.992  -4.138   1.712   5.057  17.317
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  79.1325     7.7875  10.162 3.15e-11 ***
## DOSAGE        2.2264     0.2751   8.093 4.92e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.061 on 30 degrees of freedom
## Multiple R-squared:  0.6859,Adjusted R-squared:  0.6754
## F-statistic:  65.5 on 1 and 30 DF,  p-value: 4.922e-09
```

Quelques réferences pour aller plus loin

- Thierry ANCELLE. "Statistiques". In : *Epidémiologie, Collection "sciences fondamentales", Maloine, Paris* (2002), p. 59–67.
- Cédric BATIONO. "Programmation statistique avec R (2020) Master santé numérique Ouagadougou". In : ().
- Jean BOUYER. *Epidémiologie : principes et méthodes quantitatives*. Lavoisier, 2009.
- Jean BOUYER et al. *Méthodes statistiques : médecine-biologie*. Estem, 1996.
- Juste GOUNGOUNGA. "régression linéaire (2017) Master santé publique Ouagadougou". In : ().
- Stian LYDERSEN, Morten W FAGERLAND et Petter LAAKE. "Recommended tests for association in 2× 2 tables". In : *Statistics in medicine* 28.7 (2009), p. 1159–1175.
- Dagnelie PIERRE. "Statistique théorique et appliquée 1". In : *Statistique descriptive et bases de l'inférence statistique, Bruxelles : De Boeck, DL* (2013).
- Daniel SCHWARTZ et Pierre DENOIX. "Méthodes statistiques à l'usage des médecins et des biologistes". In : (1963).