MSV601 - Modèle linéaire

Marie-Luce Taupin
marie-luce.taupin@univ-evry.fr

Laboratoire LaMME, Université d'Evry val d'Essonne
http://www.math-evry.cnrs.fr/members/mtaupin/welcome

2018-2019

Exemples : les données

Exemple 1

Etude de la durée de survie de patients atteints d'un cancer du poumon en fonction de l'âge, du sexe et du statut tabagique

Exemple 2

Etude du rendement de champs de mais en fonction du type d'engrais utilisé

Exemple 3

Etude de la tension artérielle en fonction de l'âge du patient

Exemples : les données

Exemple 4

Les données sont constituées de $n = 1429$ mesures couples circonférences-hauteur, mesures obtenues sur une parcelle d'eucalyptus âgés de 6 ans (âge de rotation avant la coupe).
But : prédire la hauteur d'un arbre à partir de sa circonférence.
Outil : trouver la relation qui lie la circonférence à la hauteur de façon à prédire la hauteur d'un arbre à partir de sa circonférence.

Exemple 5

Chez des patients ayant des problèmes cardiaques, on a mesuré la vitesse de circulation du sang (par effet Doppler) Y dans les artères coronaires. On cherche à étudier l'effet de deux variables quantitatives sur cette vitesse, à savoir le taux de cholestérol T et le poids P. On dispose des données suivantes : pour chaque patient i, $i = 1, \cdots, 20$ on mesure son poids p_i, son taux de cholestérol t_i, et sa vitesse de circulation sanguine y_i.

Exemples : les données

Exemple 6

On compare l'effet de trois traitements contre le paludisme, en mesurant le temps de clairance parasitaire chez des patients symptomatiques, répartis de façon aléatoire en trois groupes.

Exemple 7

On dispose d'informations (taux de cholestérol, fumeur ou non, etc...) sur une cohorte de 609 hommes ayant été suivis sur une période de 7 ans. Il s'agit d'étudier la variable d'intérêt "apparition ou non d'une maladie cardiaque des coronaires".

Exemples : les données

Exemple 8

Considérons un gène bi-allélique d'allèles B et b, dont on soupçonne qu'il module la concentration sanguine d'une certaine protéine. On a recruté une centaine d'individus, on les a génotypés pour le gène considéré et on a mesuré la concentration sanguine chez chacun d'eux

Modèle statistique : formalisation

Dans tous les exemples présentés ci-dessus, on dispose

- une variable expliquée : celle qu'on cherche à prédire Y
- des variables explicatives : variables qui expliquent X

Modèle statistique : formalisation

Par exemple :
exemple 1 : Y durée de vie du patient (quantitative continue positive), X l'âge, du sexe et du statut tabagique
exemple 4 : Y hauteur de l'arbre (quantitative continue), X circonférence (quantitative continue)
exemple 6 : Y temps de clairance parasitaire (qantitative continue), X : traitement contre le paludisme (qualitative)
exemple 7 : Y binaire, $Y = 1$ si pathologie des coronaires, $Y = 0$ sinon. X : taux de cholestérol, fumeur ou non...

On cherche à prédire Y en fonction de X.
Nouvel individu avec $\tilde{X} \rightarrow$ prédiction de Y notée $\tilde{Y}$.

Modèle statistique : formalisation

Contexte du cours

Dans ce cours on considère Y **quantitative continue** et X **quantitative continue** ou **qualitative**.

Modélisation

On cherche une relation du type $Y = F(X)$.

Cette relation ne sera jamais exacte

$$Y = F(X) + \text{erreur}.$$

On se restreint à un certain type de modèles : le modèle linéaire

Le modèle linéaire

Objectifs

- Expliquer les variations de Y en fonction de X
- Prédire les valeurs de Y à partir des valeurs de X

Intérêt

- Simplicité des algorithmes d'estimation et des tests
- Utilisable dans la plupart des situations

Le modèle linéaire

Types de modèles linéaires

On distingue 3 types de modèles linéaires :

- X qualitative (Analyse de la Variance ANOVA)
- X quantitatif (Modèle de régression REGRESSION)
- X des 2 types (ANCOVA)

Le modèle linéaire

Analyse de la Variance

• La variable X est qualitative : elle définit des groupes suivant les modalités qu'elle prend.

• On se demande si la variable Y a la même espérance (moyenne) dans chacun des groupes définis par les modalités de X.

• Exemple :

- X traitement contre la paludisme, trois types de traitement $=>$ trois groupes de traitement.
- Y temps de clairance parasitaire.
- On se demande si $E(Y|\text{ traitement 1}) = E(Y|\text{ traitement 2}) = E(Y|\text{ traitement 3})$.

Le modèle linéaire

Modèle de régression linéaire

- La variable X est quantitative.
- On modélise l'évolution de Y par une fonction de X.
- Exemple 1 :
  - X circonférence d'un eucalyptus.
  - Y la hauteur d'un eucalyptus.
  - On se demande comment prédire la hauteur à partir de la circonférence.

Le modèle linéaire

Modèle de régression linéaire

• Exemple 2 : Prostate Cancer The data for this example come from a study by Stamey et al. (19 89). They examined the correlation between the level of prostate-specific antigen and a number of clinical measures in men who were about to receive a radicalprostatectomy. The variables are log cancer volume (*lcavol*), log prostate weight (*lweight*), *age* , log of the amount of benign prostatic hyperplasia (*lbph*), seminal vesicle invasion (*svi*), log of capsular penetration (*lcp*), Gleason score (*gleason*), and percent of Gleason scores 4 or 5 (*pgg*45).

Le modèle linéaire

Modèle général

$$Y = X\theta + \epsilon$$

$$\text{où } \epsilon \sim N(0, \sigma^2 I_p)$$

Remarque

Le modèle linéaire est linéaire en ses paramètres (et non en X)

Régression linéaire simple

Exemples

- Tension artérielle = f(age)
- Rendement de blé = f(dose de fertilisant)
- Concentration ozone = f(température)
- Effet d'un traitement = f(dose)
- Taux de DDT = f(age du brochet)
- Hauteur=f(circonférence)
- Niveau d'antigène=f(*lcavol*,*lweight*,*age* ,*lbph*,*svi*, *lcp*,*gleason*,*pgg*45)

On se retreint ici à des f simples linéaires en les paramètres, mais pas forcément en X.

Régression linéaire simple

Le modèle

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

avec :

- X_i = circonférence, ou $X_i = \sqrt{\text{circonférence}}$, ou $X_i = \text{circonférence}^2$ de l'arbre $i, \cdots$
- Y_i hauteur de l'arbre i
- ϵ_i représente une erreur commise par la modélisation. Liée à : luminosité, positionnement de l'arbre, facteurs individuels,...

Objectifs

- Estimer les paramètre β_0 et β_1 par $\widehat{\beta}_0$ et $\widehat{\beta}_1$.
- Prédire la hauteur $\tilde{Y}$ d'un nouvel arbre à partir de sa circonférence $\tilde{X}$.

$$\tilde{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \tilde{X}.$$

Régression linéaire simple

Remarques

- Le modèle est linéaire en ses paramètres (pas nécessairement en la circonférence)
- La relation n'est pas forcément la vraie relation, elle peut être une approximation de la vraie relation.

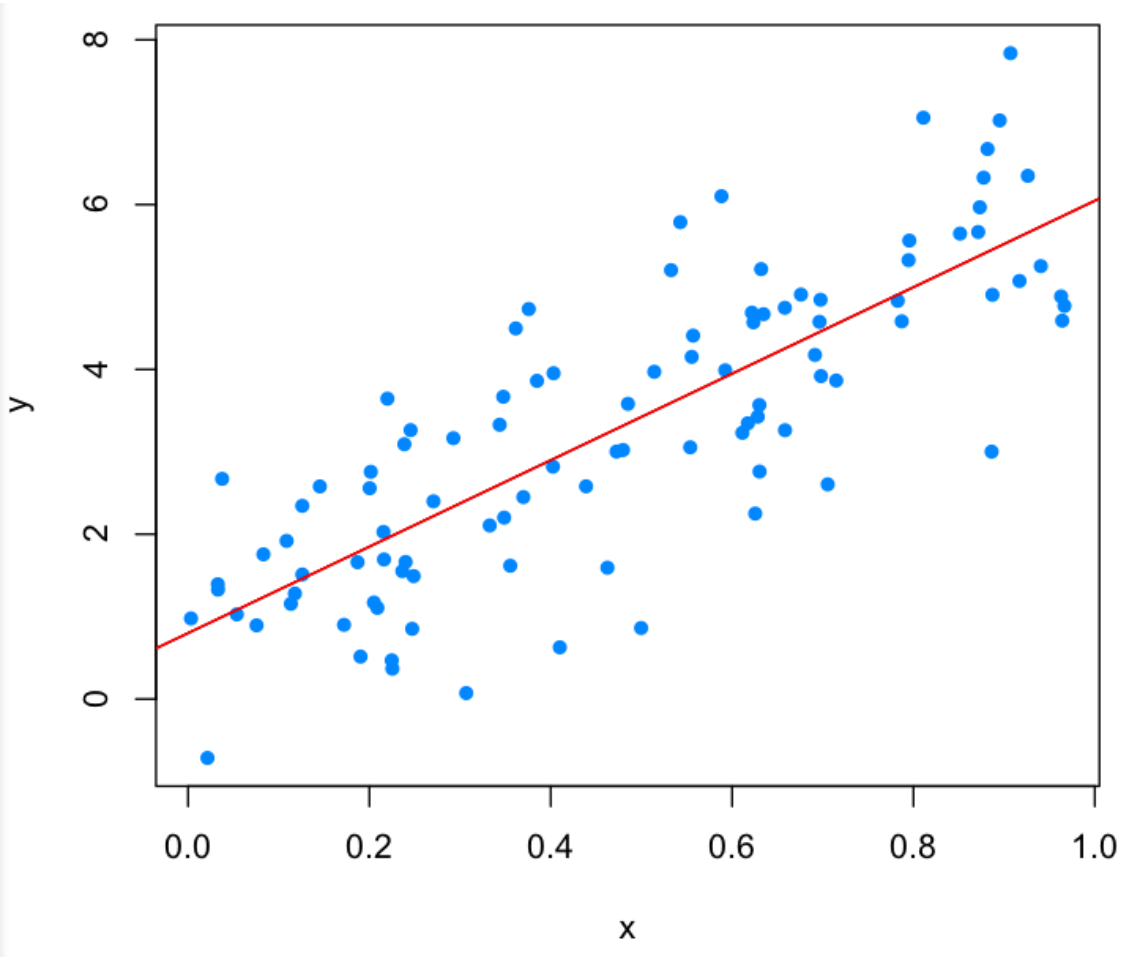
Hypothèses

Les erreurs ϵ_i sont des **variables aléatoires** indépendantes et de même loi (i.i.d.) telles que

- $E(\epsilon_i) = 0$
- $\text{Var}(\epsilon_i) = \sigma^2$ (constante)
- $\epsilon_i \perp\!\!\!\perp \epsilon_j; \forall i \neq j$
- $\epsilon_i \sim N(0, \sigma^2)$ (hypothèse nécessaire pour les tests)

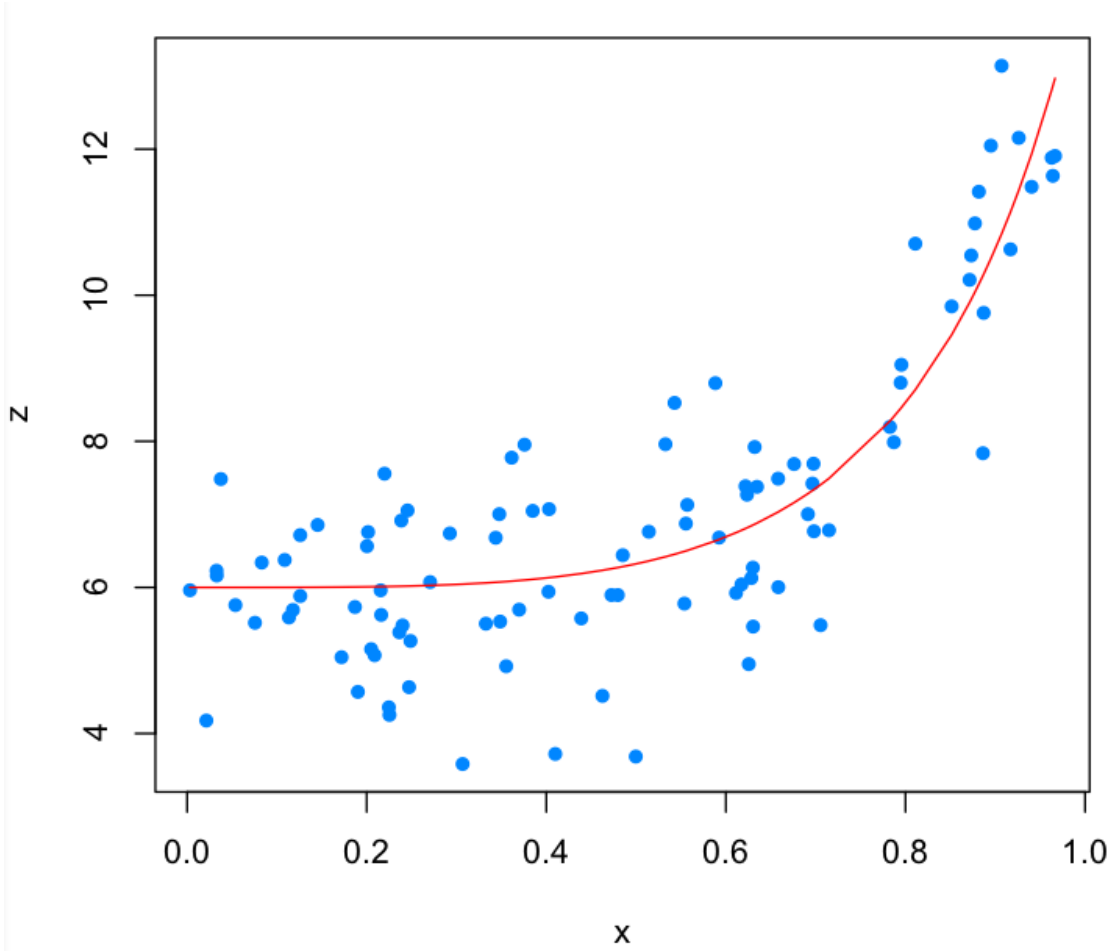
Régression linéaire simple

Représentation graphique : nuage de points



Régression linéaire simple

Représentation graphique : nuage de points



Remarque

Régression linéaire simple

Vocabulaire

- X est une variable, aléatoire ou contrôlée, dite **explicative**
- Y est une variable aléatoire dite **à expliquer**

Remarque

Si X n'est pas aléatoire, $\text{Cov}(X, Y)$ peut être calculé mais n'a pas de sens

Régression linéaire simple

Questions d'intérêt

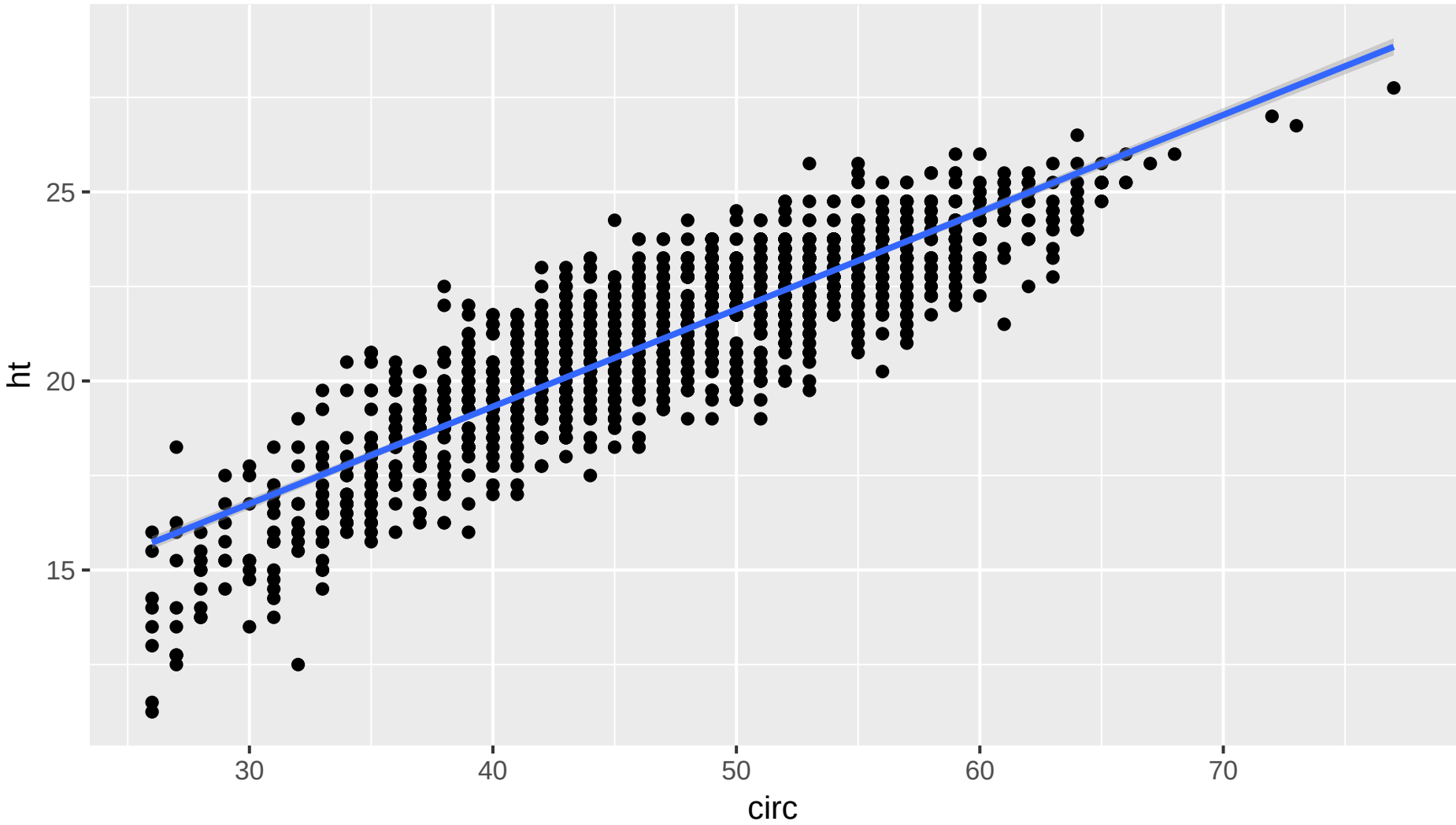
- Existe-t-il une relation entre X et Y ?
- Quelle est la forme de la relation ?
- Peut-on prédire Y à partir des valeurs de X ?

Démarche

- Estimation des paramètres du modèle (β_0,β_1, et σ^2)
- Tests (paramètres / validité du modèle)
- Sélection de modèles
- Prédictions

Application : eucalyptus

Nuage de points



Application

On considère le modèle $ht = \beta_0 + \beta_1 * circ$.

```
##
## Call:
## lm(formula = ht ~ circ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7659 -0.7802  0.0557  0.8271  3.6913
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.037476   0.179802   50.26   <2e-16 ***
## circ        0.257138   0.003738   68.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.199 on 1427 degrees of freedom
## Multiple R-squared:  0.7683,Adjusted R-squared:  0.7682
## F-statistic:  4732 on 1 and 1427 DF,  p-value: < 2.2e-16
```

Application

Eucalyptus- commentaires

- La droite de régression estimée a pour equation

$$ht = 9.03 + 0.25 * circ.$$

- la coefficient correspondant à la pente est positif et vaut 0.25.
- Pour une circonférence de 40 la hauteur prédite est

$$ht_{pred} = 9.03 + 0.25 * 40 = 19m.$$

- pour l'arbre i, l'erreur de prédiction vaut

$$ht_i - ht_{pred,i} = ht_i - (9.03 + 0.25 * circ_i).$$

Application

Eucalyptus- autres modèles

On peut essayer d'autres modèles comme

$$ht = \beta_0 + \beta_1 * \sqrt{circ},$$

ou

$$ht = \beta_0 + \beta_1 * circ^2,$$

ou

$$ht = \beta_0 + \beta_1 * circ + \beta_2 * circ^2.$$

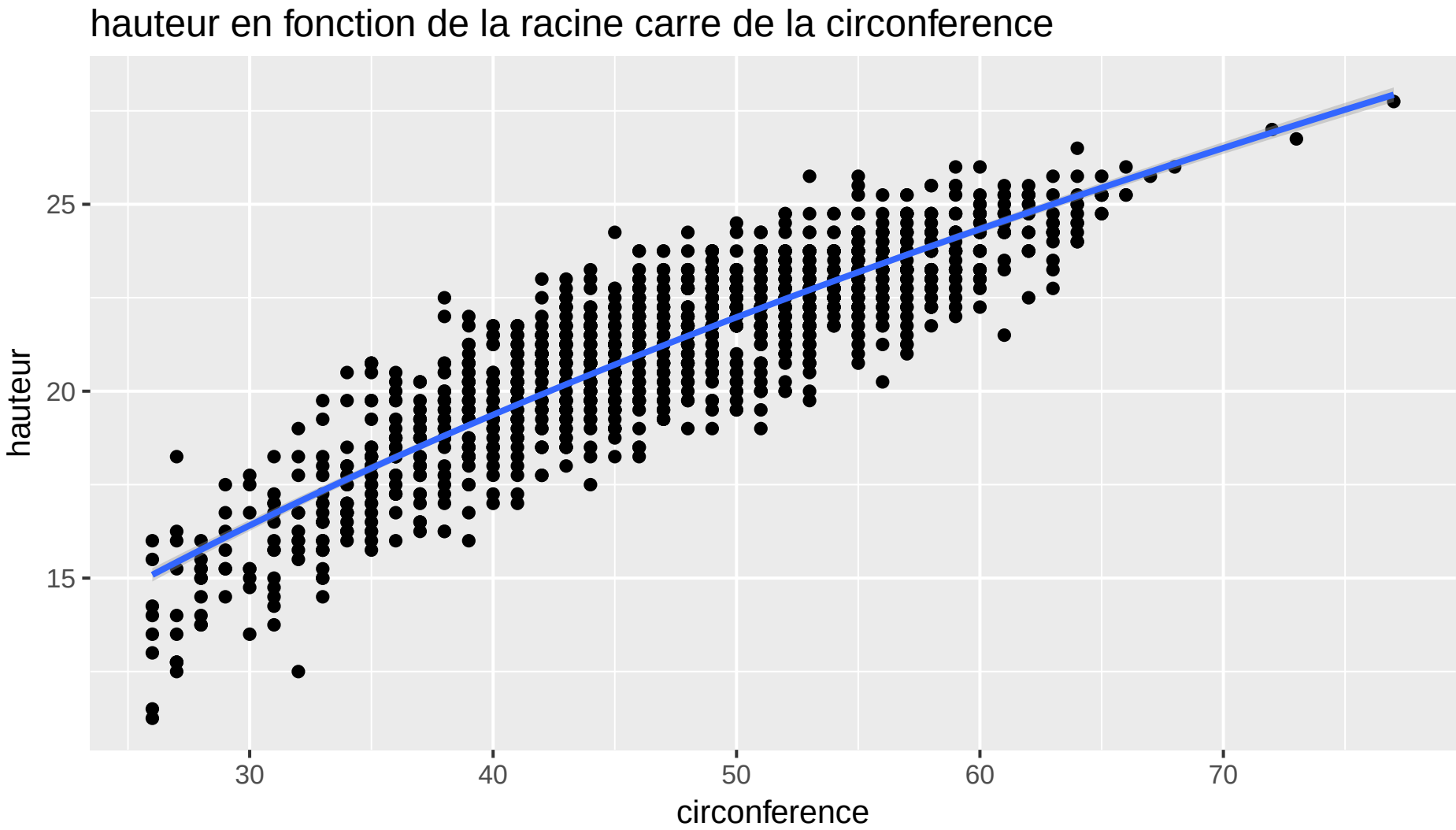
Ces modèles sont aussi des modèles linéaires en fonction de la racine carrée de la circonférence ou du carré de la circonférence.

Application : Eucalyptus- autre modèles

On considère le modèle $ht = \beta_0 + \beta_1 * \sqrt{circ}$.

```
##
## Call:
## lm(formula = ht ~ I(sqrt(circ)), data = euca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5360 -0.7249  0.0265  0.7813  3.6904
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.73036    0.33600  -8.126 9.51e-16 ***
## I(sqrt(circ))  3.49424    0.04883  71.560  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.163 on 1427 degrees of freedom
## Multiple R-squared:  0.7821,Adjusted R-squared:  0.7819
## F-statistic:  5121 on 1 and 1427 DF,  p-value: < 2.2e-16
```

Application : Eucalyptus- racine carree

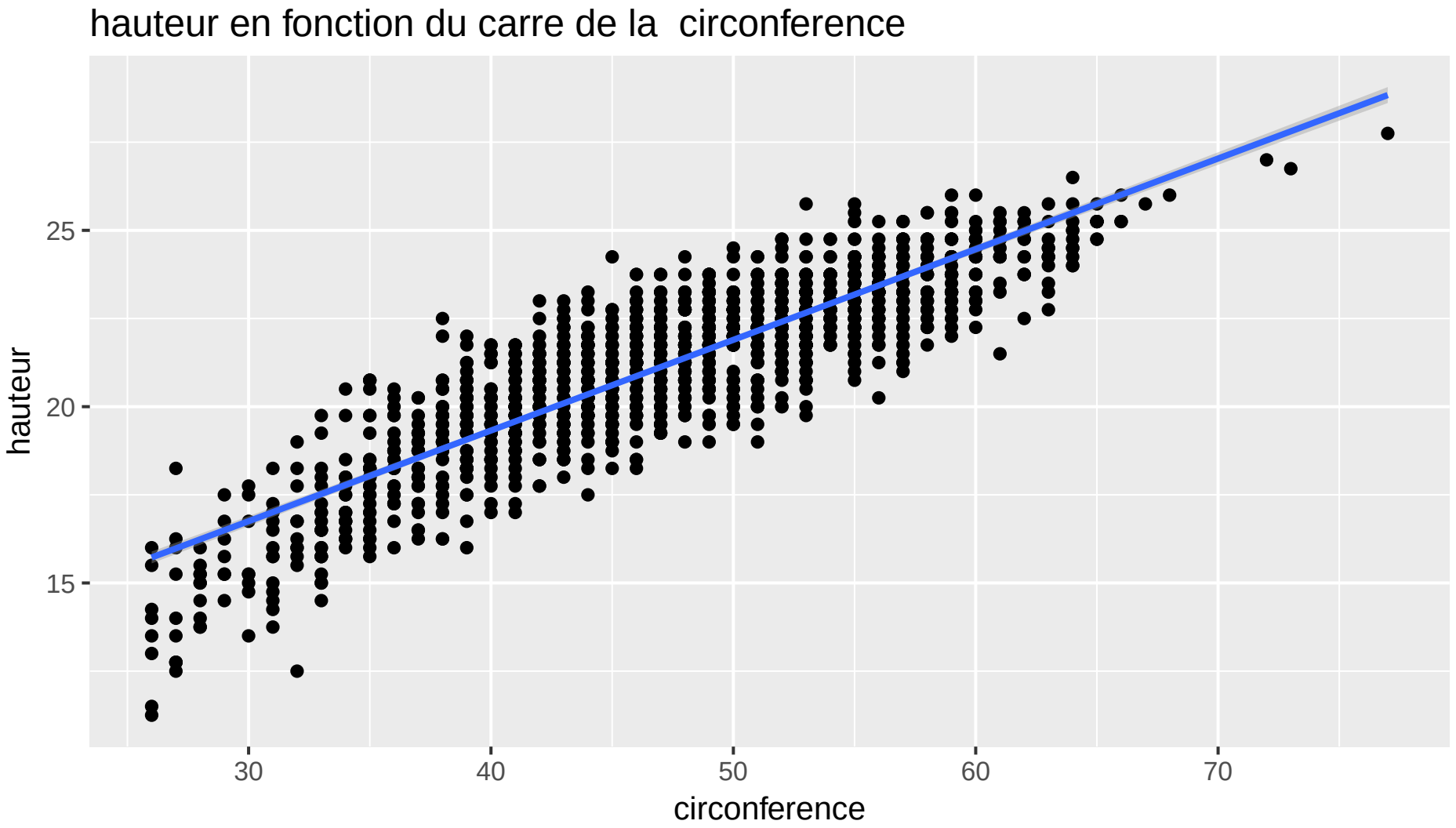


Application : Eucalyptus - en fonction du carré

On considère le modèle linéaire $ht = \beta_0 + \beta_1 * circ^2$.

```
##
## Call:
## lm(formula = ht ~ I(circ^2), data = euca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5942 -0.8184  0.0551  0.8449  3.8080
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.504e+01  1.056e-01  142.46   <2e-16 ***
## I(circ^2)   2.667e-03  4.315e-05   61.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.299 on 1427 degrees of freedom
## Multiple R-squared:  0.7281,Adjusted R-squared:  0.7279
## F-statistic:  3821 on 1 and 1427 DF,  p-value: < 2.2e-16
```

Application : Eucalyptus- en fonction du carré



Application

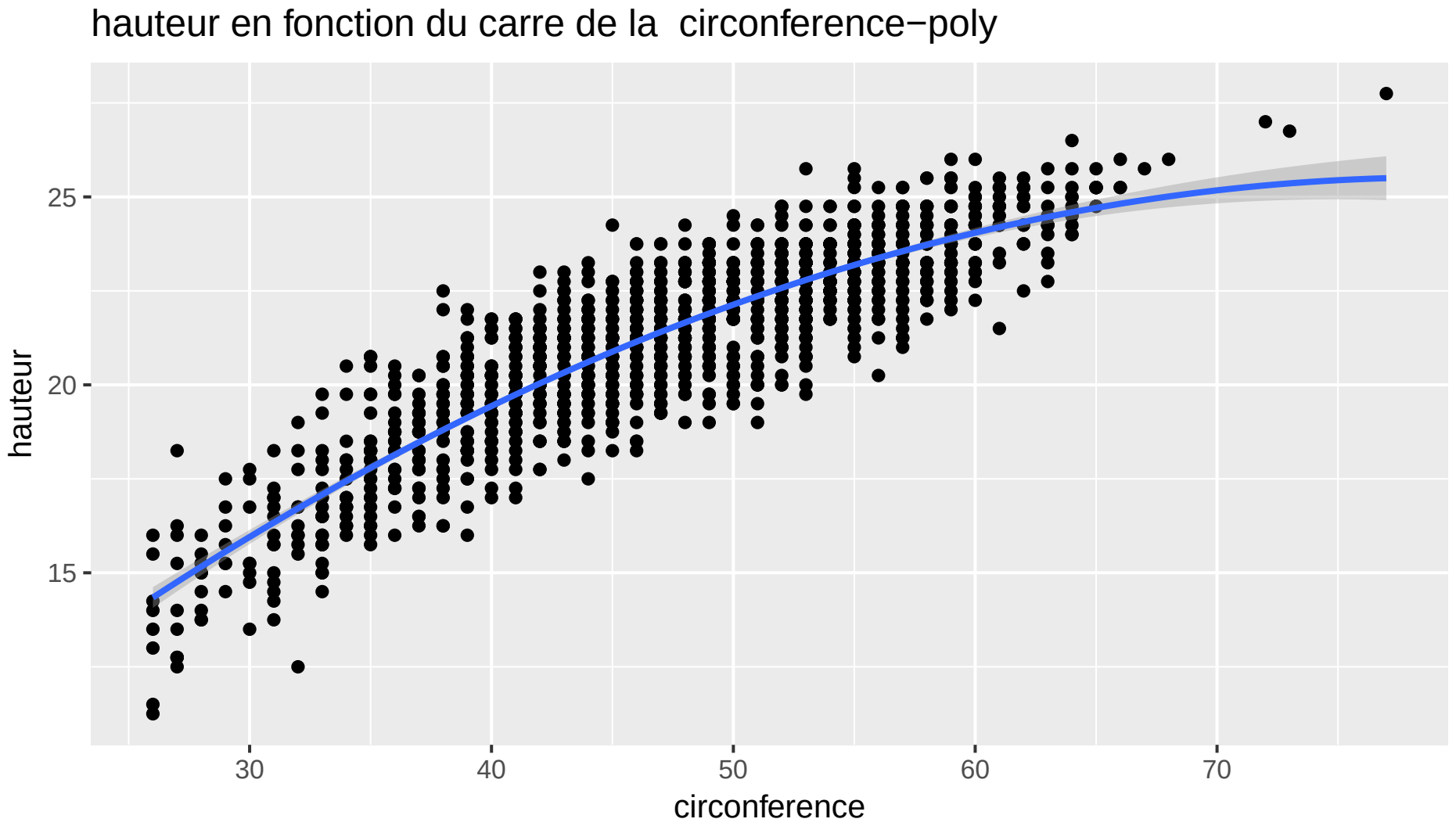
Eucalyptus- autre modèles

On considère le modèle $ht = \beta_0 + \beta_1 * circ + \beta_2 * circ^2$.

```
reg3un<-lm(ht~circ+I(circ^2),data=euca)
summary(reg3un)


##
## Call:
## lm(formula = ht ~ circ + I(circ^2), data = euca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2140 -0.6947  0.0360  0.7732  3.6970
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.8028038  0.7012035   1.145    0.252
## circ         0.6227415  0.0303984  20.486   <2e-16 ***
## I(circ^2)   -0.0039224  0.0003239 -12.110   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.142 on 1426 degrees of freedom
## Multiple R-squared:  0.7899,Adjusted R-squared:  0.7896
## F-statistic:  2681 on 2 and 1426 DF,  p-value: < 2.2e-16
```

Application : Eucalyptus- polynome de degré 2



Application

Eucalyptus-Questions

- Q1 Quels sont les modèles, comment analyser les sorties R ?
- Q2 Tous les coefficients sont-ils significatifs ?
- Q3 Quel modèle choisir ? le modèle qui prédit le mieux ? celui qui a le moins de coefficients ?

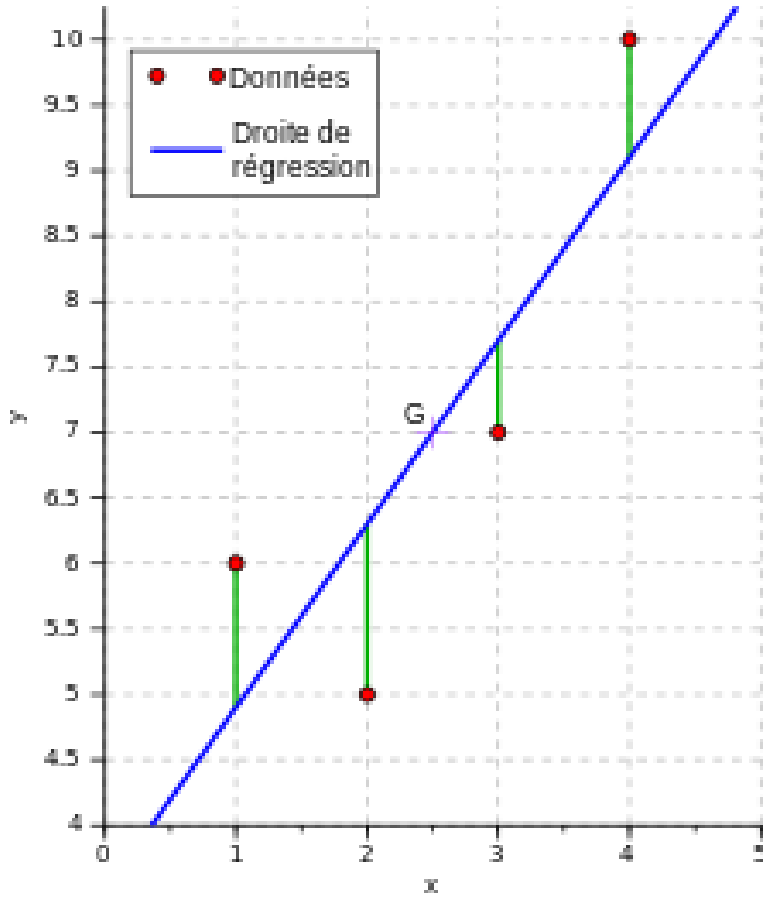
Application

Eucalyptus- réponses statistiques

- Q1 Estimation des paramètres des modèles
- Q2 Tests statistiques
- Q3 Comparaison de modèles, sélection de variables

Méthodes des moindres carrés

Illustration graphique



Méthode des moindres carrés

Formalisation

On note

- Y_i la hauteur observée pour l'arbre i
- X_i la circonférence de l'arbre i
- $\beta_0 + \beta_1 X_i$ la hauteur prédite par la droite $y = \beta_0 + \beta_1 x$
- $e_i = Y_i - (\beta_0 + \beta_1 X_i)$ l'écart entre la valeur observée Y_i et la valeur prédite $\beta_0 + \beta_1 X_i$.

On cherche β_0 et β_1 telle que la droite soit la plus proche de tous les points du nuage (observations).
On cherche β_0 et β_1 telle que la somme des carrés des erreurs soit la plus petite possible, *i.e.* telle

$$\sum_{i=1}^{n} e_i^2, \text{ soit la plus petite possible.}$$

Méthode des moindres carrés

Dans ce qui suit, on suppose : $\epsilon_i \sim N(0, \sigma^2)$

Formalisation

Les **estimateurs des moindres carrés ordinaires** (MCO) de $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont ceux qui minimisent la somme des carrés des résidus :

$$SCR(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

i.e pour tout β_0, β_1

$$SCR(\widehat{\beta}_0, \widehat{\beta}_1) \leq SCR(\beta_0, \beta_1).$$

La droite d'équation $y = \widehat{\beta}_0 + \widehat{\beta}_1 * x$ s'appelle **la droite des moindres carrés** ou **droite de régression**.

Estimation des paramètres

Proposition

Les estimateurs des moindres carrés ordinaires ont pour expression :

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

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

Remarques

- Si les résidus suivent une distribution normale, les estimateurs des moindres carrés sont les mêmes que les estimateurs du maximum de vraisemblance
- La droite des moindres carrés passe par le barycentre du nuage de points $(\overline{X}, \overline{Y})$

Estimation des paramètres

Propriétés

- $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs sans biais de β_0 et β_1
- Conditionnellement $X_1 = x_1, \cdots, X_n = x_n$, les variances des estimateurs sont :

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right)$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- Conditionnellement $X_1 = x_1, \cdots, X_n = x_n$, leur covariance est :

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- Si les résidus suivent une distribution normale, $\hat{\beta}_0$ et $\hat{\beta}_1$ suivent aussi une distribution normale

Estimation des paramètres

Propriétés

- Conditionnellement $X_1 = x_1, \cdots, X_n = x_n$, les variances des estimateurs sont données par

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right)$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

- Ces variances des estimateurs sont estimées par

$$S^2 = \frac{1}{n-2} \sum_{i=1}^n [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon}_i^2$$

$$S_0^2 = S^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right) \text{ et } S_1^2 = \frac{S^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

Estimation des paramètres

Résidus estimés

Les résidus estimés ou erreurs de prédictions sont :

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

où $\hat{Y}_i$ est la valeur prédite par le modèle :

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

Propriété

$$\sum \hat{\epsilon}_i = 0$$

Remarque : Contrairement aux résidus ϵ_i, les résidus estimés $\hat{\epsilon}_i$ ne sont pas indépendants

Tests sur les paramètres du modèle

Hypothèses testées

$$\begin{cases} (H_0) : \text{la hauteur ne dépend pas de la circonférence} \\ (H_1) : \text{la hauteur dépend linéairement de la circonférence} \end{cases}$$

Hypothèses formulées sur le paramètre β_1

$$\begin{cases} (H_0) : ht = \beta_0 \text{ Modèle constant} \\ (H_1) : ht = \beta_0 + \beta_1 * circ \end{cases} \iff \begin{cases} (H_0) : \beta_1 = 0 \\ (H_1) : \beta_1 \neq 0 \end{cases}$$

Tests sur les paramètres du modèle

Ecriture des hypothèses en terme de comparaison de modèles

Deux modèles en compétition

$$\begin{cases} (H_0) : \mathcal{M}_1 : ht = \beta_0, \quad \beta_1 = 0 \\ (H_1) : \mathcal{M}_2 : ht = \beta_0 + \beta_1 * circ. \end{cases}$$

Notion de dimension de modèle :

$$\begin{cases} (H_0) : \mathcal{M}_1 : (\beta_0, \beta_1) = (\beta_0, 0), \quad \dim(\mathcal{M}_1) = 1. \\ (H_1) : \mathcal{M}_2 : (\beta_0, \beta_1) \in \mathbb{R}^2, \quad \dim(\mathcal{M}_2) = 2. \end{cases}$$

Tests

Tests de nullité de l'un des paramètres

Les tests sont basés sur les outils

$$\frac{(\hat{\beta}_0 - \beta_0)}{S_0} \sim \mathcal{S}t(n-2) \text{ et } \frac{(\hat{\beta}_1 - \beta_1)}{S_1} \sim \mathcal{S}t(n-2).$$

Tests

Tests de nullité de β_1 ie $\beta_1 = 0$

Pour tester $(H_0) : \beta_1 = 0$ contre $(H_1), \beta_1 \neq 0$, on utilise la Statistique de test :

$$W_n = \frac{\hat{\beta}_1 - 0}{S_1} = \frac{\hat{\beta}_1}{\sqrt{\frac{S^2}{\sum_{i=1}^n (X_i - \overline{X})^2}}} \underset{H_0}{\sim} \mathcal{S}t(n-2)$$

On rejette (H_0) si $|W_n| > seuil$ où $P_{H_0}(|W_n| > seuil) \leq \alpha$ avec $P_{H_0}(|W_n| > seuil) = P_{H_0}(|\mathcal{S}t(n-2)| > seuil) \leq \alpha$ où aussi si

$$p - value \leq \alpha, \text{ avec } p - value = P(\mathcal{S}t(n-2) > |W_{n,obs}|).$$

Tests

Formulation en terme de comparaison de modèles emboités

$$\begin{cases}(H_0) : \mathcal{M}_1 : ht = \beta_0, \quad \beta_1 = 0 \\ (H_1) : \mathcal{M}_2 : ht = \beta_0 + \beta_1 * circ.\end{cases}$$

Notion de dimension de modèle :

$$\begin{cases}(H_0) : \mathcal{M}_1 : (\beta_0, \beta_1) = (\beta_0, 0), \quad \dim(\mathcal{M}_1) = 1. \\ (H_1) : \mathcal{M}_2 : (\beta_0, \beta_1) \in \mathbb{R}^2, \quad \dim(\mathcal{M}_2) = 2.\end{cases}$$

Dans le modèle $\mathcal{M}_1$, $\beta_1 = 0$ et β_0 est estimé par $\overline{Y}$ (hauteur moyenne) et la prédiction est $\hat{Y}(\mathcal{M}_1) = \overline{Y}$.
Dans le modèle $\mathcal{M}_2$, (β_0, β_1) sont estimés par $(\hat{\beta}_0, \hat{\beta}_1)$ (Moindres carrés) et la prédiction est $\hat{Y}(\mathcal{M}_2) = \hat{\beta}_0 + \hat{\beta}_1 X$.

Comparaison de modèles

Notations

- Soit $\mathcal{M}_2$ un modèle à 2 paramètres. $SCR(\mathcal{M}_2)$ est la somme des carrés résiduels associée.
- Soit $\mathcal{M}_1$ un modèle à 1 paramètres emboîté dans $\mathcal{M}_2$. $SCR(\mathcal{M}_1)$ est la somme des carrés résiduels associée.
- Soit $SCE = SCR(\mathcal{M}_1) - SCR(\mathcal{M}_2)$ la partie de la SCR expliquée par le passage du petit modèle $\mathcal{M}_1$ au grand modèle $\mathcal{M}_2$.
- Soit n le nombre total de mesures.

Comparaison de modèles emboîtés

Théorème

- $SCR(\mathcal{M}_1) \sim \sigma^2 \chi^2_{n-1}$
- $SCR(\mathcal{M}_2) \sim \sigma^2 \chi^2_{n-1}$
- $SCM = SCR(\mathcal{M}_1) - SCR(\mathcal{M}_2) \sim \sigma^2 \chi^2_{2-1}$
- $SCM \perp\!\!\!\perp SCR(\mathcal{M}_2)$

Tests

Formulation en terme de comparaison de modèles emboités

On utilise

$$T_n = \frac{SCM/(2-1)}{SCR(\mathcal{M}_2)/(n-2)} \underset{H_0}{\sim} \mathcal{F}(1, n-2), (\text{ loi de Fisher })$$

$$\text{avec } SCM = SCR(\mathcal{M}_1) - SCR(\mathcal{M}_2) = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 X_i - \overline{Y})^2$$

$$SCR(\mathcal{M}_1) = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \quad \text{et } SCR(\mathcal{M}_2) = \sum_{i=1}^{n} (Y_i - \bar{Y}_i)^2.$$

On rejette (H_0) si T_n prend de grandes valeurs soit $T_n > seuil$ où $P_{H_0}(T_n > seuil) = P_{H_0}(\mathcal{F}(1, n-2) > seuil) \leq \alpha$ où aussi si

$$p - value \leq \alpha, \text{ avec } p - value = P(\mathcal{F}(1, n-2) > T_{n,obs}).$$

Remarque : $T_n = W_n^2$. Ce sont donc les mêmes tests !

Tests

Formulation en terme de comparaison de modèles

On peut aussi écrire

$$
\begin{aligned}
T_n &= \frac{\frac{SCR(\mathcal{M}_1) - SCR(\mathcal{M}_2}{\dim(\mathcal{M}_2) - \dim(\mathcal{M}_1)}}{SCR(\mathcal{M}_2)/(n-2)} \\
&= \frac{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i(\mathcal{M}_1))^2 - \sum_{i=1}^n (Y_i - \hat{Y}_i(\mathcal{M}_2))^2}{\dim(\mathcal{M}_2) - \dim(\mathcal{M}_1)}}{\sum_{i=1}^n (Y_i - \hat{Y}_i(\mathcal{M}_2))^2/[n - \dim(\mathcal{M}_2)]}.
\end{aligned}
$$

Comparaison de modèles emboîtés

Tableau de synthèse

Source	Degrés de liberté	Sommes des carrés	Carrés moyens	F
Gain $\mathcal{M}_2/\mathcal{M}_1$	$2-1$	$(SCR(\mathcal{M}_1) - SCR(\mathcal{M}_2))$	$\frac{(SCR(\mathcal{M}_1)-SCR(\mathcal{M}_2))}{2-1}$	$\frac{(SCR(\mathcal{M}_1)-SCR(\mathcal{M}_2))/(2-1)}{\frac{SCR(\mathcal{M}_2)}{(n-2)}}$
$\mathcal{M}_2$	$n-2$	$SCR(\mathcal{M}_2)$	$\frac{SCR(\mathcal{M}_2)}{(n-2)}$	
$\mathcal{M}_1$	$n-1$	$SCR(\mathcal{M}_1)$	$\frac{SCR(\mathcal{M}_1)}{(n-1)}$	

Prédiction

Valeur prédite

Soit x_{n+1} une nouvelle observation. La valeur prédite par le modèle est : $Y_{n+1} = \beta_0 + \beta_1 X_{n+1} + \epsilon_{n+1}$. Cette valeur peut être approchée par :

$$\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 X_{n+1}$$

Remarque

Deux types d'erreurs entâchent cette prédiction :

- La non connaissance de ϵ_{n+1}
- L'incertitude sur l'estimation des paramètres β_0 et β_1

Prédiction

Proposition

Soit $\hat{\epsilon}_{n+1} = Y_{n+1} - \hat{Y}_{n+1}$ l'erreur de prévision. Conditionnellement à $(X_1 = x_1, \cdots, X_n = x_n)$, on a :

$$\mathbb{E}(\hat{\epsilon}_{n+1}) = 0$$

$$\text{Var}(\hat{\epsilon}_{n+1}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)$$

Remarques

- La variance de l'erreur de prédiction est d'autant plus grande que x_{n+1} est éloigné de la moyenne $\overline{x}$

Ecriture matricielle du modèle

Le modèle

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

Hypothèses :

- $E(\epsilon_i) = 0$
- $\text{Var}(\epsilon_i) = \sigma^2$ (constante)
- $\epsilon_i \perp\!\!\!\perp \epsilon_j; \forall i \neq j$
- $\epsilon_i \sim N(0, \sigma^2)$

Régression linéaire simple

Le modèle - écriture matricielle

$$\mathbf{Y} = \mathbf{X}\beta + \mathcal{E}$$

$$\begin{pmatrix} Y_1 \\ . \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ . \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Hypothèses :

- $E(\mathcal{E}) = \mathbf{0}$
- $\text{Var}(\mathcal{E}) = \sigma^2 \mathbf{I_n}$
- $\mathcal{E} \sim N(0, \sigma^2 \mathbf{I_n})$ (pour les tests)

Estimation des paramètres

Estimateurs des moindres carrés

L'**estimateurs des moindres carrés ordinaires** (MCO) de $\beta = (\beta_0, \beta_1)^T$ est le vecteur aléatoire de $\mathbb{R}^p$ $\hat{\beta}$ qui minimise la somme des carrés des résidus :

$$SCR(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|^2$$

Théorème

Si la matrice $\mathbf{X}^T\mathbf{X}$ est inversible, l'estimateur des moindres carrés ordinaires de β est :

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Validation du modèle

Graphe des résidus

Validation du modèle

Tests sur les résidus

- Test d'indépendance
- Test de normalité

Régression linéaire multiple

Exemples

- Concentration d'ozone=f(température,vent,nébulosité)
- Vitesse de circulation coronarienne=f(poids, taux de cholestérol)

Régression linéaire multiple

Formalisation

- On modélise l'évolution de Y par une fonction de X.
- La variable $X = (1, X_1, \cdots, X_d)$ est quantitative de dimension d, $d > 1$.
-

$$Y = X\theta + \varepsilon,$$

$$Y = \theta_0 + X_1\theta_1 + \cdots + X_d\theta_d + \varepsilon.$$

Régression linéaire multiple

Ecriture du modèle

- $$Y = X\theta + \varepsilon,$$

soit pour $i = 1, \cdots, n$

$$Y_i = \theta_0 + X_{i,1}\theta_1 + \cdots + X_{i,d}\theta_d + \varepsilon_i.$$

Régression linéaire multiple

Eucalyptus

- Exemple 1 :
  - Y la hauteur d'un eucalyptus (quantitative).
  - hauteur en fonction de la circonférence d'un eucalyptus.
  - On se demande comment prédire la hauteur à partir de la circonférence (quantitative), de la racine carrée de la circonférence, et du carré de la circonférence. On note

$$X = \begin{pmatrix} 1 \\ circ \\ \sqrt{circ} \\ circ^2 \end{pmatrix}.$$

On considère le modèle

$$ht = \theta_0 + circ * \theta_1 + \sqrt{circ} * \theta_2 + \circ^2 * \theta_3 + \varepsilon,$$

Régression linéaire multiple

Eucalyptus (2)

$$\begin{pmatrix} Y_1 \\ . \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{1,1} & X_{1,2} & X_{1,3} \\ 1 & X_{2,1} & X_{2,2} & X_{2,3} \\ \vdots & \vdots & & \\ 1 & X_{n,1} & X_{n,2} & X_{n,3} \end{pmatrix} \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ . \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Régression linéaire multiple

Le modèle

$$Y_i = \theta_0 + \theta_1 X_{i,1} + \theta_2 X_{i,1} + ... + \theta_d X_{i,d} + \epsilon_i$$

Hypothèses :

- $E(\epsilon_i) = 0$
- $\text{Var}(\epsilon_i) = \sigma^2$ (constante)
- $\epsilon_i \perp\!\!\!\perp \epsilon_j; \forall i \neq j$
- $\epsilon_i \sim N(0, \sigma^2)$ (hypothèse nécessaire pour les tests)

Régression linéaire multiple

Le modèle - écriture matricielle

$$\mathbf{Y} = \mathbf{X}\theta + \mathbf{E}$$

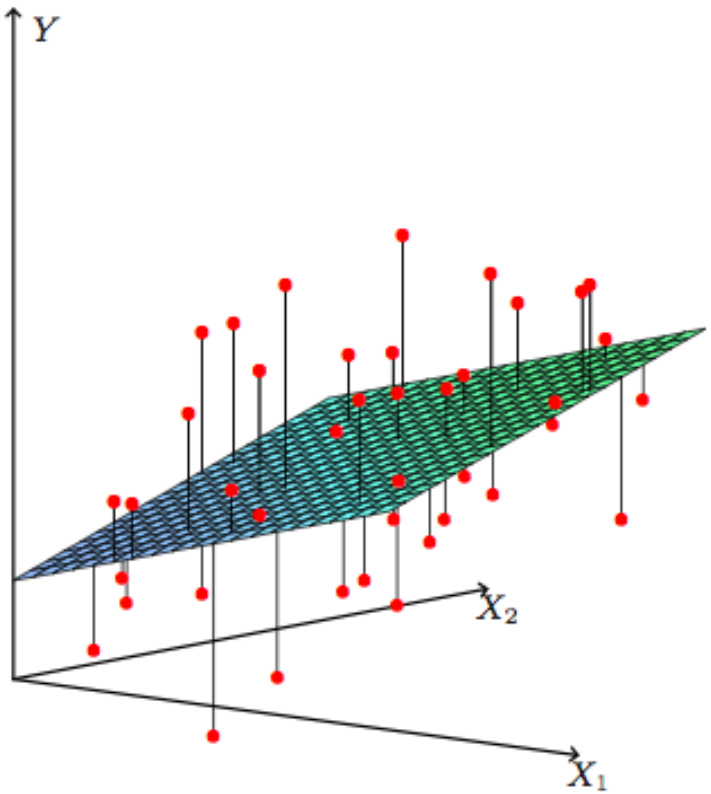
$$\begin{pmatrix} Y_1 \\ . \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1,d} \\ 1 & x_{2,1} & \dots & x_{2,d} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,d} \end{pmatrix} \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_d \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ . \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Hypothèses :

- $E(\mathbf{E}) = \mathbf{0}$
- $\text{Var}(\mathbf{E}) = \sigma^2 \mathbf{I_n}$
- $\mathbf{E} \sim N(0, \sigma^2 \mathbf{I_n})$ (pour les tests)

Estimation des paramètres

Estimateurs des moindres carrés



Hastie, Tibshirani, Friedman.The Elements of Statistical Learning. Springer, 2008

Estimation des paramètres

Estimateurs des moindres carrés

L'**estimateurs des moindres carrés ordinaires** (MCO) de β est le vecteur aléatoire de $\mathbb{R}^p$ $\hat{\beta}$ qui minimise la somme des carrés des résidus :

$$SCR(\theta) = \|\mathbf{Y} - \mathbf{X}\theta\|^2$$

Théorème

Si la matrice $\mathbf{X}^T\mathbf{X}$ est inversible, l'estimateur des moindres carrés ordinaires de θ est :

$$\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Estimation des paramètres

Propriétés

- L'estimateur $\widehat{\theta} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y}$ est un estimateur sans biais de θ
- La variance de $\hat{\theta}$ est : $\text{Var}(\widehat{\theta}) = \sigma^2(\mathbf{X^T X})^{-1}$
- σ^2 est estimé sans biais par :

$$\hat{\sigma}^2 = CMR = \frac{SCR(\beta)}{n-(d+1)} = \frac{\|\mathbf{Y} - \mathbf{X}\beta\|^2}{n-(d+1)}$$

avec

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-(d+1)}\chi^2_{n-(d+1)}$$

Comparaison de modèles

Notations

- Soit $\mathcal{M}_p$ un modèle à p paramètres parmi les d. On note $SCR(\mathcal{M}_p)$ la somme des carrés résiduels associée.
- Soit $\mathcal{M}_q$ un modèle à q paramètres emboîté dans $\mathcal{M}_p$ (donc $q < p$). On note $SCR(\mathcal{M}_q)$ est la somme des carrés résiduels associée.
- Soit $SCE = SCR(\mathcal{M}_q) - SCR(\mathcal{M}_p)$ la partie de la SCR expliquée par le passage du petit modèle $\mathcal{M}_q$ au grand modèle $\mathcal{M}_p$.
- Soit n le nombre total de mesures.

Comparaison de modèles emboîtés

Théorème

Si les résidus ϵ_i sont indépendants de loi $N(0, \sigma^2)$, alors sous $\mathcal{M}_q$:

- $SCR(\mathcal{M}_q) \sim \sigma^2 \chi^2_{n-q}$
- $SCR(\mathcal{M}_p) \sim \sigma^2 \chi^2_{n-p}$
- $SCE \sim \sigma^2 \chi^2_{p-q}$
- $SCE \perp\!\!\!\perp SCR(\mathcal{M}_p)$

Comparaison de modèles emboîtés

Tableau de synthèse

Source	Degrés de liberté	Sommes des carrés	Carrés moyens	F
Gain $\mathcal{M}_p/\mathcal{M}_q$	$p-q$	SCE	$CME = \frac{SCE}{p-q}$	$F = \frac{CME}{CMR}$
$\mathcal{M}_p$	$n-p$	$SCR(H_p)$	$CMR = \frac{SCR(\mathcal{M}_p)}{(n-p)}$	
$\mathcal{M}_q$	$n-q$	$SCR(\mathcal{M}_q)$		

Comparaison de modèles emboîtés

Vers la sélection de variables

L'idée est que l'on cherche à sélectionner les variables pertinentes. On va chercher le modèle qui a le moins de variables possibles et qui explique le mieux les données. Deux pistes

- Méthodes basées sur des tests de comparaison de modèles emboités
- Méthodes basées sur des critères de sélection de variables

Le modèle linéaire multiple

Exemple :Prostate Cancer

The data for this example come from a study by Stamey et al. (19 89). They examined the correlation between the level of prostate-specific antigen and a number of clinical measures in men who were about to receive a radicalprostatectomy. The variables are log cancer volume (*lcavol*), log prostate weight (*lweight*), *age* , log of the amount of benign prostatic hyperplasia (*lbph*), seminal vesicle invasion (*svi*), log of capsular penetration (*lcp*), Gleason score (*gleason*), and percent of Gleason scores 4 or 5 (*pgg*45).

$$Y = \text{level of prostate-specific antigen},$$

and

$$X = (lcavol, lweight, age, lbph, svi, lcp, gleason, pgg45)^T.$$

Comparaison des espérances de deux échantillons gaussiens

Test de Student

- Présupposés :
  $Y_{11}, ..., Y_{1n_1}$ iid avec $Y_{1i} \sim N(\mu_1, \sigma_1^2)$
  $Y_{21}, ..., Y_{2n_2}$ iid avec $Y_{2i} \sim N(\mu_2, \sigma_2^2)$
  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ inconnu.
- Hypothèse nulle : $H_0 : \mu_1 = \mu_2$
- Statistique de test :

$$\frac{\overline{Y_1} - \overline{Y_2}}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}} \underset{H_0}{\sim} \mathcal{S}t(n_1 + n_2 - 2)$$

avec

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(Y_{1,i} - \overline{Y_1})^2 \text{ et } S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(Y_{2,i} - \overline{Y_1})^2.$$

Comparaison des espérances de deux échantillons gaussiens

Vers l'analyse de la variance à 1 facteur :

Modèle

- Observations : Y_{ij} $i = 1, 2$, $j = 1, ..., n_i$
- Modèle : Y_{ij} indépendantes avec

$$Y_{ij} \sim N(\mu_i, \sigma^2)$$

- i représente le groupe
- j représente l'individu j dans le groupe i

L'hypothèse d'homogénéité des variances ($\sigma_1^2 = \sigma_2^2 = \sigma^2$) est appelée hypothèse d'**homoscédasticité**. Elle est cruciale !

Comparaison des espérances de 2 échantillons

Autre formulation

- Observations : Y_{ij} $i = 1, ..., k$, $j = 1, ..., n_i$
- Modèle : Y_{ij} indépendantes avec

$$Y_{1j} = \mu_1 + \epsilon_{1j}, \quad Y_{2j} = \mu_2 + \epsilon_{2j}$$

avec

$$\epsilon_{1j} \sim N(0, \sigma^2) \text{ et } \epsilon_{2j} \sim N(0, \sigma^2).$$

Remarque

- La distribution des ϵ_{ij} ne dépend pas du groupe
- μ s'interprète comme l'effet 'global', α_i comme l'effet spécifique du groupe i et ϵ_{ij} comme le résidu (bruit gaussien)

Comparaison des espérances de 2 échantillons

Formulation en terme de comparaison de modèles

Notons $\mathcal{M}_1$ le modèle tel que Y_{ij} indépendantes avec

$$Y_{ij} = \mu + \epsilon_{ij}$$

avec

$$\epsilon_{ij} \sim N(0, \sigma^2).$$

Dans ce modèle, pas d'effet du groupe sur l'espérance de Y.
Notons $\mathcal{M}_2$ le modèle tel que Y_{ij} indépendantes où

$$Y_{1j} = \mu_1 + \epsilon_{1j}, \quad Y_{2j} = \mu_2 + \epsilon_{2j}$$

avec

$$\epsilon_{1j} \sim N(0, \sigma^2) \text{ et } \epsilon_{2j} \sim N(0, \sigma^2).$$

Dans ce modèle, il y a un effet du groupe sur l'espérance de Y.

Comparaison des espérances de 2 échantillons)

Formulation en terme de comparaison de modèles

Le test revient à tester

- Hypothèse nulle : $H_0 : \mu_1 = \mu_2 = \mu$ cad le bon modèle est $\mathcal{M}_1$
- Hypothèse alternative : $H_1 : \mu_1 \neq \mu_2$ cad le bon modèle est $\mathcal{M}_2$

On compare donc les modèles $\mathcal{M}_1$ et $\mathcal{M}_2$.

Comparaison des variances de deux échantillons gaussiens

- Présupposés :
  $Y_{11}, ..., Y_{1n_1}$ iid avec $Y_{1i} \sim N(\mu_1, \sigma_1^2)$
  $Y_{21}, ..., Y_{2n_2}$ iid avec $Y_{2i} \sim N(\mu_2, \sigma_2^2)$
- Hypothèse nulle : $H_0 : \sigma_1^2 = \sigma_2^2$
- Statistique de test :

$$\frac{S_1^2}{S_2^2} \underset{H_0}{\sim} \mathcal{F}(n_1 - 1, n_2 - 1)$$

Fonction R

var.test

Exemple des forêts

Exemple des forêts :

```
data=read.table('foret3.txt',header=TRUE)
attach(data,warn.conflicts=FALSE)
```

Notations exemple forêts

- $p = 3$, n_i : le nombre d'arbres dans la forêt i avec $n = n_1 + n_2 + n_3 = 14 + 13 + 10$
- $Y_{i,j}$ hauteur de l'arbre j dans la forêt i, $1 \leq i \leq p$
- μ_i l'espérance de hauteur dans la forêt i, $1 \leq i \leq p$

$$Y_{i,j} = \mu_i + \varepsilon_{i,j} \quad \varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2).$$

- $\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$: hauteur moyenne dans la forêt i
- $\bar{Y} = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n_i} Y_{ij}$ hauteur moyenne dans les forêts

Summary

Forêt 1 :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
23.40	24.90	26.30	25.99	26.90	27.90

Forêt 2 :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
22.50	24.10	25.65	25.39	26.62	28.50

Forêt 3 :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18.90	21.43	23.00	23.13	24.57	26.70

Comparaison des espérances de p échantillons ($p > 2$)

Autre exemple

- On souhaite comparer trois traitements contre l'asthme : le traitement 2 est un nouveau traitement, que l'on souhaite mettre en compétition avec les traitements classiques 1 et 3.
- On répartit par tirage au sort les patients venant consulter dans un centre de soin et on leur affecte l'un des trois traitements. On mesure sur chaque patient la durée, en jours séparant de la prochaine crise d'asthme.

Traitement 1	Traitement 2	Traitement 3
26 ; 27 ; 35 ; 36 ; 38	29 ; 42 ; 44 ; 44 ; 45	26 ; 26 ; 30 ; 30 ; 33
38 ; 41 ; 42 ; 45 ; 50	48 ; 48 ; 52 ; 56 ; 56	36 ; 38 ; 38 ; 39 ; 46
65	58 ; 58 ; 60 ; 61 ; 63	47 ; 51 ; 51 ; 56 ; 75
	63 ; 69	

Comparaison des espérances de p échantillons ($p > 2$)

Comparaison 2 à 2 : problème des tests multiples

- Lors du test d'une seule hypothèse H_0 (au niveau α), le risque de commettre une erreur de type I est :

$$\mathbb{P}(P < \alpha | H_0) \leq \alpha$$

- Lors du test de r hypothèses $H_0^1; ...; H_0^r$ (au niveau α pour chacun des tests), le risque de commettre (au moins) une erreur de type I est :

$$\mathbb{P}(\bigcup_{1 \leq i \leq r} \{P_i < \alpha | H_0^i\}) \leq \sum_{i=1}^{r} \mathbb{P}(P_i < \alpha | H_0^i) = r\alpha$$

Conséquence : Pour assurer un niveau global de α en faisant simultanément r tests, une solution est de mettre en oeuvre chacun des r tests individuellement au niveau α/r. C'est la correction Bonferroni.

Comparaison des espérances de p échantillons ($p > 2$)

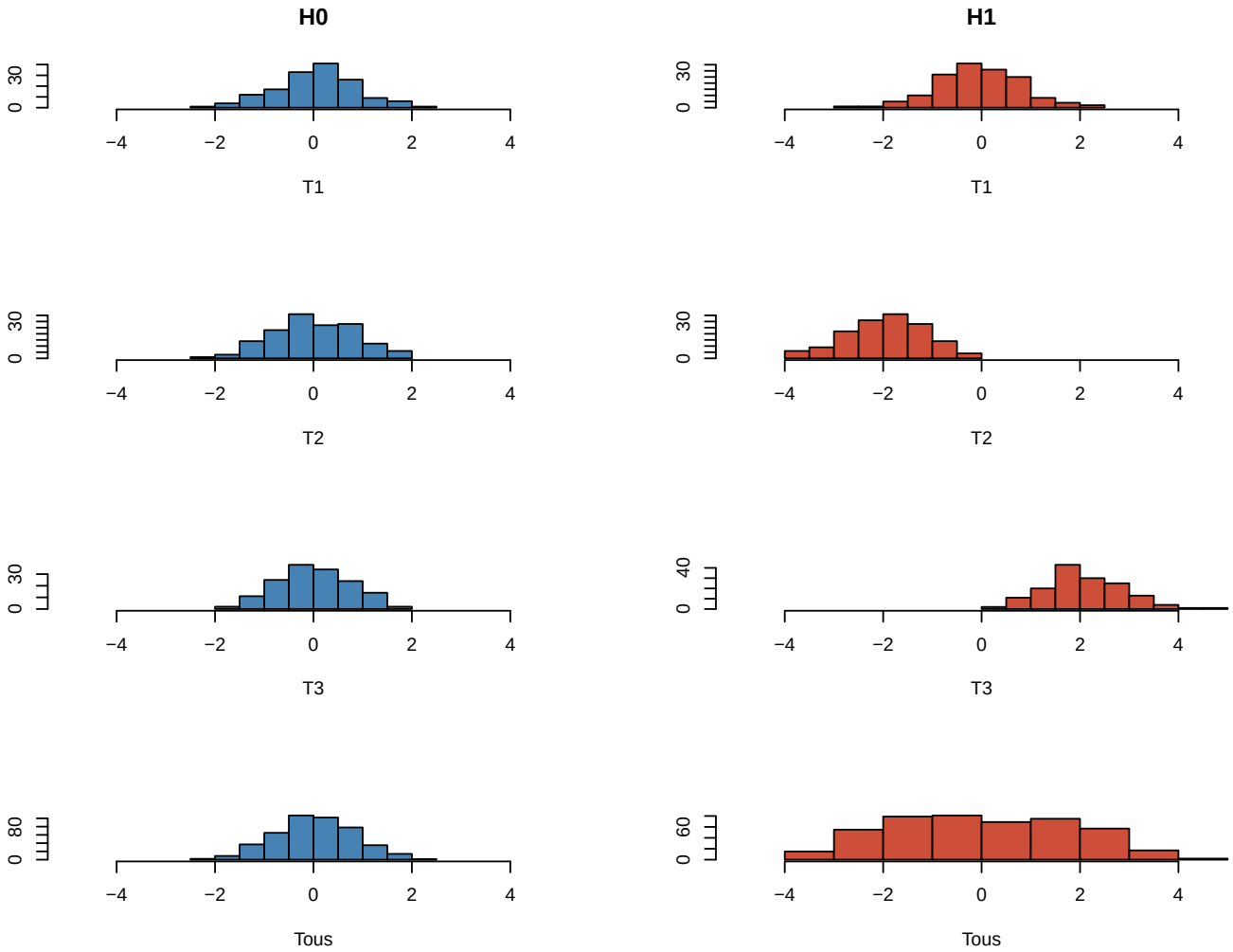
Test global

Pour comparer p espérances, au lieu de comparer toutes les espérances deux à deux (ce qui conduit à un problème de tests multiples important), on effectue un test global des hypothèses nulles et alternatives suivantes :

$$H_0 : \mu_1 = ... = \mu_p$$
$$H_1 : \exists (i,j) | \mu_i \neq \mu_j$$

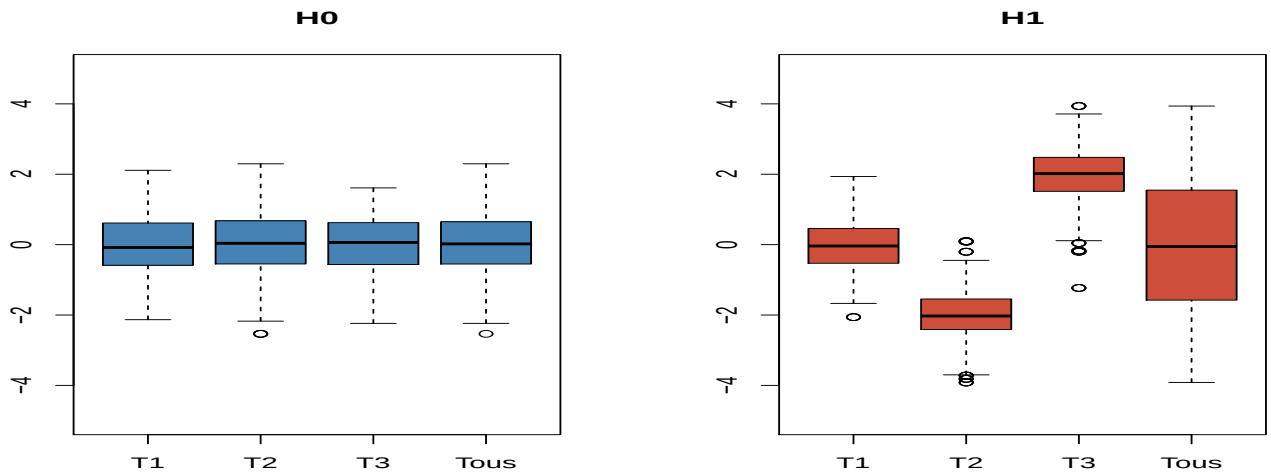
Comparaison des espérances de p échantillons ($p > 2$)

Idée générale



Comparaison des espérances de p échantillons ($p > 2$)

Idée générale



1. Trouver un estimateur de la variance commune σ^2 valable sous H_0 et sous H_1
2. Trouver un estimateur de la variance commune σ^2 valable uniquement sous H_0
3. Comparer les deux variances estimées

Comparaison des espérances de p échantillons ($p > 2$)

Modèle

- Observations : Y_{ij} $i = 1, ..., p$, $j = 1, ..., n_i$
- Modèle : Y_{ij} indépendantes avec

$$Y_{ij} \sim N(\mu_i, \sigma^2)$$

Remarque

L'hypothèse d'homogénéité des variances ($\sigma_1^2 = ... = \sigma_p^2 = \sigma^2$) est appelée hypothèse d'**homoscédasticité**.

Comparaison des espérances de p échantillons ($p > 2$)

Autre formulation

- Observations : Y_{ij} $i = 1, ..., p$, $j = 1, ..., n_i$
- Modèle : Y_{ij} indépendantes avec

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

$$\sum_{i=1}^{p} \alpha_i = 0$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

Remarque

- La distribution des ϵ_{ij} ne dépend pas du groupe
- μ s'interprète comme l'effet 'global', α_i comme l'effet spécifique du groupe i et ϵ_{ij} comme le résidu (bruit gaussien)

Comparaison des espérances de p échantillons ($p > 2$)

Formulation comme une comparaison de modèles

Dans l'ANOVA1, on compare les deux modèles suivants :

- $\mathcal{M}_1$, le modèle sous H_0 : $Y = \mu + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2)$ (1 paramètre)
- $\mathcal{M}_p$, le modèle sous H_1 : $Y = \mu_i + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2)$ (p paramètres)

où $\mathcal{M}_1$ est emboîté dans $\mathcal{M}_p$

Comparaison de modèles emboîtés

Ligne associée à un modèle

A chaque modèle peut correspondre une ligne d'un tableau de synthèse :

modèle $\mathcal{M}$	ddl	$SCR(\mathcal{M})$	$CM(\mathcal{M})$

Remarque

Tous les tests de choix de modèle dans le contexte du modèle linéaire reviennent à comparer deux telles lignes.

Comparaison des espérances de p échantillons ($p > 2$)

Sommes des carrés

Résiduels :	$SCR(i) = \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$	avec $SCR(i) \sim \sigma^2 \chi^2_{n_i - 1}$
	$SCR(\mathcal{M}_p) = \sum_{i=1}^{p} SCR(i)$	avec $SCR(\mathcal{M}_p) \sim \sigma^2 \chi^2_{n-p}$
Totaux :	$SCT = SCR(\mathcal{M}_1) = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$	avec $SCT \underset{H_0}{\sim} \sigma^2 \chi^2_{n-1}$
Factoriels :	$SCF = \sum_{i=1}^{p} n_i (\bar{Y}_i - \bar{Y})^2$	avec $SCF \underset{H_0}{\sim} \sigma^2 \chi^2_{p-1}$

Comparaison des espérances de p échantillons ($p > 2$)

Carrés moyens

Résiduel : $CMR = \dfrac{SCR(\mathcal{M}_p)}{n-p}$ estimateur sans biais de σ^2

Total : $CMT = \dfrac{SCT}{n-1} = \dfrac{SCR(\mathcal{M}_1)}{n-1}$ estimateur sans biais (sous H_0) de σ^2

Factoriel : $CMF = \dfrac{SCF}{p-1} = \dfrac{(SCR(\mathcal{M}_1) - SCR(\mathcal{M}_p)}{(p-1)}$ estimateur sans biais (sous H_0) de σ^2

Comparaison des espérances de p échantillons ($p > 2$)

Théorème fondamental (décomposition de la variance)

En utilisant les notations introduites précédemment, on a :

1. $$SCT = SCF + SCR \Longleftrightarrow SCT = SCR(\mathcal{M}_1) = SCF + SCR(\mathcal{M}_p)$$
2. avec
$$SCF \perp\!\!\!\perp SCR(\mathcal{M}_p)(orthogonal)$$

Comparaison des espérances de p échantillons ($p > 2$)

Statistique de test

$$F = \frac{SCF/(p-1)}{SCR(\mathcal{M}_p)/(n-p)} = \frac{CMF}{CMR} \underset{H_0}{\sim} \mathcal{F}(p-1, n-p)$$

Comparaison des espérances de p échantillons ($p > 2$)

Résultats ($\mathcal{M}_p$ *vs* $\mathcal{M}_1$) : table d'analyse de la variance

Source	Degrés de liberté	Sommes des carrés	Carrés moyens	F
Facteur	$p-1$	SCF	$CMF = SCF/(p-1)$	CMF/CMR
Résidus	$n-p$	$SCR(\mathcal{M}_p)$	$CMR = SCR(\mathcal{M}_p)/(n-p)$	
Total	$n-1$	$SCT = SCR(\mathcal{M}_1)$		

$$SCF = SCR(\mathcal{M}_1) - SCR(\mathcal{M}_p).$$

Comparaison des espérances de p échantillons ($p > 2$)

Exemple de l'asthme

Source	Degrés de liberté	Sommes des carrés	Carrés moyens	F
Facteur	2	1426.84	713.42	5.467
Résidus	40	5219.44	130.49	
Total	42	6646.28		

Test de Student pour deux groupes : un cas particulier de l'ANOVA

Equivalence entre test de Student et ANOVA

- Test de Student :

$$\Gamma = \left\{ T = \frac{|\bar{Y}_1 - \bar{Y}_2|}{\sqrt{CMR(\frac{1}{n_1} + \frac{1}{n_2})}} > q_{\mathcal{S}t(n_1+n_2-2);1-\alpha/2} \right\}.$$

- ANOVA :

$$\Gamma = \left\{ F = \frac{CMF}{CMR} > q_{\{\mathcal{F}_{1;n_1+n_2-2};1-\alpha\}} \right\}.$$

Remarque

$T^2 = F \Rightarrow$ Les deux régions de rejets précédentes sont identiques

Test basé sur $T \Rightarrow$ permet de faire des tests unilatères

Comparaisons partielles

Hypothèses testées

$$\begin{cases} H_0 : \mu_q = \mu_l, \ \forall \ q, l \in P \\ H_1 : \overline{H_0}(\text{complémentaire}) \end{cases}$$

où P est l'ensemble des p niveaux que l'on suppose identiques.

Remarque

Ce test est une analyse de la variance, mais réalisée sur un sous ensemble des observations disponibles.

Comparaisons partielles

SCF partielle

$$SCFP = \frac{1}{\sum_{q\in P} n_q} \sum_{q\neq l\in P} n_q n_l (\bar{Y}_q - \bar{Y}_l)^2 \underset{H_0}{\sim} \sigma^2 \chi^2_{p-1}$$

Statistique de test

$$Fpart = \frac{SCFP/(p-1)}{SCR/(n-p)} \underset{H_0}{\sim} \mathcal{F}_{p-1;n-p}$$

Région de rejet

$$\Gamma = \{Fpart > q_{\mathcal{F}_{p-1;n-p};1-\alpha}\}.$$

Comparaison de modèles emboîtés

Modèles emboîtés

Deux modèles sont dits **emboîtés** si l'un peut être vu comme un cas particulier de l'autre.

Exemple

Dans l'ANOVA1, on compare les deux modèles suivants :

- $\mathcal{M}_1$, le modèle sous H_0 : $Y = \mu + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2)$ (1 paramètre)
- $\mathcal{M}_p$, le modèle sous H_1 : $Y = \mu_i + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2)$ (p paramètres)

où $\mathcal{M}_1$ est emboîté dans $\mathcal{M}_p$

Comparaison de modèles emboîtés

Ligne associée à un modèle

A chaque modèle peut correspondre une ligne d'un tableau de synthèse :

modèle $\mathcal{M}$	ddl	$SCR(\mathcal{M})$	$CM(\mathcal{M})$

Remarque

Tous les tests de choix de modèle dans le contexte du modèle linéaire reviennent à comparer deux telles lignes.

Comparaison de modèles emboîtés

Tableau de synthèse

Modèle $\mathcal{M}_1$ emboîté modèle $\mathcal{M}_2$ (i.e. $SCR(\mathcal{M}_1) > SCR(\mathcal{M}_2)$) :

Source	Degrés de liberté	Sommes des carrés	Carrés moyens	F
$\mathcal{M}_2/\mathcal{M}_1$	$n_1 - n_2$	$SC = SCR(\mathcal{M}_1) - SCR(\mathcal{M}_2)$	$CM = SC/(n_1 - n_2)$	CM/CM_2
$\mathcal{M}_2$	n_2	$SCR(\mathcal{M}_2)$	$CM_2 = SCR(\mathcal{M}_2)/(n_2)$	
$\mathcal{M}_1$	n_1	$SCR(\mathcal{M}_1)$	$CM_1 = SCR(\mathcal{M}_1)/(n_1)$	

Question

Une augmentation du nombre de paramètres du modèle (*i.e.* de sa complexité) fait-elle significativement diminuer la variance estimée ?

Comparaison de modèles emboîtés

Cas de l'ANOVA1

On compare le modèle à 1 paramètre $\mathcal{M}_1$ au modèle à p paramètres $\mathcal{M}_p$ (avec $\mathcal{M}_1$ emboîté dans $\mathcal{M}_p$.
Par différence on obtient une ligne mesurant le gain apporté par $\mathcal{M}_p$ en comparaison de $\mathcal{M}_1$:

Gain de $\mathcal{M}_p/\mathcal{M}_1$	$ddl_1 - ddl_p$	$SCR(\mathcal{M}_1) - SCR(\mathcal{M}_p)$	$\frac{SCR(\mathcal{M}_1)-SCR(\mathcal{M}_p)}{ddl_1-ddl_p}$
modèle $\mathcal{M}_p$	ddl_p	$SCR(\mathcal{M}_p)$	$CM(\mathcal{M}_p)$
modèle $\mathcal{M}_1$	ddl_1	$SCR(\mathcal{M}_1)$	$CM(\mathcal{M}_1)$

Remarque

La ligne 'gain' correspond à la ligne associée au facteur dans le tableau de l'analyse de la variance.

Ligne associée à un modèle

Exemple

Dans le cas des données relatives aux traitements de l'asthme, la vraie question était : 'le nouveau traitement 2 est-il supérieur aux deux traitements classiques 1 et 3 ?'

Validité du modèle

Normalité

- Test de Stephens
- Test de Shapiro-Wilk

Homoscédasticité

- Test de Bartlett
- Test de Levene