Statistique élémentaire avec **R**
Partie 2 : Test d'hypothèses et régression linéaire

Julien JACQUES

Université Lumière Lyon 2

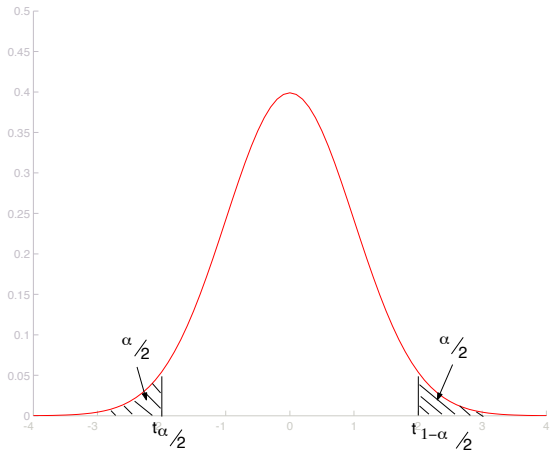
Plan

Plan

Principe d'un test statistique

Un exemple

1. Test $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$
2. Stat. de test $T = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim_{H_0} t_{n-1}$ Student à n-1 degrés de liberté
3. $\alpha = 5\%$
4. Zone de rejet $W = \{\bar{x} : |t| = \frac{|\bar{x} - \mu_0|}{\frac{s}{\sqrt{n}}} > -t_{n-1, \frac{\alpha}{2}}\}$



5. *calcul de t puis acceptation de H_0 si t est entre les bornes, rejet sinon*

Principe d'un test statistique

Les étapes

1. Identifier des hypothèses H_0 (hyp. nulle, simple) et H_1 (hyp. alternative, composite)
2. Définir un statistique de test T, dont la loi est différente sous H_0 et H_1
3. Choisir un risque de première espèce α (5%, 10%...)
4. Définir la zone de rejet W de H_0, en fonction de H_1 (test uni- ou bilatéral) et de α
5. Calculer la valeur t de la statistique de test T
6. Conclure au rejet de H_0 si $t \in W$ où à son acceptation dans le cas contraire

Principe d'un test statistique

Les risques antagonistes

Décision \ Vérité	H_0	H_1
H_0	conclusion correcte	erreur de deuxième espèce
H_1	erreur de première espèce	conclusion correcte

Table : Erreurs associés à un test

Décision \ Vérité	H_0	H_1
H_0	niveau de confiance $1-\alpha$	risque β
H_1	risque α	$1-\beta$

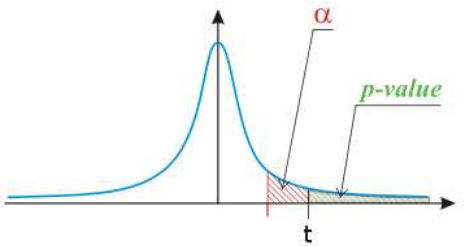
Table : Risques associés à un test

Principe d'un test statistique

La p-value p^*

- plus petite valeur de α conduisant à rejeter H_0
- probabilité sous H_0 d'observer une statistique de test aussi extrême (au sens de H_1) que le t observé
- probabilité de se tromper lorsqu'on rejette H_0

Exemple : test unilatéral $H_0 : \mu = 0$ contre $H_1 : \mu > 0$



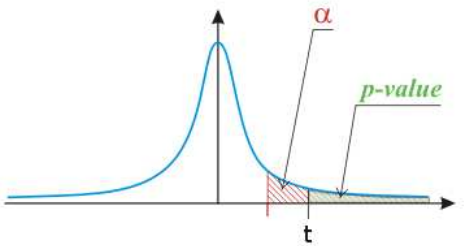
$p^* = P(T > t)$ où T stat. de test et t sa valeur sur l'échantillon

Principe d'un test statistique

La p-value p^*

- plus petite valeur de α conduisant à rejeter H_0
- probabilité sous H_0 d'observer une statistique de test aussi extrême (au sens de H_1) que le t observé
- probabilité de se tromper lorsqu'on rejette H_0

Exemple : test unilatéral $H_0 : \mu = 0$ contre $H_1 : \mu > 0$



$p^* = P(T > t)$ où T stat. de test et t sa valeur sur l'échantillon

Utilisation de la p-value p^*

- si $\alpha > p^*$: rejet de H_0
- si $\alpha < p^*$

Plan

Typologie des tests

Tests de liaison entre variables

- Tester la liaison entre deux variables quantitatives : Test de corrélation
- Tester la liaison entre deux variables qualitatives : Test d'indépendance du χ^2
- Tester la liaison entre une variable quantitative et une variable qualitative : ANOVA à 1 facteur
- Tester la liaison entre une variable quantitative et K variables qualitatives : ANOVA à K facteur

Tests de comparaison de populations indépendantes

- Test de comparaisons des variances de Fisher
- Test de comparaisons des moyennes de Student

Typologie des tests - Logiciel R

Tests de liaison entre variables

- Tester la liaison entre deux variables quantitatives : fonction `cor.test`
- Tester la liaison entre deux variables qualitatives : fonction `chisq.test`
- Tester la liaison entre une variable quantitative et une variable qualitative : fonction `aov`
- Tester la liaison entre une variable quantitative et K variables qualitatives : fonction `aov`

Tests de comparaison de populations indépendantes

- Test de comparaisons des variances de Fisher : fonction `var.test`
- Test de comparaisons des moyennes de Student : fonction `t.test`

Plan

Test de corrélation

Conditions d'application :

- X et Y deux variables aléatoires quantitatives

Hypothèses

$H_0 : \rho_{X,Y} = 0$ contre $H_1 : \rho_{X,Y} \neq 0$

Statistique de test

$T = \sqrt{n-2}\frac{R_{XY}}{\sqrt{1-R_{XY}^2}} \sim_{H_0} t_{n-2}$ où $R_{XY} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2 \sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$ est l'estimateur du coefficient de corrélation

Décision

on rejette H_0 si

$t > t_{n-2,1-\frac{\alpha}{2}}$ ou $t < t_{n-2,\frac{\alpha}{2}}$

Test d'indépendance du χ^2

Conditions d'application :

- X et Y deux variables aléatoires qualitatives à k et r modalités
  n_{ij} : nombre d'observations ayant la modalité i de X et j de Y
  $n_{i.} = \sum_{j=1}^{r} n_{ij}$ et $n_{.j} = \sum_{i=1}^{k} n_{ij}$
- $n_{ij} \geq 5$

Hypothèses

H_0 : X et Y indépendantes contre H_1 : X et Y dépendantes

Statistique de test

$$d^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(n_{ij} - \frac{n_{i.}\, n_{.j}}{n})^2}{\frac{n_{i.}\, n_{.j}}{n}} \sim_{H_0} \chi^2_{(k-1)(r-1)}$$

Décision

on rejette H_0 si
$d^2 > \chi^2_{(k-1)(r-1)1-\alpha}$

ANOVA à 1 facteur

Conditions d'application :

- X une variable quantitative, A un facteur qualitatif à K modalités
- échantillons grands ($n \geq 30$) ou gaussiens (pour chaque modalité)
- variances homogènes

Hypothèses

A influe-t-il X ?

$H_0 : \mu_1 = \ldots = \mu_K = \mu$ contre $H_1 : \exists 1 \leq i, j \leq K$ t.q. $\mu_i \neq \mu_j$

Statistique de test

$F = \frac{V_A^2}{K-1} / \frac{V_R^2}{n-K}$ où

- $V_A^2 = \frac{1}{n} \sum_{k=1}^{K} n_k (\bar{X}_k - \bar{X})^2$ est la variance expliquée par le facteur A
- V_R^2 est la variance résiduelle
- avec variance totale $V_T^2 = V_A^2 + V_R^2$

ANOVA à 1 facteur

Présentation des résultats

Facteur	Somme des carrés	degrés de liberté	carré moyen	F
A	SSA	$K-1$	$SSA/(K-1)$	$F = \frac{SSA/(K-1)}{SSR/(n-K)}$
Résidu	SSR	$n-K$	$SSR/(n-K)$	
Total	SST	$n-1$		

ou $SSA = nV_A^2$, $SSR = nV_R^2$ et $SST = nV_T^2$.

Décision

On conclue à un effet de A (rejet de H_0) si $F > F_{K-1,n-K,1-\alpha}$

ANOVA à 2 facteur

Conditions d'application :

- X une variable quantitative, A et B deux facteurs qualitatifs à J et K modalités
- échantillons grands ($n \geq 30$) ou gaussiens (pour chaque croisement de modalités)
- variances homogènes

Hypothèses

- Le facteur A a-t-il une influence sur X ?
- Le facteur B ?
- Et l'interaction entre les deux facteurs ?

ANOVA à 2 facteur

Décomposition de la variance totale

$$SST = SSA + SSB + SSAB + SSR$$

avec

$$SST = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(X_{ijk} - \bar{X}_{...})^2, \quad SSA = \sum_{j=1}^{J} n_{j.}(\bar{X}_{.j.} - \bar{X}_{...})^2, \quad SSB = \sum_{k=1}^{K} n_{.k}(\bar{X}_{..k} - \bar{X}_{...})^2,$$

$$SSAB = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk}(\bar{X}_{.jk} - \bar{X}_{.j.} - \bar{X}_{..k} + \bar{X}_{...})^2, \quad \text{et} \quad SSR = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(X_{ijk} - \bar{X}_{.jk})^2$$

où

$$\bar{X}_{.jk} = \frac{1}{n_{jk}}\sum_{i=1}^{n_{jk}} X_{ijk}, \quad \bar{X}_{..k} = \frac{1}{n_{.k}}\sum_{j=1}^{J} \bar{X}_{.jk}, \quad \bar{X}_{.j.} = \frac{1}{n_{j.}}\sum_{k=1}^{K} \bar{X}_{.jk} \quad \text{et} \quad \bar{X}_{...} = \frac{1}{n}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}} X_{ijk}.$$

ANOVA à 2 facteur

Présentation des résultats

Facteur	Somme des carrés	degrés de liberté	carré moyen	F
A	SSA	$J-1$	$SSA/(J-1)$	$F_A = \frac{SSA/(J-1)}{SSR/(n-JK)}$
B	SSB	$K-1$	$SSB/(K-1)$	$F_B = \frac{SSB/(K-1)}{SSR/(n-JK)}$
Interaction AB	$SSAB$	$(J-1)(K-1)$	$SSAB/(K-1)(J-1)$	$F_{AB} = \frac{SSAB/(K-1)(J-1)}{SSR/(n-JK)}$
Résidu	SSR	$n-JK$	$SSR/(n-JK)$	
Total	SST	$n-1$		

Décision

- On conclue à un effet de A si $F_A > F_{J-1,n-JK,1-\alpha}$
- On conclue à un effet de B si $F_B > F_{K-1,n-JK,1-\alpha}$
- On conclue à un effet de l'interaction entre A et B si $F_{AB} > F_{(K-1)(J-1),n-JK,1-\alpha}$

Plan

Test de comparaisons des variances de Fisher

Conditions d'application :

- échantillons gaussiens

Hypothèses

$H_0 : \sigma_1 = \sigma_2$ contre $H_1 : \sigma_1 \neq \sigma_2$

Statistique de test

$$F = \frac{\frac{n_1 V_1^2}{n_1 - 1}}{\frac{n_2 V_2^2}{n_2 - 1}} = \frac{S_1^2}{S_2^2} \sim_{H_0} F_{n_1-1, n_2-1} \qquad \text{avec } S_1^2 > S_2^2$$

Décision

on rejette H_0 si $\frac{S_1^2}{S_2^2} > f_{n_1-1, n_2-1, 1-\alpha}$

Test de comparaisons des moyennes de Student

Conditions d'application :

- échantillons grands ($n \geq 30$) ou gaussiens
- variances égales : $\sigma_1^2 = \sigma_2^2$

Hypothèses

$H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$

Statistique de test

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{n_1 V_1^2 + n_2 V_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim_{H_0} t_{n_1+n_2-2}$$

Décision

on rejette H_0 si

$$|\bar{x}_1 - \bar{x}_2| > -t_{n_1+n_2-2,\frac{\alpha}{2}} \sqrt{\frac{n_1 v_1^2 + n_2 v_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}.$$

Test de comparaisons des moyennes de Student

Conditions d'application :

- échantillons grands ($n \geq 30$) ou gaussien
- variances différentes : $\sigma_1^2 \neq \sigma_2^2$

Hypothèses

$H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$

Correction d'Aspin Welch

il faut remplacer le nombre de degrés de liberté de la loi de Student ($n_1 + n_2 - 2$ lorsque les variances sont égales) par l'entier le plus proche de :

$$n = \frac{1}{\frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1}} \quad \text{où } c = \frac{\frac{v_1^2}{n_1 - 1}}{\frac{v_1^2}{n_1 - 1} + \frac{v_2^2}{n_2 - 1}}$$

Test de comparaisons des moyennes de Student - cas apparié

Conditions d'application :

- échantillons grands ($n \geq 30$) ou gaussiens
- échantillons dépendants (appariés) : chaque échantillon correspond à des mesures différentes des mêmes individus

Test

on travaille sur la différence $D_i = X_{1i} - X_{2i}$ entre les 2 échantillons, et on test la nullité de la moyenne des D_i :
$H_0 : \mu = 0$ contre $H_1 : \mu \neq 0$

Test de comparaisons des moyennes de Student - cas unilatéral

Conditions d'application :

- échantillons grands ($n \geq 30$) ou gaussiens
- variances égales : $\sigma_1^2 = \sigma_2^2$ (sinon correction Aspin-Welch)

Hypothèses

$H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 > \mu_2$

Statistique de test

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1 V_1^2 + n_2 V_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim_{H_0} t_{n_1+n_2-2}$$

Décision

on rejette H_0 si $\quad \bar{x}_1 > \bar{x}_2 - t_{n_1+n_2-2,\frac{\alpha}{2}} \sqrt{\frac{n_1 v_1^2 + n_2 v_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$

Plan

Modélisation statistique

Les différents types de modélisation

Variable à expliquer	Variables explicatives	Nom de l'analyse
1 quanti.	1 quanti.	régression simple
1 quanti.	plusieurs quanti.	régression multiple
1 quanti.	plusieurs quali.	analyse de variance
1 quanti.	plusieurs quali. et quanti.	analyse de covariance

Objectifs

- prédictifs
- descriptifs : sélection des variables pertinentes, forme du modèle

Les étapes

- identifier le problème $\rightarrow$ choix du modèle statistique
- estimer les paramètres
- évaluer la qualité de la modélisation obtenue
- utiliser le modèle pour répondre à la question posée

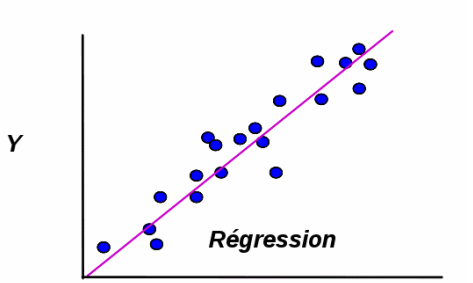
Plan

Le modèle de régression linéaire simple

Les données

Un échantillon $(X_i Y_i)_{i=1,n}$

- variable à prédire : Y
- variable explicative : X



si la liaison entre X et Y n'est pas linéaire, tester des transformations (log, puissance...)

Le modèle

$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ où $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d

Écriture matricielle :

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

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

Le modèle de régression linéaire simple

Estimation des paramètres

On cherche $\beta = (\beta_0, \beta_1)$ minimisant l'écart entre les valeurs prédites $\hat{Y}_i = \beta_0 + X_i\beta_1$ et les valeurs observées Y_i :

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

Les solutions sont

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \qquad \hat{\beta}_1 = \frac{S_{XY}}{S_X^2}.$$

où $S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$ est l'estimateur de la covariance de X et Y.

Plan

Le modèle de régression linéaire multiple

Les données

Un échantillon $(X_{i1}, \ldots, X_{ip}, Y_i)_{i=1,n}$

- variable à prédire : Y
- p variables explicatives : $X_1, \ldots, X_p$

Le modèle

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i$$

où $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \ldots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad (1)$$

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \quad (2)$$

Le modèle de régression linéaire multiple

Estimation des paramètres

On cherche $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ minimisant l'écart entre les valeurs prédites $\hat{Y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$ et les valeurs observées Y_i :

$$\min \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij})^2$$

La solution est

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

Plan

Normalité des résidus

Dans le but de faire des tests sur le modèle de régression obtenus, nous avons fait l'hypothèse de normalité des résidus $\epsilon_i = \hat{y}_i - y_i$.

Test de normalité

Il existe des tests statistiques permettant de tester l'adéquation d'une série de données (ici les résidus) à une loi normale :

- test de Shapiro-Wilk: fonction `shapiro.test`

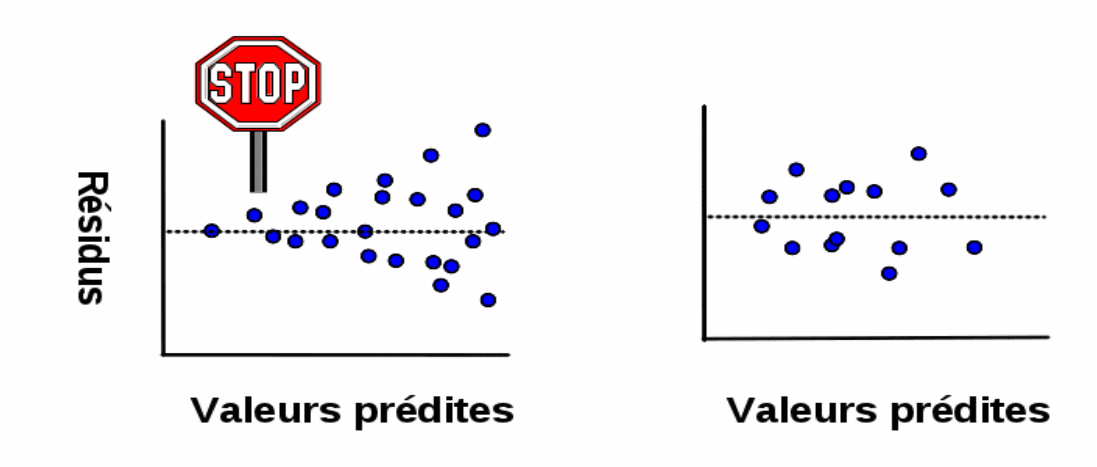
Homoscédasticité des résidus

La technique d'estimation utilisée suppose que résidus $\epsilon_i = \hat{y}_i - y_i$ ont une variance σ^2 constante (ne dépendant pas de i).

Homoscédasticité des résidus

Pour vérifier cette hypothèse, on représente généralement les résidus en fonction des variables explicatives (ou des valeurs prédites), et on vérifie visuellement que la variance est homogène sur l'ensemble de variation de chaque variable explicative

- représentation graphique



Test de non corrélation des résidus

La technique d'estimation utilisée suppose que les résidus sont non corrélés.

Test de Durbin-Watson

Le test de Durbin-Watson permet de vérifier que les ϵ_i ne sont pas corrélés.
Statistique de test :

$$d = \frac{\sum_{i=2}^{n} (\epsilon_i - \epsilon_{i-1})^2}{\sum_{i=1}^{n} \epsilon_i^2}$$

qui doit être proche de 2.

Analyse de variance de la régression

On teste l'apport du modèle de régression

Hypothèses

$H_0 : \beta_1 = \ldots = \beta_p = 0$ contre $H_1 : \exists j : \beta_j \neq 0$

Statistique de test

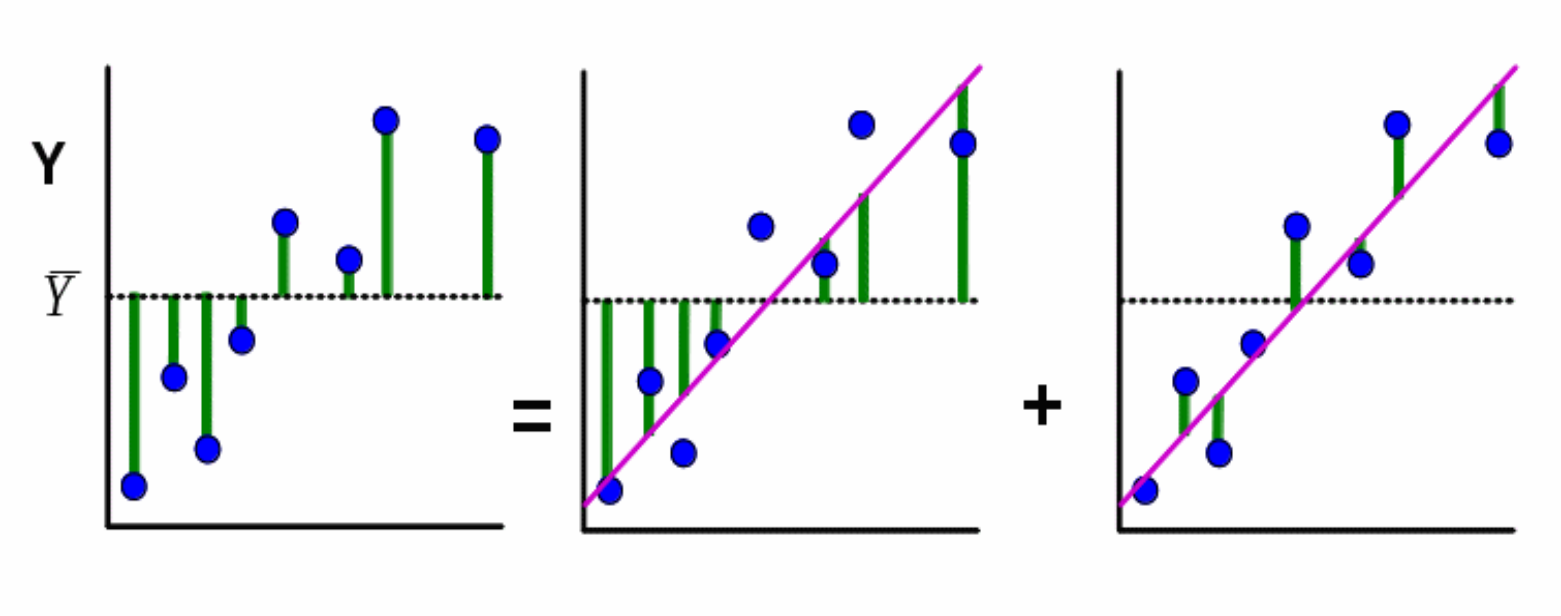
On décompose la variance de Y en $\underbrace{||\mathbf{Y} - \bar{\mathbf{Y}}||_2^2}_{SST} = \underbrace{||\hat{\mathbf{Y}} - \bar{\mathbf{Y}}||_2^2}_{SSReg} + \underbrace{||\mathbf{Y} - \hat{\mathbf{Y}}||_2^2}_{SSR}$

Source	Somme des carrés	degrés de liberté	carré moyen	F
Régression	$SSReg$	p	$MSReg = SSReg/p$	$F = \frac{MSReg}{MSR}$
Erreur	SSR	$n-p-1$	$MSR = SSR/(n-p-1)$	
Total	SST	$n-1$		

Décision

on rejette H_0 (la régression est valide) si $F > f_{p,n-p-1,1-\alpha}$

Analyse de variance de la régression



SST
variance
totale

SSReg
variance
expliquée

SSR
variance
résiduelle

Coefficient de détermination

Coefficient de détermination

Le **coefficient de détermination** R^2 :

$$R^2 = \frac{SSReg}{SST}$$

est un indicateur de la qualité du modèle de régression.
Propriétés :

- $R^2 \in [0, 1]$
- dans le cas de la régression simple : $R^2 = \rho_{XY}^2$
- plus le nombre de variables est grand, plus R^2 est grand

Coefficient de détermination ajusté

Coefficient de détermination ajusté

Le **coefficient de détermination ajusté** R^2_{adj} :

$$R^2_{adj} = \frac{(n-1)R^2 - d}{n-d-1}$$

est un indicateur de la qualité du modèle de régression, prenant en compte la complexité du modèle (nombre de variables).
Propriétés :

- $R^2_{adj} \in [0, 1]$
- plus R^2_{adj} est grand, meilleure est la régression

Tests de la nullité des paramètres du modèle

On peut également tester l'apport de chaque variable dans le modèle

Hypothèses

$H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$

Statistique de test

$T = \frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \sim_{H_0} t_{n-p-1}$

Décision

on rejette H_0 (et donc on enlève la variable du modèle) si $|t| > t_{n-1,1-\frac{\alpha}{2}}$.

Plan

Prédiction

Pour une valeur $x^* = (1, x_1^*, \ldots, x_p^*)'$ de X, la prévision de Y sera donnée par

$$\hat{y}^* = x^{*\prime}\hat{\beta}. \tag{3}$$

Un intervalle de confiance de niveau $1 - \alpha$ pour la valeur y^* sera construit à partir de cette prévision ponctuelle :

$$x^{*\prime}\hat{\beta} \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{1 + x^{*\prime}(\mathbf{X}'\mathbf{X})^{-1}x^*}. \tag{4}$$

Plan

Détection d'observations atypiques

Effet levier

L'effet levier h_i mesure l'impact de Y_i dans l'estimation $\hat{Y}_i$

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}.$$

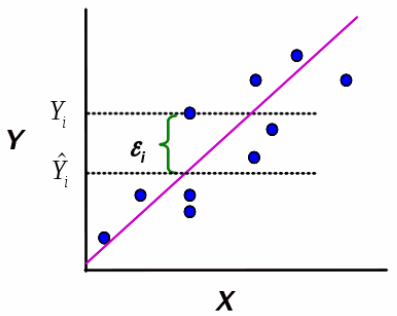
Cet impact est directement lié à l'éloignement de l'observation X_i à la moyenne des observations $\bar{X}$.

effet levier h_i grand $\Rightarrow$ observations atypiques

Détection d'observations atypiques

Résidus

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



Résidus normalisés/studentisés

$r_i = \frac{\epsilon_i}{S_{\epsilon(i)}\sqrt{1-h_i}}$ où $S_{\epsilon(i)} = \frac{n-2}{n-3} S_\epsilon - \frac{1}{n-3}\frac{\epsilon_i^2}{1-h_i}$

$|r_i| > 2 \Rightarrow$ observations atypiques

Détection d'observations atypiques

- effet levier $\Rightarrow$ éloignement d'une observation à la moyenne
- résidus normalisés $\Rightarrow$ éloignement observation / prédiction

La distance de Cook synthétisant ces deux informations.

Distance de Cook

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_{j(i)} - \hat{Y}_j)^2}{2S_\epsilon^2} = \frac{h_i}{2(1-h_i)} r_i^2$$

où $\hat{Y}_{j(i)}$: estimation de Y_j obtenue sans utiliser (X_i, Y_i).

$$D_i > 1 \Rightarrow \text{observations atypiques}$$

Régression linéaire avec R

L'analyse

1. charger les données :
   `>data=read.table('filename.dat',header=TRUE)`
2. estimer le modèle :
   `>modele=lm(y ~ .,data=data)`
3. tester la normalité des résidus :
   `>shapiro.test(modele$residuals)`
4. vérifier graphiquement l'homoscédasticité et la normalité des résidus, la présence d'individus atypiques ... :
   `plot(modele)`
5. tester l'auto-corrélation des résidus (`package lmtest`) :
   `>dwtest(modele)`
6. analyser la qualité du modèle et l'apport de chaque variable :
   `>summary(modele)`