MA431 : Mathématiques appliquées à la sécurité

Modèle linéaire

D. Barcelo

Grenoble INP ESISAR

2022/2023

Liaisons entre variables statistiques

Comment traduit-on que deux variables statistiques sont liées entre elles ?

1. Si les deux variables statistiques sont qualitatives ?

2. Si les deux variables statistiques sont quantitatives ?

3. Et s'il y a plus de deux variables ?

Liaisons entre variables statistiques

Comment traduit-on que deux variables statistiques sont liées entre elles ?

1. Si les deux variables statistiques sont qualitatives ?
   Tableau de contingence
   Test de l'indépendance des variables (test du χ^2 par exemple)
2. Si les deux variables statistiques sont quantitatives ?

3. Et s'il y a plus de deux variables ?

Liaisons entre variables statistiques

Comment traduit-on que deux variables statistiques sont liées entre elles ?

1. Si les deux variables statistiques sont qualitatives ?
   Tableau de contingence
   Test de l'indépendance des variables (test du χ^2 par exemple)
2. Si les deux variables statistiques sont quantitatives ?
   Représentation graphique des variables puis recherche de corrélation.
3. Et s'il y a plus de deux variables ?

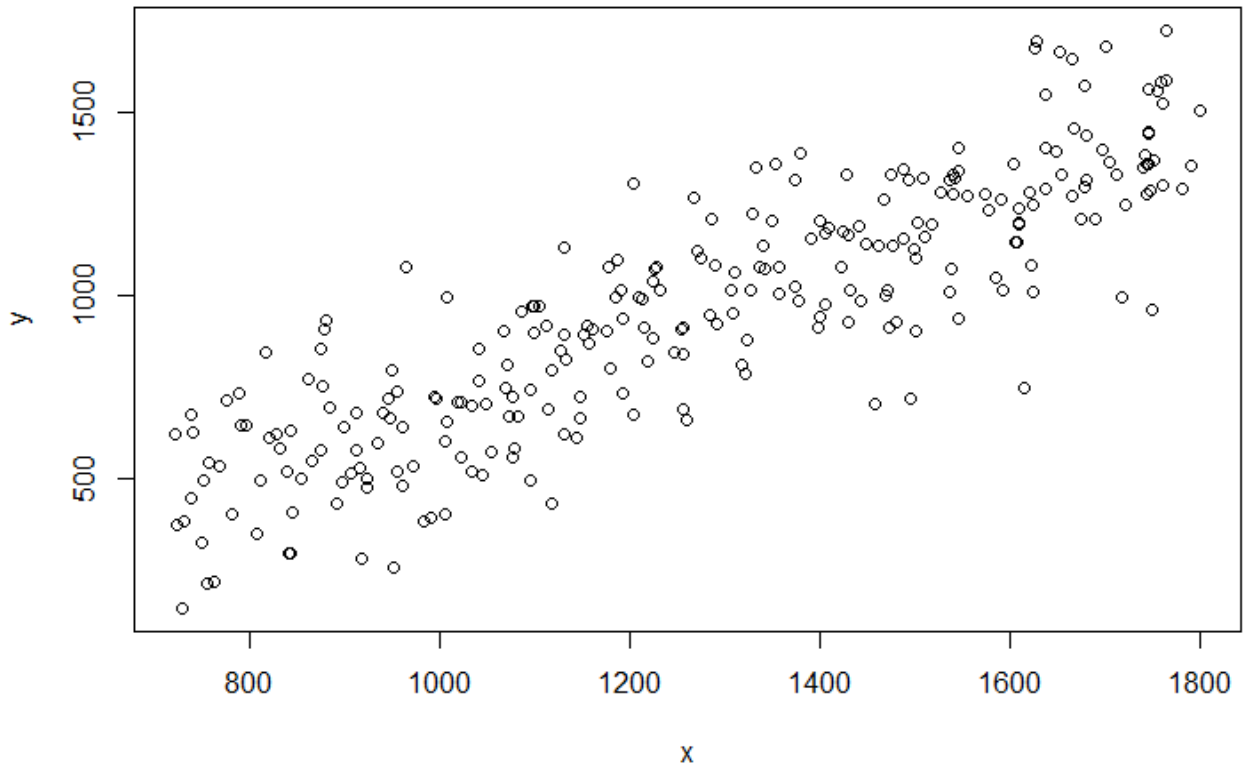
Liaisons entre variables statistiques

Comment traduit-on que deux variables statistiques sont liées entre elles ?

1. Si les deux variables statistiques sont qualitatives ?
   Tableau de contingence
   Test de l'indépendance des variables (test du χ^2 par exemple)
2. Si les deux variables statistiques sont quantitatives ?
   Représentation graphique des variables puis recherche de corrélation.
3. Et s'il y a plus de deux variables ?
   A voir suivant les cas.

Corrélation

Observation de deux variables statistiques continues X et Y
Représentation graphique d'un échantillon en nuage de points.



Coefficient de corrélation

Coefficient de corrélation linéaire

Soit x et y deux variables statistiques quantitatives observées sur n individus. On appelle coefficient de corrélation linéaire de x et y la quantité notée r_{xy} et définie par :

$$r_{xy} = \frac{Cov(x;y)}{\sigma_x \sigma_y}$$

avec $Cov(x;y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$

Coefficient de corrélation linéaire

Propriétés du coefficient de corrélation linéaire

- $\boxed{-1 \leqslant r_{xy} \leqslant 1}$.
- r_{xy} : mesure le lien entre les données observées pour x et y.
- Si $|r_{xy}|$ est proche de 1 alors les données sont fortement corrélées linéairement.
- Corrélation positive ou négative suivant le signe de r_{xy}. Indique le sens du lien entre les variables/
- Corrélation linéaire parfaite : $|r_{xy}| = 1$.
- Aucune corrélation linéaire : $r_{xy} = 0$.

Lien avec la sécurité

- Quantités simples à calculer,
- Recherche des liaisons entre différentes variables,
- Base des attaques par analyse différentielle de consommation de puissance,
- Coefficient de corrélation utilisé en cryptanalyse pour les attaques dites de corrélation.

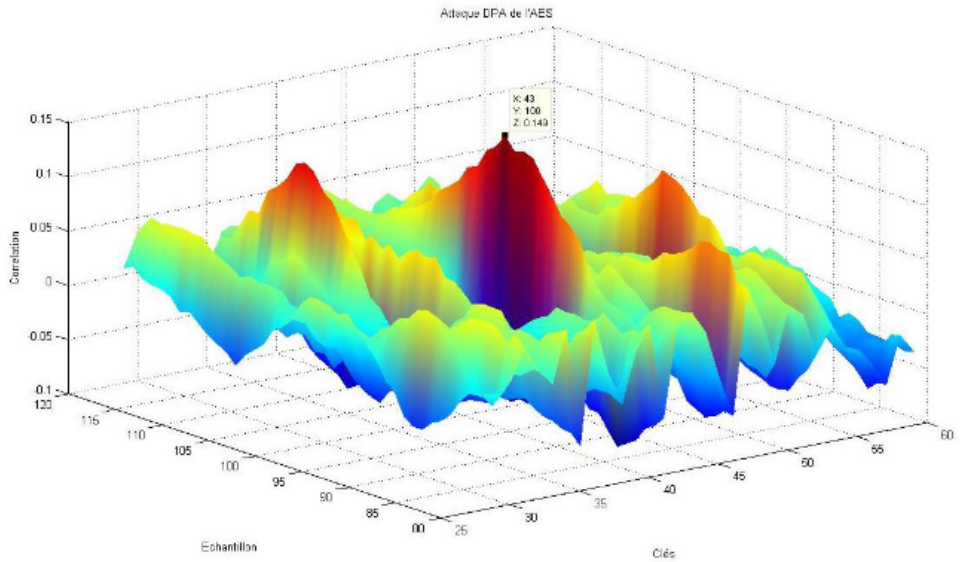
Attaque par corrélation

- Méthode de cryptanalyse utilisée contre le chiffrement par flot,
- le chiffrement par flot utilise des générateurs pseudo-aléatoires,
- Objectif : exploiter l'existence d'une éventuelle corrélation entre la sortie du générateur pseudo-aléatoire et celle d'un des registres utilisés pour déchiffrer les données.

(*Source : wikipedia*)

Lien avec la sécurité

Possibilité d'utilisation des calculs de corrélation pour les attaques dites physiques qui ciblent les implantations matérielles de fonctions cryptographiques, cf graphique d'un exercice d'attaque par analyse de la consommation de puissance d'un chiffreur.
(*Source : http://perso.univ-st-etienne.fr/bl16388h/publication2_fichiers/Bossuet_J3EA_2012.pdf*)



Calcul de la corrélation obtenue pour toutes les sous-clés possibles et pour tous les échantillons de mesures de consommation de puissance.

Modèle linéaire gaussien

Le modèle linéaire gaussien

Objectif : décrire une variable quantitative continue à l'aide d'une variable quantitative continue.

- $Y=$ variable à expliquer,
- $X=$ variable explicative,

Recherche d'une relation linéaire entre Y et X : $Y = f(X) + \varepsilon$
Avec f une application linéaire (voire affine) et ε une variable aléatoire gaussienne.

$$Y = aX + b + \varepsilon$$

Modèle linéaire gaussien

Le modèle linéaire gaussien

En pratique, on dispose de :.

- $(y_1, ..., y_n)$ échantillon de taille n de la variable Y,
- $(x_1, x_2, ..., x_n)$ n mesures de la variable X.

On doit donc avoir :

Modèle linéaire gaussien

Le modèle linéaire gaussien

En pratique, on dispose de :.

- $(y_1, ..., y_n)$ échantillon de taille n de la variable Y,
- $(x_1, x_2, ..., x_n)$ n mesures de la variable X.

On doit donc avoir :

$$\boxed{\forall i \in \{1, ..., n\}, y_i = ax_i + b + \varepsilon_i}$$

ε_i = erreurs (ou perturbations, ou résidus) possibles.
On peut noter $\hat{y}_i$ la prévision et alors $\varepsilon_i = y_i - \hat{y}_i$.

Le modèle linéaire gaussien

OBJECTIF : déterminer f

Le modèle linéaire gaussien

PROBLEME :

- Echantillon des variables X et Y.
- Valeurs des coefficients calculées : estimations ponctuelles des véritables valeurs de a et b.
- Estimateurs $\hat{a}$ et $\hat{b}$ de a et b.

Hypothèses du modèle

Hypothèses sur les résidus :

Hypothèses du modèle

Hypothèses sur les résidus :

1. Les perturbations ε_i sont d'espérances nulles :

$$\boxed{\forall i \in \{1, ..., n\}, \mathbb{E}(\varepsilon_i) = 0}$$

Hypothèses du modèle

Hypothèses sur les résidus :

1. Les perturbations ε_i sont d'espérances nulles :
$$\boxed{\forall i \in \{1, ..., n\}, \mathbb{E}(\varepsilon_i) = 0}$$
2. Les variances des perturbations sont supposées égales (homoscédasticité) :
$$\boxed{\forall i \in \{1, ..., n\}, \forall j \in \{1, ..., n\}, \mathbb{V}(\varepsilon_i) = \mathbb{V}(\varepsilon_j) = \sigma^2}$$

Hypothèses du modèle

Hypothèses sur les résidus :

1. Les perturbations ε_i sont d'espérances nulles :
$$\boxed{\forall i \in \{1, ..., n\}, \mathbb{E}(\varepsilon_i) = 0}$$
2. Les variances des perturbations sont supposées égales (homoscédasticité) :
$$\boxed{\forall i \in \{1, ..., n\}, \forall j \in \{1, ..., n\}, \mathbb{V}(\varepsilon_i) = \mathbb{V}(\varepsilon_j) = \sigma^2}$$
3. Les perturbations sont non corrélées :
$$\boxed{\forall i \in \{1, ..., n\}, \forall j \in \{1, ..., n\}, i \neq j, Cov(\varepsilon_i; \varepsilon_j) = 0}$$

Hypothèses du modèle

Hypothèses sur les résidus :

1. Les perturbations ε_i sont d'espérances nulles :
$$\boxed{\forall i \in \{1, ..., n\}, \mathbb{E}(\varepsilon_i) = 0}$$
2. Les variances des perturbations sont supposées égales (homoscédasticité) :
$$\boxed{\forall i \in \{1, ..., n\}, \forall j \in \{1, ..., n\}, \mathbb{V}(\varepsilon_i) = \mathbb{V}(\varepsilon_j) = \sigma^2}$$
3. Les perturbations sont non corrélées :
$$\boxed{\forall i \in \{1, ..., n\}, \forall j \in \{1, ..., n\}, i \neq j, Cov(\varepsilon_i; \varepsilon_j) = 0}$$
4. Les perturbations suivent des lois normales :
$$\boxed{\forall i \in \{1, ..., n\}, \varepsilon_i \sim \mathcal{N}(0; \sigma)}$$

Remarques sur le modèle linéaire

- On assimile la variable statistique Y et le vecteur $(Y_i)_{i\in[\![1;n]\!]}$ des n observations.
- (x_i) sont connus (donc constants) et on veut prévoir les valeurs des Y_i.
- Y_i : variables aléatoires de même espérance et de même variance.
- Connaitre $X = x_i$ ne permet pas exactement de déterminer la valeur de Y mais au moins une valeur moyenne : $\mathbb{E}(Y/X = x_i)$.

◂ démo

Que doit-on déterminer ?

Lorsque le modèle est connu, il se pose plusieurs questions :

- Peut-on déterminer a, b et σ^2 ?
- Ces valeurs sont-elles fiables ? Si oui, dans quelles limites ?
- Peut-on effectuer des prévisions à l'aide de ce modèle ?
- Ce modèle est-il applicable aux données observées ?

Que doit-on déterminer ?

Statistiquement, les questions précédentes se traduisent par

- Comment estimer a, b et σ^2 ?
- Quelles lois suivent les estimateurs ?
- Peut-on construire des intervalles de confiance à l'aide de ces estimateur ?
- Quels tests peut-on réaliser pour valider la dépendance linéaire des deux variables ?

Recherche des coefficients par la méthode des moindres carrés

Déterminer a et b (et σ^2).

Méthode des moindres carrés ordinaires (MCO).

Minimiser : $$\boxed{SCR(a,b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2}$$

Coefficients empiriques : $\bar{x_n}, \bar{y_n}, s_x^2, s_y^2, cov(x,y), r_{xy}$

Attention, certains sont des constantes et d'autres des réalisations de variables aléatoires.

Recherche des coefficients par la méthode des moindres carrés

Après calculs, on obtient que $SCR(a, b)$ est minimal en :

$$\begin{cases} \hat{a} = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \dfrac{cov(x; y)}{s_x^2} \\ \hat{b} = \bar{y_n} - a\bar{x_n} \end{cases}$$

$\hat{a}$ et $\hat{b}$: estimations de a et b obtenues à partir d'un échantillon.
Valeurs fluctuent en fonction de l'échantillon choisi.

Les estimations

A partir des résultats précédents, on peut définir :

Les estimations

A partir des résultats précédents, on peut définir :

- Les valeurs estimées : $\forall i \in \{1, ..., n\}, \hat{y}_i = \hat{a}x_i + \hat{b}$; on note $\hat{Y}$ le vecteur des estimations,

Les estimations

A partir des résultats précédents, on peut définir :

i) Les valeurs estimées : $\forall i \in \{1, ..., n\}, \hat{y}_i = \hat{a}x_i + \hat{b}$; on note $\hat{Y}$ le vecteur des estimations,

ii) Les résidus empiriques : $\forall i \in \{1, ..., n\}, \hat{\varepsilon}_i = Y_i - \hat{y}_i$; on note $\hat{\varepsilon}$ le vecteur des résidus,

Les estimations

A partir des résultats précédents, on peut définir :

i) Les valeurs estimées : $\forall i \in \{1, ..., n\}, \hat{y}_i = \hat{a}x_i + \hat{b}$; on note $\hat{Y}$ le vecteur des estimations,

ii) Les résidus empiriques : $\forall i \in \{1, ..., n\}, \hat{\varepsilon}_i = Y_i - \hat{y}_i$; on note $\hat{\varepsilon}$ le vecteur des résidus,

iii) L'estimateur de la variance : $S^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{y}_i)^2$;

Estimateurs

$\hat{a}$ et $\hat{b}$ sont les meilleurs estimateurs linéaires non biaisés de a et b. (on dit qu'ils sont BLUE)

Estimateurs : variables aléatoires avec espérances, variances et lois.

Caractéristiques de $\hat{a}$:

$$\mathbb{E}(\hat{a}) = a \text{ et } \mathbb{V}(\hat{a}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x_n})^2}$$

$\hat{a}$ suit une loi normale de variance inconnue.

Estimation par $\hat{S}_a^2$: $\hat{S}_a^2 = \dfrac{s^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$

s^2 la variance des résidus calculée sur l'échantillon.

Loi de Student à n-2 degrés de libertés.

$$\frac{\hat{a} - a}{\hat{S}_a} \sim T_{n-2}$$

Caractéristiques de $\hat{b}$:

$$\boxed{\mathbb{E}(\hat{b}) = b}$$

$\hat{b}$ suit une loi normale de variance inconnue.

Estimation par $\hat{S}_b^2$: $\hat{S}_b^2 = \frac{1}{n-2} s^2 \left(\frac{1}{n} + \frac{\bar{x_n}^2}{\sum_{i=1}^{n} (x_i - \bar{x_n})^2} \right)$

s^2 la variance des résidus calculée sur l'échantillon.

Loi de Student à n-2 degrés de libertés.

$$\boxed{\frac{\hat{b} - b}{\hat{S}_b} \sim T_{n-2}}$$

Intervalles de confiances :

Prévision de Y en un point x_0.

Prévision

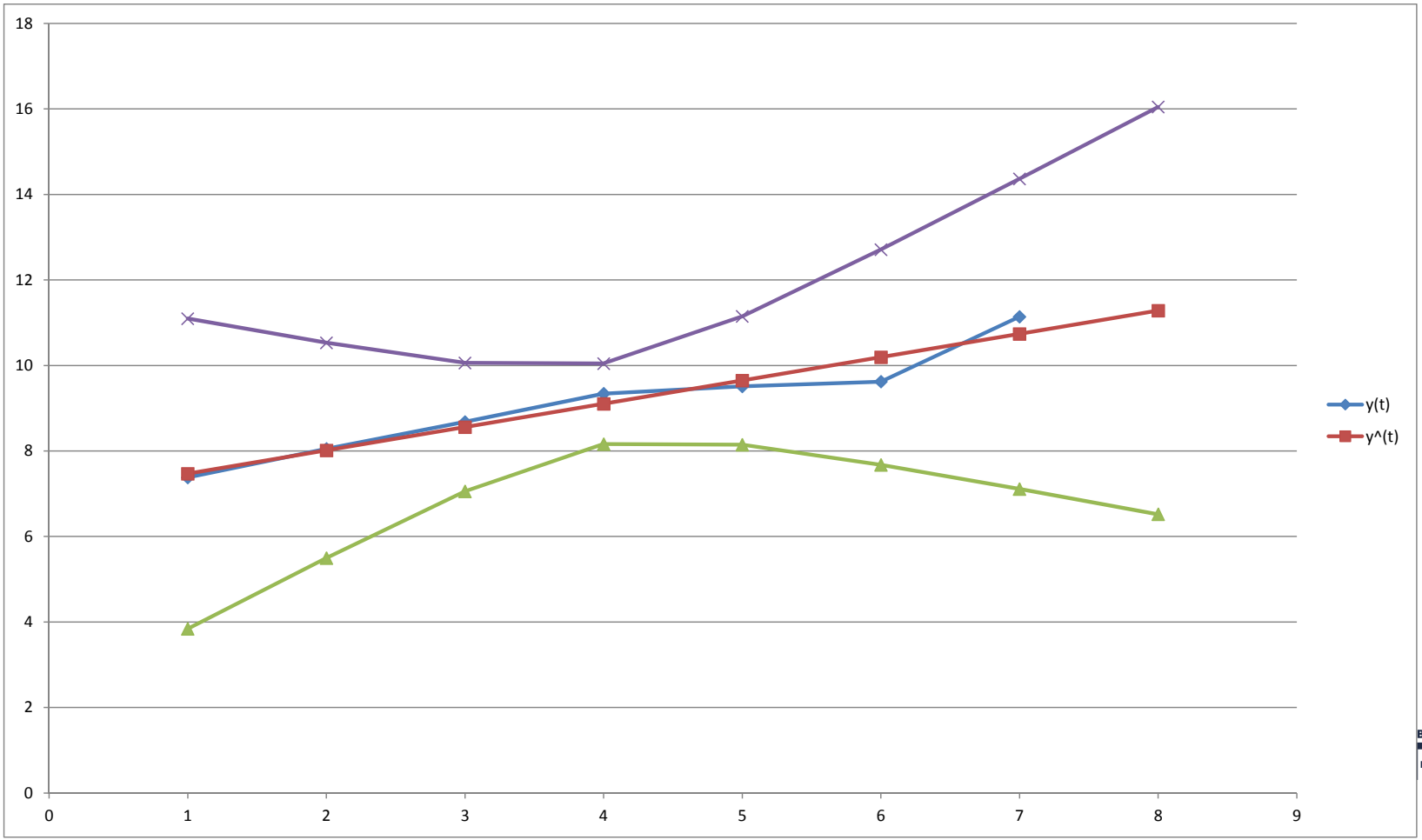
- $\hat{y}_0 = ax_0 + b$: prévision de Y et de sa moyenne $E(Y)$.
- Intervalle de confiance à α de $E(Y)$:

$$\hat{y}_0 \pm t_{1-\frac{\alpha}{2},n-2} s \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x_n})^2}{\sum_{i=1}^{n}(x_i - \bar{x_n})^2}\right)}$$

- Intervalle de confiance à α de Y :

$$\hat{y}_0 \pm t_{1-\frac{\alpha}{2},n-2} s \sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x_n})^2}{\sum_{i=1}^{n}(x_i - \bar{x_n})^2}\right)}$$

Intervalles de confiances :



Respect des hypothèses sur les résidus

Respect des hypothèses sur les résidus

- Vérifier la nullité, en moyenne, des résidus.

Respect des hypothèses sur les résidus

- Vérifier la nullité, en moyenne, des résidus.
- Vérifier l'absence d'autocorrélation des résidus (test de Durbin-Watson).

Respect des hypothèses sur les résidus

- Vérifier la nullité, en moyenne, des résidus.
- Vérifier l'absence d'autocorrélation des résidus (test de Durbin-Watson).
- Vérifier l'homoscédasticité à l'aide d'un test statistique (test de Levene)

Respect des hypothèses sur les résidus

- Vérifier la nullité, en moyenne, des résidus.
- Vérifier l'absence d'autocorrélation des résidus (test de Durbin-Watson).
- Vérifier l'homoscédasticité à l'aide d'un test statistique (test de Levene)
- Vérifier la normalité de manière empirique (droite de Henry) ou bien à l'aide d'un test de normalité type Kolmogorov-Smirnov.

Analyse de la variance

Test de la validité du modèle par l'analyse de variance.

Calcul de la variance expliquée par le modèle, la variance résiduelle et la variance totale. Comparaison de la variance expliquée à la variance résiduelle.

	Variations	degrés de libertés	Variance
Régression	$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_n)^2$	1	Variance expliquée (VE)
Résiduelle	$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	$n-2$	Variance Résiduelle (VR)
Totale	$\sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2$	$n-1$	Variance Totale (VT)

démo

Analyse de la variance

$\frac{VE}{VR}(n-2)$ suit une loi de Fisher à (1 ;n-2) degrés de libertés.
Test pour savoir si la variance expliquée est significativement supérieure à la variance résiduelle. (validation de la pertinence du modèle)

Analyse de la variance

$\frac{VE}{VR}(n-2)$ suit une loi de Fisher à (1 ;n-2) degrés de libertés.
Test pour savoir si la variance expliquée est significativement supérieure à la variance résiduelle. (validation de la pertinence du modèle)
Hypothèse H_0 : hypothèse que tous les coefficients sont nuls.

Rejet de H_0 : validation du modèle par non nullité d'au moins un des coefficients.
(test F est significatif)

Exemple

Exemple de régression linéaire.

Exemple

On observe simultanément sur 268 pièces leur résistance à la traction ainsi que leur limite élastique. Voici ce que donne la commande lm(formula = y x) sous R

Residuals : Interprétation

Min	1Q	Median	3Q	Max
-535.53	-117.39	-1.97	111.61	412.61

Coefficients : Interprétation

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-253.65968	44.77225	-5.666	3.8e-08 ***
x	0.95244	0.03428	27.787	< 2e-16 ***

Signif. codes : 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error : 172.5 on 266 degrees of freedom

Multiple R-squared : 0.7438, Adjusted R-squared : 0.7428

F-statistic : 772.1 on 1 and 266 DF, p-value : < 2.2e-16

Interprétation

Regression non linéaire

Autre modèle de regression, non linéaire, de la forme :

$$\boxed{\forall i \in \{1, ..., n\}, Y_i = f(x_i) + \varepsilon_i}$$

regression non linéaire : f non linéaire.
Modèle non gaussien : Hypothèse des résidus i.i.d. mais sans obligation d'une loi normale.

Regression linéaire multiple

Modèle pour expliquer une variable à l'aide de plusieurs variables explicatives.

Regression linéaire multiple

Relation linéaire entre :

- une variable Y
- p variables $X_1, \ldots, X_p$

On dispose de n observations.

$$\forall i \in [\![1; n]\!] \; Y_i = \sum_{j=1}^{p} a_j x_{i,j} + b + \varepsilon_i$$

avec $(\varepsilon_i)_{i \in [\![1;n]\!]}$ n variables aléatoires i.i.d. suivant une loi normale centrée.

Regression linéaire multiple

Matriciellement :

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix} \begin{pmatrix} b \\ a_1 \\ \vdots \\ a_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Regression linéaire multiple

Donc, en posant :

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix} \quad A = \begin{pmatrix} b \\ a_1 \\ \vdots \\ a_p \end{pmatrix} \text{ et } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\boxed{Y = AX + \varepsilon}.$$

Regression linéaire multiple

On peut montrer, en cherchant à minimiser distance euclidienne de $\mathbb{R}^n$, que l'on a : Démo

$$\left({}^tX\,X\right)A = {}^t\,X\,Y$$

Si la matrice $\left(X\,{}^tX\right)$ est inversible, on a :

$$A = \left({}^tX\,X\right)^{-1}\,{}^tX\,Y$$

Même étude sous R avec même interprétation des coefficients.

Interprétation géométrique : projection orthogonale de Y sur l'espace engendré par X.

exemple

On a testé une regression polynômiale de degré 5 sur des données.
Voici les résultats de la regression :

Residuals	Min	1Q	Median	3Q	Max
	-113.844	-13.259	1.094	17.181	48.031
Coefficients :	Estimate	Std. Error	t value	$Pr(>\|t\|)$	
(Intercept)	-380.5	592.2	-0.643	0.527	
xpol	202.8	250.3	0.810	0.426	
xpol2	-38.07	39.06	-0.975	0.339	
xpol3	3.059	2.858	1.070	0.295	
xpol4	0.8839	0.09899	8.929	4.28e-09	
xpol5	0.001648	0.001308	1.260	0.220	
Residual	standard error :	35.39	on 24 degrees	of freedom	
Multiple	R-squared :	1	Adjusted	R-squared :	1
F-statistic :	9.336e+07	on 5 and 24 DF	p-value :	$<$ 2.2e-16	

Démonstrations et interprétations

Espérance et variance des Y_i

On a : $Y_i = aX + b + \varepsilon_i$.

Cours

Espérance et variance des Y_i

On a : $Y_i = aX + b + \varepsilon_i$. Par linéarité de l'espérance on a :

$$\mathbb{E}(Y_i/X = x_i) = \mathbb{E}(ax_i + b + \varepsilon_i) = ax_i + b$$

Cours

Espérance et variance des Y_i

On a : $Y_i = aX + b + \varepsilon_i$. Par linéarité de l'espérance on a :

$$\mathbb{E}(Y_i/X = x_i) = \mathbb{E}(ax_i + b + \varepsilon_i) = ax_i + b$$

A l'aide des propriétés de la variance, on obtient :

$$\mathbb{V}(ax_i + b + \varepsilon_i) = \mathbb{V}(\varepsilon_i) = \sigma^2$$

Cours

Espérance et variance des Y_i

On a : $Y_i = aX + b + \varepsilon_i$. Par linéarité de l'espérance on a :

$$\mathbb{E}(Y_i/X = x_i) = \mathbb{E}(ax_i + b + \varepsilon_i) = ax_i + b$$

A l'aide des propriétés de la variance, on obtient :

$$\mathbb{V}(ax_i + b + \varepsilon_i) = \mathbb{V}(\varepsilon_i) = \sigma^2$$

Une transformation affine d'une loi normale suit également une loi normale :

$$Y_i/X = x_i \sim \mathcal{N}\left(ax_i + b; \sigma^2\right)$$

Cours

Coefficient de détermination

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{y}_i)^2$$

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(Y_i - ax_i - b)^2$$

$$EQM = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - ax_i - \bar{Y}_n + a\bar{x}_n\right)^2$$

$$EQM = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}_n - a(x_i - \bar{x}_n)\right)^2$$

$$EQM = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}_n\right)^2 + \frac{a^2}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 - \frac{2a}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}_n\right)(x_i - \bar{x}_n)$$

Coefficient de détermination

$$EQM = S_Y^2 + \frac{Cov(x;Y)^2}{s_x^4} s_x^2 - 2\frac{Cov(x;Y)}{s_x^2} Cov(x;Y)$$

$$EQM = S_Y^2 - \frac{Cov(x;Y)^2}{s_x^2}$$

$$EQM = S_Y^2 - \frac{Cov(x;Y)^2}{s_x^2}$$

$$EQM = S_Y^2 \left(1 - \frac{Cov(x;Y)^2}{s_x^2 S_Y^2}\right)$$

$$EQM = S_Y^2 \left(1 - r_{x,Y}^2\right)$$

Variance résiduelle

$$VE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - \bar{Y}_n\right)^2$$

$$VE = \frac{1}{n}\sum_{i=1}^{n}\left(ax_i + \bar{Y}_n - a\bar{x}_n - \bar{Y}_n\right)^2$$

$$VE = \frac{a^2}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2$$

$$VE = \frac{Cov(x;Y)^2}{s_x^4}s_x^2 \text{ donc } VE = \frac{Cov(x;Y)^2}{s_x^2}$$

Or $EQM = S_Y^2 - \frac{Cov(x;Y)^2}{s_x^2}$ avec $EQM = VR$. D'où :

$$VR = VT - VE$$

Cours

Exemple : Interprétation

Min	1Q	Median	3Q	Max
-535.53	-117.39	-1.97	111.61	412.61

Répartition des résidus, différents quantiles, minimum et maximum.
Si l'hypothèse de normalité des résidus est vérifiée alors la moyenne et la médiane doivent être égales, donc proches de 0.

◂ Exemple

Exemple : Interprétation

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-253.65968	44.77225	-5.666	3.8e-08 ***
x	0.95244	0.03428	27.787	< 2e-16 ***

(Intercept) : $\hat{b}$ calculé sur l'échantillon (donc estimation ponctuelle de b)
Std. Error : estimation de son écart type (pour les intervalles de confiance).
t value : statistique calculée dans le test de Student $\mathcal{H}_0 = \{b = 0\}$ contre $\mathcal{H}_1 = \{b \neq 0\}$
Pr(>|t|) : probabilité critique de ce test (ou p_{value}).

x : $\hat{a}$ calculé sur l'échantillon (donc estimation ponctuelle de a). (de même les coefficients que pour b).

◂ Exemple

Exemple : Interprétation

Residual standard error : 172.5 on 266 degrees of freedom
Multiple R-squared : 0.7438, Adjusted R-squared : 0.7428
F-statistic : 772.1 on 1 and 266 DF, p-value : $<$ 2.2e-16

Residual standard error représente la valeur de S (donc l'estimation ponctuelle de σ. Le degré de libertés $(n-2)$ est précisé également.
Multiple R-squared représente la valeur de r_{xy}^2 (ou le coefficient de détermination).
Adjusted R-squared représente le coefficient de détermination ajusté en fonction du nombre de variables explicatives. Il permet de comparer des modèles avec un nombre différent de variables explicatives pour savoir si on doit les prendre en compte, il est moins sensible que R^2 au nombre de variables et peut être négatif. $R^2_{\text{ajusté}} = 1 - \dfrac{(1-R^2)(n-1)}{n-p-1} < R^2$
F-statistic est la statistique du test de Fisher effectué pour savoir si la regression est pertinente. La **p-value** représente le probabilité de rejeter à tort l'hypothèse que le test n'est pas pertinent.

Regression linéaire multiple

Minimiser : $\|Y - AX\|^2$.

Equivaut à minimiser : $\sum_{i=1}^{n} \left(y_i - b - \sum_{j=1}^{p} a_j x_{i,j} \right)^2$.

On doit donc annuler les dérivées partielles de cette fonction en $p+1$ variables, soit résoudre le système :

Regression linéaire multiple

$$\begin{cases} \sum_{i=1}^{n} -2\left(y_i - b - \sum_{j=1}^{p} a_j x_{i,j}\right) = 0 \\ \sum_{i=1}^{n} -2x_{i,1}\left(y_i - b - \sum_{j=1}^{p} a_j x_{i,j}\right) = 0 \\ \vdots \\ \sum_{i=1}^{n} -2x_{i,p}\left(y_i - b - \sum_{j=1}^{p} a_j x_{i,j}\right) = 0 \end{cases} \Leftrightarrow \begin{cases} \sum_{i=1}^{n} 1.y_i - 1.\left(b + \sum_{j=1}^{p} a_j x_{i,j}\right) = 0 \\ \sum_{i=1}^{n} x_{i,1}y_i - x_{i,1}\left(b + \sum_{j=1}^{p} a_j x_{i,j}\right) = 0 \\ \vdots \\ \sum_{i=1}^{n} x_{i,p}(y_i - x_{i,p}\left(b + \sum_{j=1}^{p} a_j x_{i,j}\right) = 0 \end{cases}$$

Regression linéaire multiple

Matriciellement, on a donc à résoudre : ${}^tX\,Y - \left({}^tX\,X\right)A = 0$.
On obtient donc : ${}^tX\,Y = \left({}^tX\,X\right)A$.
Donc, si la matrice $\left({}^tX\,X\right)$ est inversible, on a :

$$A = \left({}^tX\,X\right)^{-1}\,{}^tX\,Y$$

cours