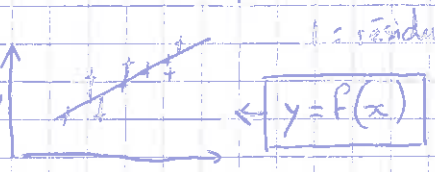
Regression linéaire multiple

15/05/2012

On veut expliquer y par rapport à x.



$\hat{y}_i = bx + a$ $\quad y_i = bx + a + \varepsilon_i$

R^2

régresseur = variables explicatives exemple : T°, tps, taille...
= paramètre de fabrication
régresseurs notés $x_1, x_2, x_3 \dots x_k$

réponse = critère que qualité

"Tout le monde sait ce que c'est une rognure !"

On a un plan expérimental : "on va faire des essais de fabrication."

Plan expé :

essai	I	u_1	u_2	$\dots$	u_j	$\dots$	u_k
1	1	u_{11}					
2	1	u_{21}					
$\vdots$	$\vdots$	$\vdots$					
i	1	u_{i1}					
$\vdots$	$\vdots$	$\vdots$					
n	1	u_{n1}					

On doit trouver tous les β de la formule : $\hat{y} = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + \dots + \beta_j u_j + \dots + \beta_k u_k$

$k' = k+1$ coef. Eqm

IV Regression selon le modèle complet

En écriture matricielle, $y = u\beta$. La matrice u ("des effets") permet de calculer ces coef. β. On construit la matrice u en ajoutant une colonne "de plus 1"

"I est le vecteur identitaire"

Le produit matriciel $u'u$ est inversible : $(u'u)^{-1} u'y = (u'u)^{-1}(u'u)\beta$

$$\beta = (u'u)^{-1} u'y$$

Les dimensions de u sont : $\dim u = [n ; k+1]$

Les dimensions de son u' : $\dim u' = [k+1 ; n]$

$u'u$ est la matrice d'information

$\dim u'u = [k+1 ; k+1]$ et $\dim (u'u)^{-1} = [k+1 ; k+1]$

$\dim y = [n ; 1]$; $\dim u'y = [k+1 ; 1]$; $\dim (u'u)^{-1} u'y = [k+1 ; 1] = \dim \beta$

exemple p. 7

CAS de matrice inverse calculée vite :

si matrice d'info : $\begin{pmatrix} 6 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 7 \end{pmatrix}$ alors son inverse : $\begin{pmatrix} \frac{1}{6} & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{5} & 0 \\ 0 & 0 & 0 & \frac{1}{7} \end{pmatrix}$

exemple p. 7 (on considère u_1 → Température, u_2 → temps, u_3 → dose conservateurs)

avec la matrice 7 on pose $\hat{y} = -7,887 + 0,456\,u_1 + 0,878\,u_2 - 0,045$ (1)

Quand la T° ↗ de 1°C, le log du nb de bactéries ↗ de 0,456
le tps ↗ de 1 unité, le log du nb de bactéries ↗ de 0,878
on ↗ de 1% la qtt de conservateurs, le log du nb de bactéries ↘ de 0,045

NB: le −7,887, ordonnée à l'origine, n'a pas de sens. Il est mathématique.

→ Pour valider de modèle (1), on utilise l'ANOVA.
→ on veut ensuite être sûr que les facteurs sont représentatifs
pour ce faire, on fait un test sur les coef. du modèle.

On dresse donc un tableau | y_i | $\hat{y}_i$ | e_i | (cf. p. 7)

et on rajoute une ligne | SCE → $SCE_y = SCE_{\hat{y}} + SCE_e$

La $SCE_{\hat{y}}$ permet de mesurer la part de variation des réponses y qu'on peut expliquer avec le modèle.

La SCE_e mesure la part de varia° des y qu'on est pas capable d'expliquer. C'est de la fluctuation interne naturelle qui va permettre de calculer le CM_e.

Propriétés magiques : $\Sigma e_i = 0$ $\quad \bar{e} = 0$ $\quad SCE_e = SC_e$

$\Sigma y_i = \Sigma \hat{y}_i$ et donc $\bar{y}_i = \bar{\hat{y}}_i$

$$\begin{aligned}
\hat{y}_1 &= b_0 + b_1 u_{11} + \dots + b_j u_{1j} + \dots + b_k u_{1k} \\
\hat{y}_2 &= b_0 + b_1 u_{21} + \dots + b_j u_{2j} + \dots + b_k u_{2k} \\
&\vdots \\
\hat{y}_i &\\
&\vdots \\
\hat{y}_n &
\end{aligned}$$

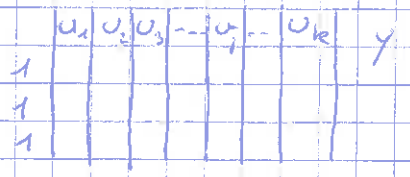
on a donc
n équations (n essais)
k+1 inconnues (k+1 coef)

NB: on doit avoir autant d'équations que d'inconnues

15/05/2012 Hier, on faisait varier des paramètres u_i.
Avec les observations, on dresse des equations

15/0 On a donc un plan d'expérience

puis une matrice des effets
en ajoutant une colonne de 1



Avec toutes les étapes, on sait une matrice $B = (U'U)^{-1} U'Y$

la réponse estimée : $\hat{y} = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + \dots + \beta_j u_j + \dots + \beta_k u_k$

Point culture = les matrices

IV 2 Validation du modèle

MAINTENANT, on voudrait savoir si le modèle est valable.
On calcule les SCE : $SCE_y = SCE_{\hat{y}} - SCE_e$

$H_0 = \frac{CM_{\hat{y}}}{CM_e} = 1$ $\qquad$ $H_1 = \frac{CM_{\hat{y}}}{CM_e} > 1 \rightarrow$ peu de variations inexpliquées

ANOVA → variation	SCE	ddl	CM	
- expliquée par modèle	$SCE_{\hat{y}}$	$k'-1$	$CM_{\hat{y}}$	Δ
- pas expliquée (résidus)	SCE_e	$n-k'$	CM_e	
- totale	SCE_y	$n-1$		

NB : * $SCE_y = \sum y_i^2 - n\bar{y}^2$

$SCE_{\hat{y}} = \sum \hat{y}_i^2 - n\bar{\hat{y}}^2$

$SCE_e = \sum (e_i - \bar{e})^2 = \sum e_i^2 = SC_e$

en exam, sont donnés
- $\sum y_i$
- $\sum y_i^2$
- $\sum \hat{y}^2$

$\bar{y} = \bar{\hat{y}}$

* si on a k facteurs, combien a-t-on de coef β ? → $k+1 = k'$
donc on pose $ddl = k'-1$
* $ddl_e = n-1-(k'-1)$

On veut prouver H_1 : si $CM_{\hat{y}} \gg CM_e$, ça veut dire que le modèle a peu de variations non expliquées. DONC c'est bien !

Si $F_{calc} < F_{théo}$, on ne peut pas prouver la validité du modèle linéaire additif et du 1er degré.

Si $F_{calc} > F_{théo}$, on met en évidence que le modèle additif linéaire du 1er degré avec moins de x% de risque d'erreur.

Δ si $CM_e > CM_{\hat{y}}$, on ne calcule pas de F_{calc}

IV 3 Coefficient de détermination

Maintenant, on va voir quelle proportion des variations on explique avec notre modèle.

R^2 est un coef. de détermination multiple (propre à un modèle)
↳ ≠ partiel

exemple : $R^2 = 0,8$ signifie que 80% des variations de y sont expliquées par les variations des b_j selon le modèle additif du 1er degré.

← elle a dit ce mais c'est vrai ?

$$R^2 = \frac{SCE_{\hat{y}}}{SCE_y}$$

IV 4 Test de signification sur chaque paramètre β_j

On sait autant de tester t que de coef. On pose

$H_0 : \beta_j = 0$ $\quad H_1 : \beta_j \neq 0$ avec $j = 1 \dots k$

On étudie une pop. avec $\hat{y} = b_0 + \sum b_j v_j$ $\quad b_0$ et b_j varient inévitablement avec les expériences. On pose $\hat{L}(b_j) = St(\nu)\left(E(b_j) ; \sigma(b_j)\right)$

$b_j = \beta_j \quad E(b_j) = \beta_j$

L'écart type de chaque coef b_j est = à la racine carré de la variance de b_j

Chaque variance de b_j se trouve ds la diagonale de la matrice des variances - covariances et notée Var-B

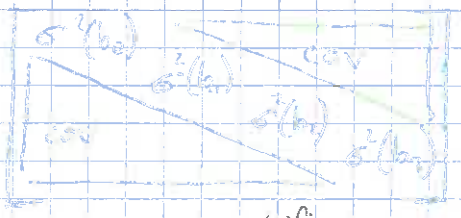
$$Var\,B = \sigma_\varepsilon^2 (v'v)^{-1}$$ avec σ_ε^2 l'erreur expé (CM_e), $(v'v)^{-1}$ la matrice de dispersion.

Var B = constante x matrice donc chaque terme de la matrice est simplement multiplié par la constante.

exemple : avec la matrice de dispersion (= inverse) p. 7

$0,166 \times 51,882 = 8,612 = \sigma^2(b_0)$
$0,166 \times 0,014 = 0,002 = \sigma^2(b_1)$
$0,166 \times 0,097 = 0,016 = \sigma^2(b_2)$

Ainsi, Var B :



← vérifier

de là on tire $t_{calc} = \frac{b_j - \beta_j}{\sigma^2(b_j)} = \frac{b_j}{\sigma^2(b_j)}$

Si $t_{calc} > t_{théo}$ on a mis en évidence un effet du facteur b_j sur y avec α% de risque d'erreur.

↖ vérifier

Δ qd $\beta \neq 0$ ça veut juste dire qu'il ne passe pas par l'origine. Retenir que b_0 il est pas comme les autres.

Si α n'est pas donné, on cherche les seuils de significativité avec les ** et NS etc

IV 5 Estimations par intervalle de confiance

$$P\left[b_j - t_{1-\frac{\alpha}{2}}(n-k') \times \sigma(b_j) < \beta_j < b_j + t_{1-\frac{\alpha}{2}}(n-k') \times \sigma(b_j)\right] = 1-\alpha$$

Quand on augmente U_j d'une unité, la réponse y a α% de risque de ne pas varier d'une quantité comprise entre la limite inférieure et supérieure de la variable d'estimation.

exemple : $P[10 < \beta_1 < 12] = 0,95$

si U_1 ↗ de 1°C, la qtt de p org a 5% de chances de ne pas varier en 10 et 12 unités.

si on a $P[-1 < \beta_j < +2] = 0,95$ ça veut dire que la réponse peut aussi bien ↗ que ↘ ou rester =
→ dans ces cas, on dit que β_j est non significatif

la covariance mesure le lien entre 2 variances (on en a de part et d'autre de la diag)

22/05/2012

III Méthodologie

2 types principaux → régression pas à pas (ascendante ou descendante)
↳ régression complet

V Régression selon le pas à pas

V La regression pas à pas

V 1 La matrice des corrélations

NB : on verra des R^2 partiels, et aucun rapport avec les R^2

exemple p. 9 : ① on étudie le 1er facteur, et on l'analyse

N°	I	U_2	$\hat{Y}$	Y	e
1	1				
⋮	⋮				
18	1				

matrice des effets

$$B = (\dot{U}U)^{-1}\dot{U}y$$

$$\hat{y} = 120,955 - 0,359 U_2$$

H_0 : $CM(\hat{y}_{/U_2}) / CM_e = 1$

H_1 : $\frac{CM_{fact}}{CM_e} > 1$

② On étudie le facteur suivant : U_3

N°	I	U_2	U_3	y	$\hat{y}$	e
1	1	210	4			
⋮	⋮	⋮	⋮			
18	1	290	18			

matrice des effets

$B = (U'U)^{-1} U'y$

$\hat{y} = 106,223 - 0,359 U_2 + 1,339 U_3$

$H_0 : CM_{\hat{y}/U_3} / CM_e = 1$ ← not sure

	SCE	ddl	CM	Fcalc
Reg	$SCE_{\hat{y}} = 4712$	2	$CM_{\hat{y}}$	$F_{total} = \frac{CM_{\hat{y}}}{CM_e}$
U_2	$SCE_{\hat{y}/U_2} = 3306$	1	$CM_{\hat{y}/U_2}$	④
U_3/U_2	$SCE_{\hat{y}/U_3/U_2} = 1406$	1	$CM_{\hat{y}/U_3/U_2}$	②
résid	$SCE_e = 665$	15	CM_e	
tot	$SCE_y = 5377$	17		

④ $F = \frac{CM_{\hat{y}/U_2}}{CM_e}$ ② $F = \frac{CM_{\hat{y}/U_3/U_2}}{CM_e}$

$R^2_{multiple} = \frac{SCE_{\hat{y}}}{SCE_y}$ $R^2_{y/U_2} = \frac{SCE_{\hat{y}/U_2}}{SCE_y}$ $R^2_{y/U_3/U_2} = \frac{SCE_{y/U_3/U_2}}{SCE_y}$

$R^2_{\hat{y}/U_3,U_2} = \frac{SCE_{\hat{y}/U_2 U_3}}{SCE_{\hat{y}}}$

* notation pour " U_3 entrée après U_2 "

Exercice 3 poly

(voir tableau)

2) a) $H_0 : \beta_0 = 0$ $\quad y = \beta_0 + \beta_1 v_1 + \varepsilon$

$H_1 : \beta_1 \neq 0$ $\quad \left(t_{calc} = \frac{b_1}{\sigma(b_1)} \quad (\text{à comparer à } t_{théo}) \right)$ mais ici on va plutôt utiliser un test F

	SCE	ddl	CM	Fcalc
$\hat{Y}/v_1$	981,326	1		25,76
résiduel	761,955	20		
total	1743,281	21		

$F_{théo} = F_{0,095}(1;20) = 4,35$

b) Cette fois on étudie $y = \beta_0 + \beta_1 v_1 + \beta_2 v_2 + \varepsilon$

Si on avait b_2 on pourrait étudier $t_{calc} = \frac{b_2}{\sigma(b_2)}$

Mais on a pas b_2 alors test F ↓

	SCE	ddl	CM	Fcalc
v_1	9,8,326	1		
v_2/v_1	190,326	1	190,23	6,32
résiduel	571,723	19	30,09	
total	1743,282	21		

$\rightarrow t_{calc} = \sqrt{6,32}$

$F_{théo} = F_{0,095}(1;19) = 4,38$

$t_{théo} = \sqrt{4,38}$

isaralyon
Une école d'ingénieurs au coeur de la vie

ISARA – 2ème année – STATISTIQUE
Cours : Partie 7
Mme B. Bottollier-Lemallaz

REGRESSION LINEAIRE MULTIPLE

SOMMAIRE

OBJECTIFS

- Construire les matrices U, U', U'U, U'Y associées à une régression linéaire multiple du 1er degré,
- Calculer (calculs matriciels) et interpréter concrètement les coefficients de la régression,
- Etablir la liste des Y estimés et celles des résidus,
- Calculer les SCE des y, des y estimés et celle des résidus,
- Réaliser et interpréter l'analyse de la variance associée au modèle de régression linéaire,
- Calculer et interpréter les coefficients de détermination (partiels et multiple) et de corrélation
- Calculer et interpréter la matrice des variances-covariances à partir de la matrice $(U'U)^{-1}$ donnée si nécessaire,
- Calculer les écarts type des coefficients de la régression,
- Appliquer les calculs d'estimations aux coefficients de la régression linéaire,
- Interpréter les intervalles d'estimation des coefficients de la régression,
- Réaliser et interpréter les tests de signification des coefficients de la régression,
- Réaliser et interpréter la régression pas à pas (ascendante ou descendante).

NB : Le calcul de la matrice inverse devra être réalisé sur Excel pour les exercices d'entraînement, par contre cette matrice sera donnée, si nécessaire, le jour de l'évaluation.

I – Les données

On a souvent besoin d'examiner la façon dont une variable quantitative est reliée à d'autres variables quantitatives.
La régression multiple généralise la régression simple en étudiant la liaison stochastique entre une variable aléatoire Y (critère de qualité d'un produit fabriqué) et k variables certaines U_1, U_2, Uj, U_k (paramètres de fabrication de ce produit) au sein d'une population donnée dont on observe un échantillon aléatoire. On suppose en outre que les variables indépendantes U_j sont mesurées sans erreur. Le fait d'introduire des variables supplémentaires peut diminuer la valeur de la variance résiduelle et par là améliorer l'analyse.

N°	Plan expérimental							Réponses
1	U_{11}	U_{12}	....	U_{1j}	....	U_{1k}		y_1
2	U_{21}	U_{22}	....	U_{2j}	....	U_{2k}		y_2
...								
...	....	....	....	....	....	....		....
i	U_{i1}	U_{i2}	....	U_{ij}	....	U_{ik}		yi
...	....	....	....	.	....	....		....
...	....	....	....	.	....	....		....
n	U_{n1}	U_{n2}	....	U_{nj}	...	U_{nk}	.	y_n

II – Le modèle linéaire additif

Dans le cas où Y est reliée linéairement à une seule variable U, il s'agit de **régression linéaire simple**, si Y est reliée linéairement à plusieurs variables Uj, il s'agit de **régression linéaire multiple**.
Y est la **variable dépendante** à expliquer ou variable de réponse et Uj sont les **variables explicatives** ou **variables indépendantes** ou encore **régresseurs**.
En fabrication : U : paramètres de fabrication, Y : critère de qualité du produit fabriqué

Le modèle mathématique fixé a priori : $\hat{y}_i = \beta_0 + \Sigma \beta j U_{ij}$

Le modèle à déterminer à partir du dispositif expérimental et à valider : $\hat{y}_i = \mathbf{b_0 + \Sigma bj U_{ij}}$

C'est un modèle à k' coefficient, ici k' = k +1

On mesure $y_i = b_0 + \Sigma bj U_{ij} + e_{ij}$ avec $e_i = \hat{y}_i - y_i$

Les coefficients βj sont les **coefficients de régression**. Ils sont estimés β^j = bj par la méthode des moindres carrés ordinaires (MCO). Un tel critère entraîne que la **somme des résidus soit nulle**.

III – Méthodologie

Une régression linéaire est toujours délicate à interpréter car les régresseurs ne sont généralement pas indépendants, on peut donc procéder de différentes façons:

↳ Régression progressive appelée aussi régression pas à pas ascendante (ou descendante) qui consiste à ajouter (ou supprimer) une variable explicative dans la mesure où les coefficients de détermination des 2 régressions successives sont significativement différents. Pour déterminer l'ordre d'introduction des régresseurs, on examine la matrice des corrélations. Dans le procédé ascendant on introduit la variable Uj la plus corrélée avec Y, puis on introduit comme seconde variable celle qui, après Uj, augmente le plus R^2 à condition que sa contribution soit significative. On peut introduire un troisième régresseur ....etc.
Dans le procédé descendant on élimine la variable Uj la moins corrélée avec Y, puis la 2ème moins corrélée...etc.
Avec cette méthode on peut déterminer, par le calcul, la part des variations des Y expliquées par chacun des régresseurs.

pour trouver les régresseurs ⇒

↳ Régression avec le modèle complet où l'équation contiendra alors les k régresseurs
Si, dans la régression ascendante toutes les variables Uj sont introduites (ou aucune n'est supprimée dans la régression descendante), le résultat est identique à celui obtenu avec le modèle complet.

IV - La régression selon le modèle complet

→ on calcule tous les coefs avec la matrice B, APRES → on vérifie

IV - 1 - Calcul des k' coefficients du modèle

Ils mesurent les effets de chaque facteur Uj sur la réponse Y.
Dans l'échantillon de n observations : $y_i = b_0 + b_1 U_{i1} + b_2 U_{i2} + ...+ b_j U_{ij} + ...+ b_k U_{ik} + e_i$

☞ **Dans la population :** $Y_i = \beta_0 + \beta_1 U_{i1} + \beta_2 U_{i2} + ...+ \beta_j U_{ij} + ...+ \beta_k U_{ik}$

Chaque βj est estimé ponctuellement par la valeur calculée bj dans l'échantillon.
Pour $j \neq 0$, bj mesure la variation de Y consécutive à une variation d'une unité de Uj, les autres facteurs de la régression restant constants. On appelle bj « l'effet du régresseur ».

Matrice des coefficients : $B = (U' U)^{-1} U' Y$
U est la matrice des effets
Y est la matrice des réponses
U' est la matrice transposée de U
$(U'U)^{-1}$ est la matrice de dispersion, c'est la matrice inverse du produit matriciel (U'U)

En notation matricielle $Y = UB + E$

↳ $L(\varepsilon_i) = N(0 ; \sigma_\varepsilon)$
↳ Il n'existe aucune corrélation entre les erreurs
↳ Les variables Uj sont des grandeurs certaines indépendantes entre elles
↳ Le nombre d'observations n doit être supérieur ou égal à k'

Y [n, 1]	Matrice des effets = U [n, k+1] = [n, k']						
y_1	1	U_{11}	U_{12}	....	U_{1j}	....	U_{1k}
y_2	1	U_{21}	U_{22}	....	U_{2j}	....	U_{2k}
....	....	....	....	....	....	....	....
y_i	1	U_{i1}	U_{i2}	....	U_{ij}	....	U_{ik}
....	....	....	....	....	.	....	....
....	....	....	....	....	.	....	....
y_n	1	U_{n1}	U_{n2}	....	U_{nj}	....	U_{nk}

B [k', 1]
b_0
b_1
....
b_j
....
....
b_k

E [n, 1]
e_1
e_2
....
e_i
....
....
e_n

Matrice d'information = Matrice U'U : **!!!Construction très importante à comprendre, pré requis indispensable en 3ème année !!!!**					Matrice U'Y
n	ΣU_1	ΣU_2	.....	ΣU_k	Σy
ΣU_1	ΣU_1^2	$\Sigma U_1\ U_2$.	....	$\Sigma U_1\ U_k$	$\Sigma U_1 y$
ΣU_2	$\Sigma U_2\ U_1$	ΣU_2^2.	....	$\Sigma U_2\ U_k$	$\Sigma U_2 y$
....	....	....	....	....	....
ΣU_k	$\Sigma U_k\ U_1$	$\Sigma U_k\ U_2$	....	ΣU_k^2	$\Sigma U_k y$

IV - 2 - Validation du modèle

Il s'agit de répondre à la question : la régression linéaire est-elle significative dans son ensemble?
On calcule les valeurs de Y estimées d'après le modèle puis on établit la liste des résidus.
On détermine la SCE des Y estimés et celle des résidus.

On pose: H0 : CM $\hat{y}$ / CM e = 1

H1 : CM $\hat{y}$ / CM e >1 le modèle est satisfaisant (test unilatéral)

Tableau de l'analyse de variance (ANOVA) :

origine de la variation	SCE (*)	ddl(**)	CM = SCE / ddl	F calculé
Régression	SCE $\hat{y}$	k' – 1	CM $\hat{y}$ = SCE $\hat{y}$ / (k' – 1)	CM $\hat{y}$ / CM e
Résiduelle	SCE e = Σ ei^2	n - k'	CM e = SCE e/(n – k')	
Total	SCE y	n - 1		

(*) SCE y = SCE $\hat{y}$ + SCEe (**) ddl total = ddl de la régression + ddl résiduel

☞ **Astuces** pour les calculs : $\Sigma yi = \Sigma yi\hat{}$ et $\Sigma ei = 0$

Pour α fixé a priori, si F calculé > $F_{1-\alpha}$ (k' - 1 ; n – k') on rejette l'hypothèse nulle avec α % de risque d'erreur et on considère que le modèle est satisfaisant dans son ensemble. Dans le cas contraire on conserve H0, on n'a pas pu prouver que le modèle était satisfaisant dans son ensemble.

IV - 3 - Coefficient de détermination

On définit **un coefficient de détermination multiple** égal au rapport entre la variance expliquée par l'ensemble des régresseurs et la variance totale de Y

$$R^2 = SCE\,\hat{y} / SCE\,y$$

IV - 4 - Test de signification sur chaque paramètre βj

On va calculer la matrice Var B : la **matrice des variances covariances** selon le principe:

$$Var(B) = \sigma^2_\varepsilon (U'U)^{-1}$$

σ^2_ε = CMe = SCE e / (n – k') n – k' = ddl résiduel

var b_0	cov b_0b_1	cov b_0b_2	....	cov b_0b_j	....	cov b_0b_k
cov b_1b_0	var b_1	cov b_1b_2	....	....	....	cov b_1b_k
cov b_2b_0	cov b_2b_1	var b_2	....	....	....	cov b_2b_k
....	....	....	....	....	....	....
cov b_jb_0	....	....	....	var b_j	....	....
....	....	....	....	....	....	....
cov b_kb_0	....	covb_2b_k	....	....	....	var b_k

On pose: H0: $\beta j = 0$

H1 : $\beta j \neq 0$ effet du régresseur Uj (test bilatéral)

Critère statistique: t calculé = (bj – βj) / σ (bj) = bj / σ (bj)

On trouve la valeur de σ (bj) en prenant la racine de la variance du coefficient dont il est question dans la diagonale de la matrice des variances-covariances VarB.

Pour α fixé : si t calculé < $t_{1-\alpha/2}$ (n – k') on conserve l'hypothèse nulle, ce qui signifie que les variations de Uj n'expliquent pas de façon significative les variations de Y compte tenu de la valeur du risque de première espèce retenue.
Pour α non fixé : Si on rejette H0 on a un risque égal à la p-value de le faire à tort (pour la calculer il faut être dans la situation où la variable suit une loi Normale pour utiliser la table, ou alors, disposer d'Excel).

IV - 5 - Estimations par intervalle de confiance de chaque coefficient βj

On est donc en présence du modèle: $Y_i = \Sigma\beta_j U_{ij} + \beta_0 + \varepsilon_i$

où ε est une erreur aléatoire normale, de moyenne nulle et de variance $\sigma^2(\varepsilon)$, les βj sont les coefficients de la régression dans la population concernée et ils sont estimés à partir d'un échantillon.
Dans les n observations aléatoires indépendantes, les Uij sont des valeurs exactes et certaines des Uj, les yi sont des réalisations de la variable aléatoire Y. Il résulte que les bj sont des variables aléatoires qui suivent une loi de Student à (n – k') ddl avec E(bj) = βj
On peut ainsi déduire les intervalles de confiance bilatéraux pour chaque βj :

$$P\,[bj - t_{1-\alpha/2}\,(n-k') * \sigma\,(bj) < \beta j < bj + t_{1-\alpha/2}\,(n-k') * \sigma\,(bj)] = 1 - \alpha$$

Chaque intervalle a α% de risque de ne pas contenir la vraie valeur βj.

Y a α% de risque de ne pas varier d'une quantité comprise entre les limites de l'intervalle de βj quand Uj varie de +1 unité.

V - La régression pas à pas ascendante

→ DABORD on regarde les régresseurs significatifs
Ensuite on les intègre

V - 1 - La matrice des corrélations

Dans cette matrice on trouve tous les coefficients de corrélation entre toutes les variables du modèle prises 2 à 2. Elle nous permet de déterminer l'ordre d'introduction des régresseurs.

V - 2 - Coefficient de détermination

On définit **un coefficient de détermination multiple** égal au rapport entre la variance expliquée par l'ensemble des régresseurs et la variance totale de Y

$$R^2 = SCE\,\hat{y} \,/\, SCE\,y$$

On peut définir le coefficient de détermination **partiel** de chaque régresseur Uj qui permet d'évaluer la contribution de chacun, selon son ordre d'introduction.

FACE, on calcule !

VI - Exemple- Régression linéaire multiple : Modèle complet

matrice des effets U

Matrice 1 = matrice U			
1	84	9	7
1	71	6	4
1	73	5	3
1	78	9	11
1	69	5	0
1	81	11	6
1	68	5	2
1	71	8	5
1	80	9	11
1	75	7	5

matrice des réponses Y

Matrice 2 = mat Y
38,01
29,35
29,19
35,06
28,14
38,53
28,06
30,85
36,28
32,15

(matrice des effets)' U'

Matrice 3 = matrice U'									
1	1	1	1	1	1	1	1	1	1
84	71	73	78	69	81	68	71	80	75
9	6	5	9	5	11	5	8	9	7
7	4	3	11	0	6	2	5	11	5

matrice d'information

sommes des carrés des colonnes

Matrice 4 = produit U'U			
10	750	74	54
750	56522 *1	5638 *2	4181
74	5638	588	451 *4
54	4181	451	406

matrice inverse (se trouve sur excel, donné en exam SAU si la matrice d'info a des zéros dans la diagonale et des 0 autour)

Matrice 5 = matrice $(U'U)^{-1}$			
51,882	-0,829	1,108	0,404
-0,829	0,014	-0,023	-0,005
1,108	-0,023	0,097	-0,017
0,404	-0,005	-0,017	0,022

Matrice 6 = produit U'Y		Matrice 7 = matrice B
325,62	← comme ligne de 1 sur U', 325,62 = somme des y	**-7,887**
24616,91 →		**0,456**
2482,88		**0,878**
1858,08		**-0,045**

somme des produits $u_2 \times y$: $84 \times 38{,}01 + 71 \times 29$ etc

$B = (U'U)^{-1}(U'Y)$

matrice 7 = matrice 5 × matrice 6

	y observé	y estimé	ei
	38,01	38,00 *3	0,012
	29,35	29,57	-0,222
	29,19	29,65	-0,460
	35,06	35,08	-0,024
	28,14	27,96	0,179
	38,53	38,43	0,098
	28,06	27,42	0,644
	30,85	31,28	-0,433
	36,28	36,00	0,284
	32,15	32,23	-0,079
SCE	149,971	148,978	0,993

Tableau 1 : Listes des réponses estimées et des résidus

*3 $\hat{y} = -7{,}887 + 0{,}456 \times 84 + 0{,}878 \times 9 - 0{,}045 \times 7$
$= 38{,}004$

*1 $56522 = 84 \times 84 + 71^2 + 73^2 + 78^2 + 69^2 + 81^2 + 68^2 + 71^2 + 80^2 + 75^2$

*2 $451 = 9 \times 7 + 6 \times 4 + 5 \times 3 + 9 \times 11 + 5 \times 0 + 11 \times 6 + 5 \times 2 + 8 \times 5 + 9 \times 11 + 7 \times 5$

PILE, on vérifie !

ANOVA				
	SCE	ddl	CM	F
régression	148,978	3	49,66	300
aléatoire	0,993	6	0,166	
totale	149,971	9		
p-value	6 E-07			
R^2	0,9934			

$F_{théo} = F_{1-\alpha}$

(et pas $F_{\alpha-1}$)

Tableau 2 = Test de validité du modèle

$(U'U)^{-1}$: Cov

Matrice 8 = Var B			
8,612	-0,138	0,184	0,067
-0,138	**0,002**	-0,004	-0,001
0,184	-0,004	**0,016**	-0,003
0,067	-0,001	-0,003	**0,004**

	Colonne 1	Colonne 2	Colonne 3	
val abs bj	**σ(bj)**	**t calculé**	**p-value**	
7,887	2,935 *	2,69	0,0362	*
0,456	0,048 *	9,56 *	0,0001	***
0,878	0,127	6,93	0,0004	***
0,045	0,061	0,73	0,4908	NS

$* = \sqrt{8,612}$

$* = \sqrt{0,002}$

$* = \frac{0,456}{0,048}$

Tableau 3 : Tests de significativité des coefficients

↑ on utilise les $t_{théo}$ des tables

(matrice des corrélations)

sinon loi $St(ddl_e)$

Matrice 9 = Matrice des R^2			
r^2	U1	U2	U3
Y	0,9361	0,8866	0,5797
U1	1	0,705	0,5515
U2	0,7047	1	0,5716
U3	0,5515	0,5716	1

← le + grand Y → premier U entré

$: 2Y sont pareils, on prend le moins corrélé...

Le fact. exp. de U_3 : on fait n'est pas significatif ici, dans le domaine étudié.
→ on peut simplifier le modèle

VII - Exemple- Régression linéaire multiple : Méthode ascendante

	U1	U2	U3	U4	U5	Y
1	90	210	4	9	8	30
2	90	210	4	5	12	40
3	90	210	18	5	8	60
4	90	210	18	9	12	60
5	90	290	4	5	8	10
6	90	290	4	9	12	10
7	90	290	18	9	8	20
8	90	290	18	5	12	40
9	100	250	11	7	10	30
10	100	250	11	7	10	20
11	110	210	4	5	8	40
12	110	210	4	9	12	30
13	110	210	18	9	8	50
14	110	210	18	5	12	60
15	110	290	4	9	8	10
16	110	290	4	5	12	10
17	110	290	18	5	8	20
18	110	290	18	9	12	20

Tableau des données

	U1	U2	U3	U4	U5	Y
U1	1					
U2	0	1				
U3	0	0	1			
U4	0	0	0	1		
U5	0	0	0	0	1	
Y	-0,102	-0,784 *	0,511	-0,170	0,102	1

Matrice des corrélations

Ordre d'introduction	U2	U3	U4	U5	U1

(1) Introduction de U2 : Y^ = 120.955 – 0.359 U2

source de variation	SCE	ddl	CM	F calc	F théo(5%)	R^2
régression (U2)	3306,250	1	3306,25	25,54	4,49	0,6148
résiduelle	2071,528	16	129,47			
totale	5377,778	17				

Tableau de l'ANOVA : test de validation du modèle

	Coefficients	Erreur-type	Statistique t			
Constante	120,955	17,980	6,727	***	$t_{0,975}(16)$	2,12
U2	-0,359	0,071	-5,053	***	$t_{0,995}(16)$	2,921
					$t_{0,9995}(16)$	4,015

Tableau des tests sur les coefficients

(2) **Puis introduction de U3** : Y^ = 106.223 – 0.359 U2+1.339 U3

parce que la covariance entre U_2 et U_3 est nulle, on a une somme ici

87% des variations de Y sont expliquées par le modèle avec U_2 et U_3 (1er deg

source de variation	SCE	ddl	CM	F calc	F théo (5%)	R^2	
régression	4712,500	2	2356,25	53,13	3,68	0,8763	multiple
U2	3306,25	1	3306,25	74,55	4,54	0,6148	partiels
U3/U2	1406,250	1	1406,25	31,71	4,54	0,2615	
résiduelle	665,278	15	44,35				
totale	5377,778	17					

Tableau de l'ANOVA : test de validation du modèle

64% des variations de y sont expliquées par U_2

	Coefficients	Erreur-type	Statistique t			
Constante	106,223	10,844	9,796	***	$t_{0,975}(15)$	2,131
U2	-0,359	0,042	-8,634	***	$t_{0,995}(15)$	2,947
U3	1,339	0,238	5,631	***	$t_{0,9995}(15)$	4,073

Tableau des tests sur les coefficients

(3) **Puis introduction de U4** : Y^ = 117.16 – 0.359 U2+1.339 U3 – 1.562 U4

U4 est-il significatif ? Pour le savoir, soit on compare les F ($F_{calc} < F_{théo}$ → NS) soit on teste t (voir ci

source de variation	SCE	ddl	CM	F calc	F théo (5%)	R^2	
régression	4868,750	3	1622,92	44,64	3,34	0,9053	multiple
U2	3306,25	1	3306,25	90,93	4,60	0,6148	
U3/U2	1406,25	1	1406,25	38,68	4,60	0,2615	partiels
U4/U3/U2	156,250	1	156,25	4,30	4,60	0,0291	
résiduelle	509,028	14	36,36				
totale	5377,778	17					

Tableau de l'ANOVA : test de validation du modèle

	Coefficients	Erreur-type	Statistique t			
Constante	117,160	11,146	10,511	***	$t_{0,975}(14)$	2,145
U2	-0,359	0,038	-9,536	***	$t_{0,995}(14)$	2,977
U3	1,339	0,215	6,219	***	$t_{0,9995}(14)$	4,14
U4	-1,562	0,754	-2,073	NS		

Tableau des tests sur les coefficients

$2,073^2 = 4,30$ soit le F_{calc} !

Le modèle que l'on peut conserver est : Y^ = 106,223 - 0,359 U2 + 1,339 U3

isaralyon
Une école d'ingénieurs au coeur de la vie

ISARA – 2ème année – STATISTIQUE
Exercices : Partie 7
Mme B. Bottollier-Lemallaz

REGRESSION LINEAIRE MULTIPLE

Exercice 1

Pour expliquer les variations d'acidité d'une tomme, on étudie le pH Y en fonction de 3 paramètres U1, U2 et U3, tels que.

Code pour Xj	U1	U2	U3
-1	Lait de chèvre	Présure 10 mg/l	Dose de ferment 1%
0	Mélange Chèvre + vache	Présure 11 mg/l	Dose de ferment 3%
+1	Lait de vache	Présure 12 mg/l	Dose de ferment 5%

Un des facteurs étant qualitatif, et pour faciliter l'ensemble des calculs, on décide de traduire les conditions expérimentales en variables naturelles (notées Uj) en conditions expérimentales en variables codées (notées Xj)
La matrice des essais et celle des résultats obtenus sont :

N°	X1	X2	X3	pH
1	-1	-1	-1	6,6
2	0	0	0	6
3	-1	1	1	5,4
4	0	0	0	5,8
5	-1	1	-1	6,2
6	0	0	0	6,1
7	-1	-1	1	5,6
8	0	0	0	5,7
9	1	-1	1	5,9
10	1	1	-1	6,3
11	1	-1	-1	6,5
12	1	1	1	5,9

Analyser les résultats (calculs des coefficients du modèle, validation du modèle, R^2 total, tests sur les coefficients, intervalles d'estimation des coefficients) selon un modèle de régression linéaire multiple complet. (NB : Si vous réfléchissez bien aux propriétés des matrices, il n'est pas utile d'avoir le logiciel Excel pour réaliser les calculs matriciels de cet exercice !!!)

Exercice 2 : (A faire sur Excel. Cf conseils dernière page de ce TD)

Le but de cet exercice est de savoir si le montant des salaires de 10 employés d'une entreprise suit un modèle linéaire multiple fonction du résultat au test d'embauche, du nombre d'années d'expériences et de la note mise par le supérieur hiérarchique. Pour cela vous déterminerez l'équation du modèle, vous établirez le tableau de

l'analyse de la variance et vous effectuerez les tests de signification sur les coefficients de la régression en réalisant une régression pas à pas ascendante. Calculer et interpréter tous les coefficients de détermination.

FICHE N°	résultats au test d'embauche (note / 10)	nombre d'années d'expérience	évaluation du supérieur (note / 10)	salaire annuel ($000)
	U1	U2	U3	Y
1	1	2	1	22,2
2	7,5	15	7	52
3	3,5	7	5	33,7
4	10	20	8	63,5
5	10	19	5	59,4
6	6	12	5	44,7
7	9,5	21	10	65,1
8	4	8	5	35,8
9	3,5	6	5	33,1
10	3,5	7	5	34,2

Exercice 3 (en cours)

Un bureau conseil en ressources humaines a effectué une étude sur le niveau d'anxiété Y, mesuré sur une échelle de 1 à 50, de 22 cadres d'entreprises au cours d'une période de 2 semaines. On veut examiner si les facteurs suivants peuvent influencer sur le niveau d'anxiété des cadres :

U1 : pression artérielle systolique

U2 : score au test évaluant les capacités managériales

U3 : niveau de satisfaction du poste occupé (mesuré sur une échelle de 1 à 25)

1°) Compléter le tableau de l'analyse de la variance ci-après qui indique l'apport de chaque variable introduite dans l'ordre indiqué (on supposera que toutes les conditions sont remplies pour avoir le droit d'analyser sur un plan statistique) :

CM/CMR

source de variation	SCE	ddl	CM	F	R^2 → SCEŷ/SCEy
Régression, $U_1U_2U_3$	1 300,989	3	433,663	17,64	0,7463
U_1	981.326	1	981,326	39,94	0,5639
U_2 / U_1	190.232	1	190,232	7,74	0,1091
U_3 / U_2 / U_1	129.431	1	129,431	5,27	0,0742
résiduelle	442,292	18	24,5718 ←		
totale	1743.281	21			

et $F_{0,95}(1;18) = 4,41$

Interpréter les résultats.

2°) Tester les hypothèses suivantes au seuil 0.05 :

a) $H_0 : \beta_1 = 0$ dans le modèle $Y = \beta_0 + \beta_1 U_1 + \varepsilon$

b) $H_0 : \beta_2 = 0$ dans le modèle $Y = \beta_0 + \beta_1 U_1 + \beta_2 U_2 + \varepsilon$

c) $H_0 : \beta_3 = 0$ dans le modèle $Y = \beta_0 + \beta_1 U_1 + \beta_2 U_2 + \beta_3 U_3 + \varepsilon$

Quel modèle retenez-vous ?

ANNALES Examen -Juin 2010 → savoir faire en 1h

L'entreprise SINTRON fabrique un matériau en matière plastique qui est utilisé dans la fabrication d'un outil agricole. Le département de contrôle de la qualité de l'entreprise a effectué une étude qui avait pour but d'établir dans quelle mesure la résistance à la rupture (en kg/cm²) de ce matériau pouvait être affectée par son épaisseur ainsi que par sa densité. Douze essais ont été effectués et les résultats sont présentés dans le tableau ci-dessous :

Essai N°	Résistance y	Epaisseur	Densité
1	37.8	4	4
2	22.5	4	3.6
3	17.1	3	3.1
4	10.8	2	3.2
5	7.2	1	3.0
6	42.3	6	3.8
7	30.2	4	3.8
8	19.4	4	2.9
9	14.8	1	3.8
10	9.5	1	2.8
11	32.4	3	3.4
12	21.6	4	2.8
somme	265.600	37.000	40.200
SC	7299,280	141.000	136,780

Vous allez réaliser l'étude de la régression linéaire multiple selon le modèle complet.
Les calculs sur ordinateur ont donné 4 matrices que vous devez identifier et compléter (2pts).

Matrice (U'U) → informations

12	37	40,2
37	141	126,7
40,2	126,7	136,780

Matrice (U'Y) →

265,600
981,40
926,61

Matrice (U'U)⁻¹ → inverse

5,4726	0,0550	-1,6593
0,0550	0,0429	-0,0559
-1,6593	-0,0559	0,5467

Matrice B, des coef

-30,0809
4,9047
11,0721

β_1 → à ce point 0
β_2 → si épaisseur ↑ 1 alors y ↑ de 4,9047
β_3 → si densité ↑ 1 alors y ↑ de 11,0721

Var B = $(U'U)^{-1} \times CM_e$

1°) (2 pts) A quel modèle correspond cette étude ? Interpréter <u>concrètement</u> chaque terme.
2°) (2 pts) Compléter le tableau suivant :

N°	Y^i	ei
1	33,8262	3,9738
2	29,3973	-6,8973
3	18,9566	-1,8566
4	15,1591	-4,3591
5	8,0400	-0,8400
6	41,4212	0,8788
7	31,6117	-1,4117
8	21,6469	-2,2469
9	16,8976	-2,0976
10	5,8256	3,6744
11	22,2782	10,1218
12	20,5397	1,0603
somme	265,60	0
SC	7083,4709	216,80308

Mais pour faire une ANOVA, il faut vérifier l'homoscédasticité des variances. Or, not possible ici, selon mes calculs...

3°) (5 pts) Tester, pour un niveau de confiance égal à 0.95, la validité du modèle. Quelle est votre conclusion ? → ANOVA (Fakile! SCE, ddl, CM, F_c, F_t → $\frac{CM_E}{CM_R} \gg 1$ donc et $F_c \gg F_t$ → modèle OK
4°) (1.5 pts) Quel pourcentage de variation dans la résistance à la rupture est expliqué par le modèle ? $R^2 = \frac{SCE_Y}{SCE_Y} = 0,848$ → 84,8% des variations expliquées.
5°) (6 pts) Sous forme de tableau tester la significativités des effets des facteurs. Au 1/1000ème, donner un intervalle d'estimation de ces effets pour un risque de 1%. Quelle est votre conclusion ?
6°) (1.5 pts) D'après le tableau des corrélations ci-dessous, dans quel ordre introduiriez-vous les régresseurs dans la méthode pas à pas ascendante ? Justifier.

	Résistance	Epaisseur	Densité
Résistance	1		
Epaisseur	0,8308	1	
Densité	0,6730	0,3649	1

→ d'abord l'épaisseur, car sa corrélation est + importante

5) b_j	$\sigma(b_j)$	t_c	Signif'	loi la dépassé α : 0,00 ?	lim inf	
4,9047	1,0142	4,838	***		1,61	8,20
11,0721	3,6206	3,058	*	loi il a pas atteint $\alpha = 0,01$	-0,70	22,83

($\sqrt{U'U} \times CM_E$)

↑ "à chaque fois qu'on dépasse, on ajoute une étoile"

α	$t_{1-\frac{\alpha}{2}}(ddl_e)$
0,05	2,262
0,01	3,2500
0,001	4,781

← loi St

Examen- Juin 2011 (durée 1h)

Exercice d'application (30 min) :

Pour optimiser les quantités de carpes pêchées pendant la période estivale dans 9 étangs de la Dombes, un auteur étudie trois facteurs pouvant influencer les rendements des pêches. Le premier facteur est l'orientation (sinus d'un angle) de l'entrée de la zone de piégeage. Le deuxième facteur est un rapport entre la hauteur d'eau en sortie dans la zone de piégeage et celle de l'entrée. Pour finir, le dernier facteur est la charge initiale de carpes rapportées à l'hectare.

La réponse mesurée Y représente la quantité de carpes pêchées lors de la première pêche estivale, on admettra qu'elle suit une loi normale. Le plan d'expérience utilisé ainsi que les résultats observés sont donnés dans les tableaux ci-dessous.

N°	U1	U2	U3	Y
Etang	Orientation	Sortie/entrée	Charge	Poids pêché (kg)
1	-0,96	0,4	187,5	528
2	0,65	0,1	180	298
3	-0,87	0,21	250	447
4	0	0,62	242,86	1287
5	-0,13	0,43	130,77	760
6	-0,3	0,48	225	501
7	0,58	0,07	133,33	17
8	-0,87	0,33	193,55	350
9	0,89	0,5	666,67	806
Somme	***-1,01***	***3,14***	***2209,68***	***4994,00***
SC	***4,0933***	***1,3776***	***756450,4028***	***3824792,00***

1) Les 4 matrices calculées pour déterminer les coefficients du modèle sont dans l'annexe 1 ; A vous de les compléter et de reconnaitre ces matrices pour écrire l'équation du modèle retenu :

.....

2) Sous forme d'un tableau, valider la pertinence du modèle en complétant et en vous aidant des résultats de l'annexe 2. On sait que la p-value du test = 0.0375.
Quelles sont vos hypothèses et conclusions ?

3) Discutez la significativité du coefficient du facteur U2. Justifiez.

4) Sachant que l'analyse entière a montré que les facteurs U1 et U3 n'ont pas montré des effets significatifs, quelle réponse apporteriez- vous à cette étude ?

Annexe 1

$(O'O)^{-1}$ inverse

0,65020714	0,081195025	-0,889011597	-0,000895316
0,08119502	0,351047959	0,474902839	-0,000845096
-0,88901160	0,474902839	4,824997036	-0,003018416
-0,00089532	-0,000845096	-0,003018416	0,000007550

O'Y

4339,00
528,47
2219,780
1396406,150

O'O info

9,0000	-1,0100	3,1400	2209,6800
-1,0100	4,0933	-0,5031	137,2791
3,1400	-0,5031	1,3776	866,8439
2209,6800	137,2791	866,8439	756450,4028

B

-119,402609
94,052
1804,801
-0,183

← OK!!!

Annexe 2

	Y	Y^	e
	528	577,9084	-49,9084
	298	189,2629	108,7359
	447	232,1335	214,8665
	1287	1055,2308	231,7692
	760	720,5551	39,4449
	501	777,6036	-276,6036
	17	137,1365	-120,1365
	350	459,0152	-109,0152
	806	844,9901	-38,9901
somme	4994	4393,22749	0,1685
SC	***3824792,00***	***3604699,4810***	
SCE	1053 646,889	833 564,2169	219 818,1921

U1 U2 U3

0,96 0,4 187,5

0,65 0,1 180

0,87 0,11 250 } → je retrouve 232 !!

← épargnez-moi les additions quoi

ANOVA

	SCE	ddl	CM
$\hat{Y}$		3	
e		5	
Y		8	

GUIDE DE TRAVAIL SUR EXCEL

(à suivre mot à mot)

Pour nommer une cellule ou une plage :

Sélectionner la ou les cellules de nombres.

Formule - Définir un nom

Tapez le nom de votre choix (pas de caractères spéciaux, ni accent, ni lettre unique...) et .....**VALIDEZ !**

Matrice X à saisir (avec sa colonne de 1), la nommer : matx.
Matrice Y à saisir la nommer : maty.

Remarque : pour supprimer un nom aller dans : Formules - Noms définis- Gestionnaire de noms

Pour déterminer la transposée X' d'une matrice X

Préparer la zone d'accueil de la matrice X' transposée de X, en sélectionnant une plage de cellules vides, les dimensions doivent être exactes.

Formule - Insertion Fonction - Recherche et Matrice - Transpose

Tapez : matx - CTRL SHIFT ENTER - nommer X' : tmatx

Pour déterminer un produit matriciel X'X

Préparer la zone d'accueil du produit X'X en sélectionnant une plage de cellules vides à la dimension exacte.

Formule - Insertion Fonction - Math et Trigo - Produitmat

Mat 1 = tmatx Mat2 = matx
CTRL SHIFT ENTER nommer X'X : prod
Remarque : même procédure pour le calcul de X'Y et de $(X'X)^{-1}$ *(X'Y)

Détermination d'inverse matriciel $(X'X)^{-1}$

Préparer la zone d'accueil de l'inverse du produit X'X en sélectionnant une plage de cellules vides à la dimension exacte.

Formule - Insertion Fonction - Math et Trigo - Inversemat

Matrice = Prod
CTRL SHIFT ENTER nommer $(X'X)^{-1}$: inv

Détermination de la liste des Y estimés.

Préparer la zone d'accueil de la liste des Y estimés en sélectionnant une plage de cellules vides à la dimension exacte à coté de la colonne des Y.

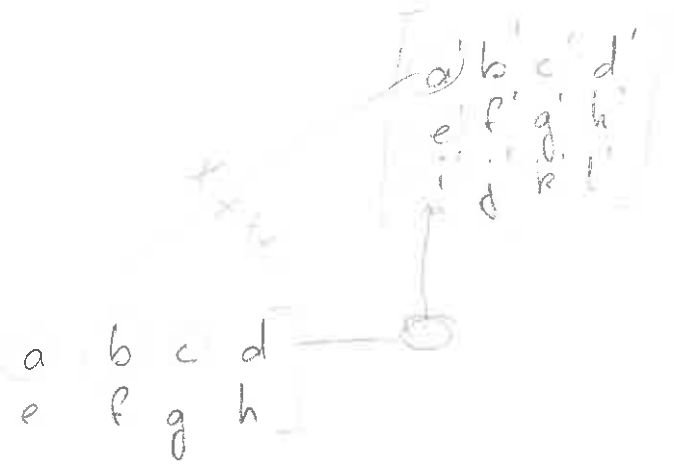
Formule - Insertion Fonction - Statistique - Tendance

Y connu = Y
X connu = sélectionner la plage des régresseurs sans la colonne de 1
X nouveau = (ne rien mettre)
Constante = vrai
CTRL SHIFT ENTER

Calculs divers:

Dans : « Formule - Insertion Fonction - Statistique » vous trouverez : MOYENNE, SOMME.CARRES.ECARTS ; ECARTTYPEP, VARP, ...etc ainsi que les Loi Normale, Loi Student, Loi F, Loi KHI DEUX pour déterminer les p-alues des tests.

Pour le calcul de VarB il s'agit de la multiplication algébrique d'une constante par une matrice.


a b c d
e f g h
a b c d
e f g h

Régression linéaire multiple : Méthode ascendante
(avec l'exemple du cours)

① Dresser la matrice des corrélations

Avec les coef. de détermination : $R^2 = \frac{cov_{xy}}{s_x s_y}$ $\quad cov_{xy} = \frac{SPE_{xy}}{n}$

Pour calculer le SPE_{xy} il faut : - calculer les écarts $x_j - \bar{x} = e_{x_j}$
- calculer les écarts $y_j - \bar{y} = e_{y_j}$
- $SPE_{xy} = e_{x_1} \times e_{y_1} + e_{x_2} \times e_{y_2} \dots + e_{x_n} \times e_{y_n}$

Dans l'exemple : * $SPE_{xy} = (-1) \cdot (-1) + (-10) \cdot \dots + (+10) \cdot \dots = -306$

* $cov_{xy} = \frac{SPE_{xy}}{n} = \frac{-306}{18} = -16,667$

* $s_x = \sqrt{\frac{SCE_x}{n}} = 3,628$ $\quad s_y = \sqrt{\frac{5291,771}{18}} = 17,285$

☆ $R^2 = \frac{-16,667}{3,628 \times 17,285} = -0,532$

NB : une covariance nulle entre 2 facteur indique que ces 2 facteurs sont indépendants.

② Choisir l'ordre d'introduction des facteurs

→ on introduit en priorité celui dont le R^2 avec y est le plus grand en valeur absolue

↳ si 2 facteur on le même R^2 en valeur absolue, on choisit en priorité le moins corrélé avec les autres facteurs

③ Introduction du 1er facteur

Là je suppose qu'on doit faire le modèle complet pour calculer les β ... Dans la mesure du possible. (en économie, impossible de faire sans connaissances et sans Excel)

Donc on obtient un modèle de type $\hat{y} = \beta_0 + \beta_1 U_1$

↳ on teste ce modèle avec l'ANOVA

↳ on teste le coefficient

④ Introduction du facteur suivant

C'est pareil que ③, sauf que là on testera les coefficients.

→ Cette étape est à répéter jusqu'à ce que le dernier coef. testé soit non significatif. Alors on arrête tout et on garde le modèle sans le coef. NS.

15/05/2012

TD STATS
Régression linéaire multiple

Exercice 1

① On étudie l'acidité (pH) d'un fromage selon 3 paramètres :
- U_1 → qualitatif → lait de chèvre, vache ou mélange → on code −1, 0, 1
- U_2 → quantitatif → présure
- U_3 → " → dose de ferment

Donnée : la matrice expérimentale (NB : c'est bien la matrice, parce qu'elle est codée, matrice expérimentale ≠ plan expérimental)

Le modèle en variables codées sera $\hat{y} = \beta_0 + \sum_1^3 \beta_j X_j$

La matrice des effets est donc X
La matrice $B = (X'X)^{-1}X'Y$

Voilà ce qu'on doit trouver :
$\hat{y} = 6 + 0{,}1X_1 - 0{,}1X_2 - 0{,}35X_3$

Matrice U'

$$\begin{matrix} 1&1&1&1&1&1&1&1&1&1&1&1 \\ -1&0&-1&0&-1&0&-1&0&1&1&1&1 \\ -1&0&1&0&1&0&-1&0&-1&1&-1&1 \\ -1&0&1&0&-1&0&1&0&1&-1&-1&1 \end{matrix}$$

Matrice U'U

$$\begin{matrix} 12&0&0&0 \\ 0&8&0&0 \\ 0&0&8&0 \\ 0&0&0&8 \end{matrix}$$

Matrice $(U'U)^{-1}$

$$\begin{matrix} \frac{1}{12}&0&0&0 \\ 0&\frac{1}{8}&0&0 \\ 0&0&\frac{1}{8}&0 \\ 0&0&0&\frac{1}{8} \end{matrix}$$

Matrice U'Y $\begin{bmatrix} 72 \\ 0{,}8 \\ -0{,}8 \\ -2{,}8 \end{bmatrix}$ Matrice B $\begin{bmatrix} 6 \\ 0{,}1 \\ -0{,}1 \\ -0{,}35 \end{bmatrix}$ ← oh oui ! !!

NB : les vecteurs étaient orthogonaux 2 à 2 donc la somme des produits vaut 0 (pas compris)

② ANOVA

$\sum y_i^2 = 433{,}42$ → $SCE_y = 433{,}42 - 12 \times 6^2 = 1{,}42$

$\bar{y} = 6$

$\hat{\bar{y}} = \bar{y} = 6$

poursuivi au verso

Pour la somme algébrique des X_i affectés des signes de code divisé par la somme au carré des X_j

Point culture : qd y a que des 0, on parle d'essais "au centre"

* Interprétation concrète

$\hat{y} = 6 + 0{,}1X_1 - 0{,}1X_2 - 0{,}35X_3$

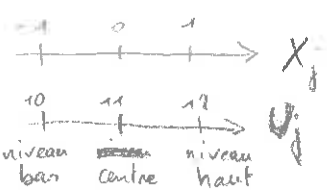
- X_1 : le pH ↗ de 0,1 qd U_1 ↗ de 1 donc le pH ↗ de 0,2 qd U_1 passe de chèvre à vache
- X_2 : quand X_2 ↗ de 1 unité (1mg) alors le pH ↘ de 0,1
- X_3 : le pH ↘ de 0,35 quand U_3 ↗ de 1, soit quand on augmente de 20 %

si on remplaçait "vache" par 10, "vache/chèvre" par 20, "chèvre" par 30) % de MG

On recherche le modèle avec les vraies valeurs

on doit trouver $\hat{y} = 7,425 - 0,01 U_1 - 0,1 U_2 - 0,175 U_3$

$X_j = \frac{U_j - U_{0j}}{\Delta U_j}$



$\Delta U_j = 12 - 11 = 1$

ΔU_j est le pas de variation

$X_2^- = \frac{10-11}{1} = -1$ $\quad X_3^- = \frac{1-3}{2} = -1$

$X_2^0 = \frac{11-11}{1} = 0$ $\quad X_3^0 = \frac{3-3}{1} = 0$

$X_2^+ = \frac{12-11}{1} = +1$ $\quad X_3^+ = \frac{5-3}{2} = +1$

Du coup, en remplaçant :

$$\hat{y} = 6 + 0,1\left(\frac{U_1 - 20}{10}\right) - 0,1\left(\frac{U_2 - 11}{1}\right) - 0,35\left(\frac{U_3 - 3}{2}\right)$$

$$\hat{y} = 6 + 0,01(U_1 - 20) - 0,1(U_2 - 11) - 0,175(U_3 - 3)$$

$$\hat{y} = 6 + 0,01U_1 - 0,2 - 0,1U_2 + 1,1 - 0,175U_3 + 0,525$$

$$\hat{y} = 7,425 + 0,01U_1 - 0,1U_2 - 0,175U_3$$

le pH de 0,01 qd la MG ↗ de 1%

le pH ↘ de 0,1 qd U_2 ↗ de 1

le pH ↘ de 0,175 quand U_3 ↗ de 1%

↳ ici le pH β_0 n'a pas de sens parce qu'il ne sert à rien

② ANOVA

$SC_y = 433,14$ $\quad \Sigma y_i = 72$

les $\hat{y}_i$:

6,35
6
5,45
6
6,15
6
5,65
6
5,85
6,35
6,55
5,65

et $SC_y = 433,42$

$\Delta\theta$	SCE	ddl	CM	F_{calc}
Totale	1,42	12-1=11		
régression	1,14	4-1=3	0,38	10,86
Fluctu. nat.	0,28	11-3=8	0,035	

$F_{calc} < F_{théo}$, on rejete H_0 avec moins de 5% de risque d'erreur

③ Test de signification

$$Var(B) = \sigma_\varepsilon^2 (U'U)^{-1} = 0{,}035 \begin{bmatrix} \frac{1}{12} & 0 & 0 & 0 \\ 0 & \frac{1}{8} & 0 & 0 \\ 0 & 0 & \frac{1}{8} & 0 \\ 0 & 0 & 0 & \frac{1}{8} \end{bmatrix}$$

$$\sigma^2(b_4) = \sigma^2(b_2) = \sigma^2(b_3) = \frac{0{,}035}{8}$$

$$\sigma^2(b_4) = (0{,}066)^2$$

$H_0 : \beta_j = 0$ et $H_1 : \beta_j \neq 0$ et $t_{calc} = \frac{b_j}{\sigma(b_j)}$

on compare t_{calc} et $t_{théo}$ en valeur abs

b_j	$\sigma(b)$	t_{calc}		$t_{théo}(8)$		
0,1	0,066	1,51	0,05	2,306	NS	
-0,1	0,066	-1,51	0,01	3,355	NS	
-0,35	0,066	-5,29	0,001	5,041	***	

⇒ Seule la dose de ferment a un effet sur le pH.

⇓

on peut le simplifier

$P[-0{,}053 < \beta_1 < 0{,}253] = 0{,}95$

$P[-0{,}503 < \beta_2 < -0{,}197] = 0{,}95$

⇒ on a β_1