MODÈLE LINÉAIRE À m VARIABLES EXPLICATIVES

10.1

Généralités sur la régression

Pendant les cours précédents on s'est surtout intéressé à étudier des inférences sur des grandeurs statistiques issues des échantillons, eu égard aux mêmes grandeurs statistiques concernant les populations. Dans ce cours, on examinera la relation entre une variable et une autre ou plusieurs autres variables. Nous allons établir un modèle pour cette relation que l'on utilisera pour des buts de compréhension ou de prédiction. On notera par Y la v.a. à expliquer et par $X_1, \ldots, X_m$ les v.a. indépendantes qui seront utilisées par le modèle.

Le modèle mathématique aura la forme :

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_m X_m + b + e$$

où $e \sim (0, \sigma^2)$. Ce modèle s'appelle *régression multilinéaire* de Y en fonction de $X_1, \ldots, X_m$. Pour calculer cette régression il nous faut une série de n mesures ($n \gg m$) :

$$x_{i1}, x_{i2}, \ldots, x_{im} \ , \ y_i \quad ; \quad i = 1, 2, \ldots n$$

De plus on fait l'hypothèse selon laquelle les valeurs des variables X_i sont mesurées sans erreur.

Dans la suite nous noterons :

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} & 1 \\ x_{21} & x_{22} & \cdots & x_{2m} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} & 1 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \\ b \end{bmatrix}$$

$$\mathbf{x_1} = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix}, \ \cdots, \ \mathbf{x_m} = \begin{bmatrix} x_{1m} \\ x_{2m} \\ \vdots \\ x_{nm} \end{bmatrix}$$

$$m_y = \frac{1}{n}\sum_{i=1}^{n} y_i \,, \ m_{x_1} = \frac{1}{n}\sum_{i=1}^{n} x_{i1} \,, \ \cdots, \ m_{x_m} = \frac{1}{n}\sum_{i=1}^{n} x_{im}$$

10.2

Modèle linéaire

Le modèle linéaire sur l'ensemble des observations $[\mathbf{y} \ \mathbf{X}]$ s'exprime par la relation :

$$\mathbf{y} = \mathbf{X} \cdot \mathbf{a} + \mathbf{e}$$

avec $\mathbf{a} = [a_1, \ldots, a_m, b]^\top$. Le problème de la régression est l'estimation des valeurs des composantes du vecteur $\mathbf{a}$ de sorte que $\mathbf{e}^T\mathbf{e} = \sum_{i=1}^{n} e_i^2$ soit minimal.

La solution est donnée par :

$$\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

où

$$\mathbf{X}^+ = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

la pseudoînverse de $\mathbf{X}$.

Pour les calculs on aura besoin de la matrice des variances-covariances :

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{YY} & \mathbf{V}_{XY}^T \\ \mathbf{V}_{XY} & \mathbf{V}_{XX} \end{bmatrix}$$

10.3

Propriétés de l'estimation

On peut démonter que $\hat{\mathbf{a}}$ est un estimateur sans biais, i.e. $E(\hat{\mathbf{a}}) = \mathbf{a}$.
La variance de $\hat{\mathbf{a}}$ est :

$$V(\hat{\mathbf{a}}) = s^2(\mathbf{X}^T\mathbf{X})^{-1}$$

avec

$$s^2 = \frac{1}{n-m-1}\sum_{i=1}^{n} e_i^2 = \frac{n}{n-m-1}V(\mathbf{e})$$

où $V(\mathbf{e})$ est la variance de $\mathbf{e}$.

Pour le terme constant nous avons :

$$V(b) = \frac{s^2}{n}\left[1 + \mathbf{m}_X^T\, V_{XX}^{-1}\, \mathbf{m}_X\right]$$

Nous avons aussi $\hat{\mathbf{a}} - \mathbf{a} = (X^TX)^{-1}\, X^T\mathbf{e}$.

10.4

Coefficient de corrélation multiple

C'est le rapport

$$R^2 = \begin{cases} \dfrac{\text{variance expliquée par la regression}}{\text{variance totale}} \\ \dfrac{\text{variance totale - variance résiduelle}}{\text{variance totale}} \end{cases}$$

Nous avons :

$$R^2 = \frac{\mathbf{V}_{XY}^T\, \mathbf{V}_{XX}^{-1}\, \mathbf{V}_{XY}}{\mathbf{V}_{YY}}$$

D'après la définition de R^2, nous avons que

$$1 - R^2 = \frac{V(\mathbf{e})}{V(\mathbf{y})}$$

et, par conséquent :

$$s^2 = \frac{n}{n-m-1}\, V(\mathbf{y})\, (1-R^2)$$

10.5 Tests de signification

Tests sur R^2

Si

$$\frac{n-m-1}{m} \cdot \frac{R^2}{1-R^2} \geq F_{\alpha;m,n-m-1}$$

alors il existe au moins un coefficient de regression qui est significatif.

Comparaison d'un coefficient à une valeur de référence

Soit $\widehat{a}_i$ l'estimation d'un coefficient du modèle linéaire. On compare ce coefficient à une valeur de référence a_i^0 (d'habitude on prend $a_i^0 = 0$, c-à-d. on cherche à savoir si $\widehat{a}_i$ est significativement différent de zéro). Si

$$\frac{\hat{a}_i - a_i^0}{s\sqrt{V_{XX}(i,i)}} = \frac{\hat{a}_i - a_i^0}{\sqrt{V(\hat{a}_i)}} \leq t_{\alpha;n-m-1}$$

alors $\hat{a}_i$ est proche de la valeur de référence a_i^0 avec confiance $1-\alpha$.

Intervalle de confiance pour les coefficients

Soit $\widehat{a}_i$ l'estimation d'un coefficient du modèle linéaire. L'intervalle de confiance pour cette estimation, avec confiance $1-\alpha$, est donné par

$$\mathrm{IdC}\,(\widehat{a}_i) = \left[\widehat{a}_i - t_{\alpha;n-m-1}\sqrt{V(\widehat{a}_i)}, \widehat{a}_i + t_{\alpha;n-m-1}\sqrt{V(\widehat{a}_i)}\right]$$

Test sur un ensemble des variables explicatives

Considérons les q premières variables $x_1, \cdots, x_q$ et soit $\mathbf{X}_q$ la matrice correspondante des données. Si

$$\frac{n-m-1}{m-q+1} \cdot \frac{\mathbf{y}^T\left[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T - \mathbf{X}_q(\mathbf{X}_q^T\mathbf{X}_q)^{-1}\mathbf{X}_q^T\right]\mathbf{y}}{\mathbf{y}^T\left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right]\mathbf{y}} \geq F_{\alpha;m-q+1,n-m-1}$$

alors toutes les variables (et pas seulement les q premières) sont significatives avec confiance $1-\alpha$.

10.6

Linéarisation des données

L'hypothèse fondamentale faite pour la regression linéaire est que le modèle qui décrit les données est linéaire. Dans le cas contraire il faut faire des transformations pour essayer d'obtenir des modèles linéaires. Cependant, toutes les données ne sont pas linéarisables.

Parmi les nombreuses équations que l'on emploie pour exprimer les relations entre les différentes variables, on peut mentionner quelques modèles dont la linéarisation est facile.

Fonction	**Transformation**	**Forme linéaire**	**Figure**
$y = \alpha x^{\beta}$	$y' = \log y,\ x' = \log x$	$y' = \log\alpha + \beta x'$	fig. 10.1 (a)
$y = \alpha e^{\beta x}$	$y' = \ln y$	$y' = \ln\alpha + \beta x$	fig. 10.1 (b)
$y = \alpha + \beta \log x$	$x' = \log x$	$y = \alpha + \beta x'$	fig. 10.1 (c)
$y = \dfrac{x}{\alpha x - \beta}$	$y' = \dfrac{1}{y},\ x' = \dfrac{1}{x}$	$y' = \alpha - \beta x'$	fig. 10.1 (e)
$y = \dfrac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$	$y' = \ln\left(\dfrac{y}{1-y}\right)$	$y' = \alpha + \beta x$	fig. 10.1 (e)

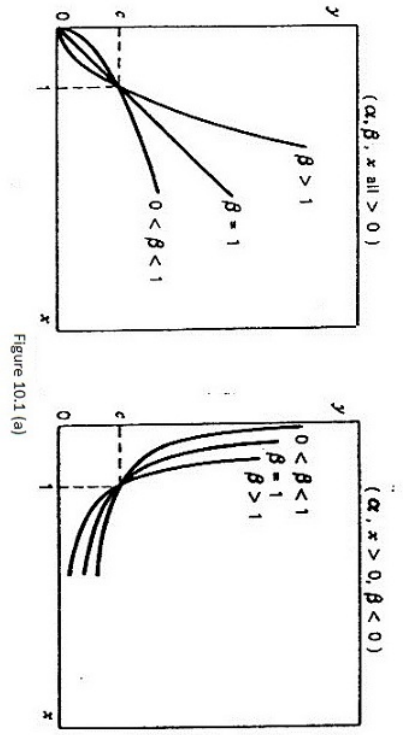

Figure 10.1 (a)

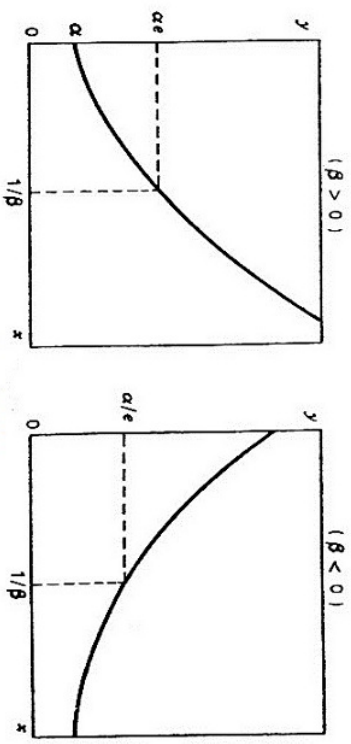

Figure 10.1 (b)

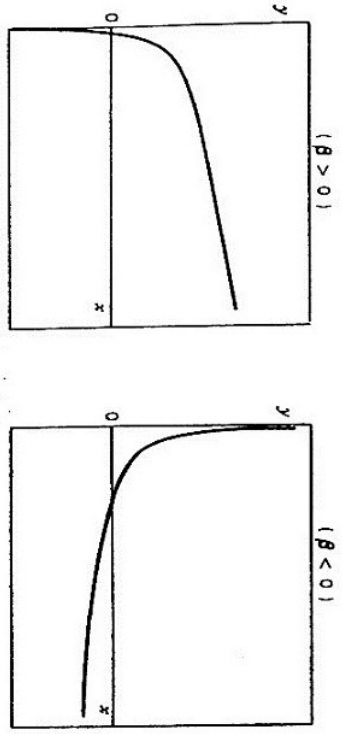

Figure 10.1 (c)

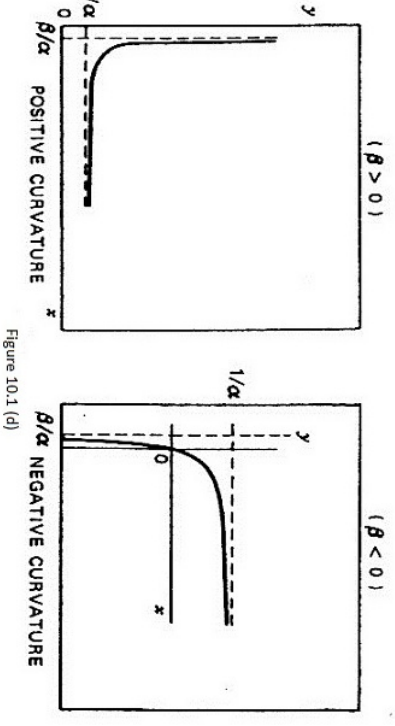

Figure 10.1 (d)

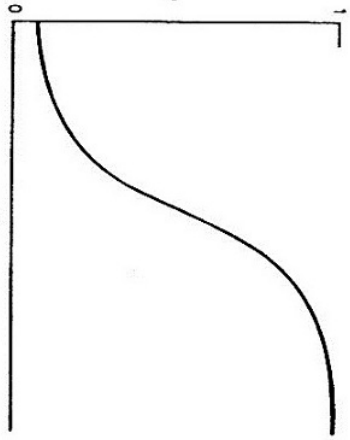

Figure 10.1 (e)

10.7

Détection des relations non linéaires

Pour détecter les non-linéarités on examine le graphique obtenu par les points $\widehat{y}$ et $(y-\widehat{y})$.

C'est le graphique des résidus.

Les résidus doivent être distribués suivant une loi normale de moyenne 0.

Quelques types des problèmes rencontrés sont présentés ci-après :

Description	**Figure**
Régression (modèle) valable	fig. 10.2 (a)
Variance non constante. Problèmes avec les grandes valeurs de y. Il faut appliquer une transformation pour y. Transformations utilisées : $y' = \sqrt{y}, y' = \log y, y' = \log(y+1)$	fig. 10.2 (b)
Variance non constante. Problèmes avec les petites valeurs de y. Il faut appliquer une transformation pour y. Transformations utilisées : $y' = \frac{1}{y}, y' = \frac{1}{y+1}$	fig. 10.2 (c)
Variance non constante. Situation fréquente quand y est un pourcentage entre 0% et 100%. Il faut appliquer une transformation pour y. Transformations utilisées : $y' = \log y, y' = \log(y+1)$	fig. 10.2 (d)
Variance non linéaire. Présence des relations non linéaires.	fig. 10.2 (e) et 10.2 (f)
Variance non linéaire et non constante. Présence des relations non linéaires.	fig. 10.2 (g) et 10.2 (h)

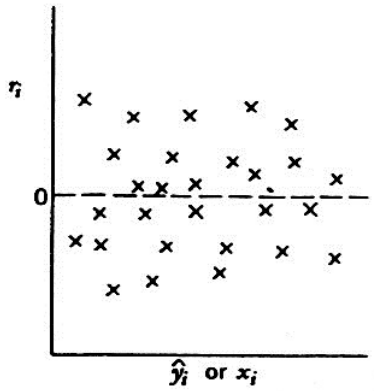

Figure 10.2 (a)

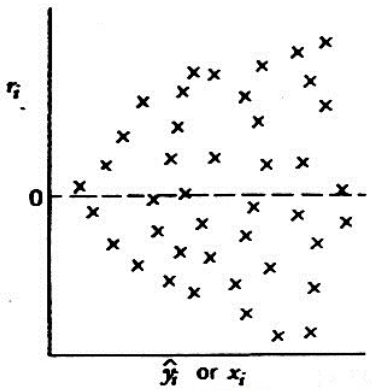

Figure 10.2 (b)

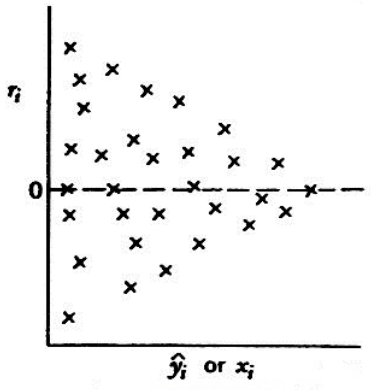

Figure 10.2 (c)

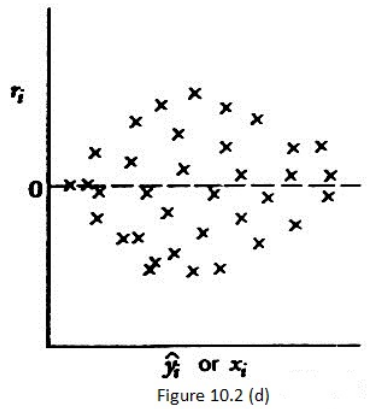

Figure 10.2 (d)

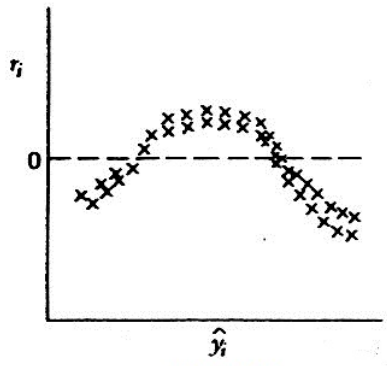

Figure 10.2 (e)

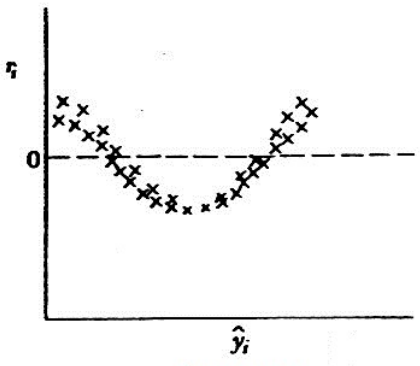

Figure 10.2 (f)

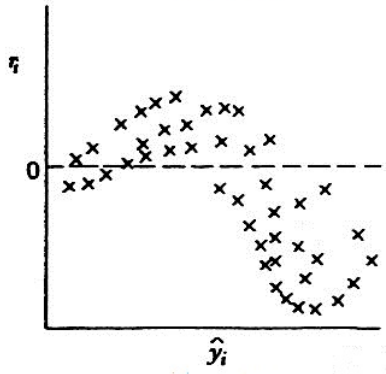

Figure 10.2 (g)

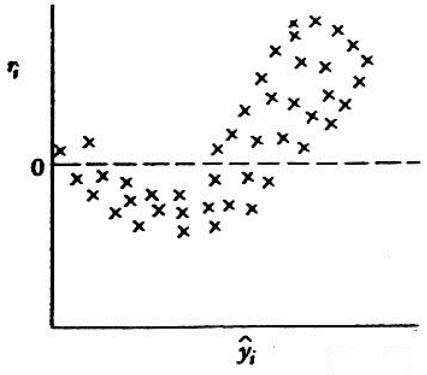

Figure 10.2 (h)

10.8

Co-linéarité entre les variables explicatives

Pour obtenir un modèle valable, on doit avoir toutes les hypothèses fondamentales vérifiées et aussi que les variables (X_i) explicatives doivent être indépendantes.

Dans le cas où les valeurs de deux ou plus variables explicatives sont liées entre elles on parle des problèmes de co-linéarité.

Détection de co-linéarité

On calcule la corrélation linéaire entre les variables explicatives , prises deux à deux, par :

$$\rho = \frac{cov(X_i, X_j)}{\sigma_{x_i}\sigma_{x_j}}$$

Exemple :

- Y, consommation des familles
- X_1, niveau de scolarité
- X_2, revenu des familles

Effets d'une forte co-linéarité entre les variables explicatives

(1) Les valeurs des coefficients de régression deviennent *très sensibles*. L'*ajout ou le retrait* des quelques observations provoquent des *grands changements*

(2) La précision de l'estimation des coefficients est très affectée. On constate que la variance des a_i, $S^2(a_i)$ sont très grandes.

(3) Le *test de Student* peut donner que les coefficients *ne sont pas significatifs* malgré que le test de *Fisher montre l'existence des coefficients significatifs*.

Méthodes pour réduire l'effet de co-linéarité

(1) On fait un choix judicieux des variables explicatives à inclure dans le modèle

(2) On regarde la matrice de corrélation des variables explicatives :
Si une co-linéarité existe entre deux variables, on supprime l'une de deux ou on fait la moyenne pondérée de deux.

(3) On utilise une méthode rigoureuse pour construire le modèle.

10.9

Méthodes pour construire une équation de régression multiple

On vise deux objectifs :

(1) Construire un modèle avec un R^2 élevé, des coefficients significatifs et S^2 faible.

(2) Obtenir un modèle facile à utiliser et économique.

Les méthodes sont :
- Toutes les régression possibles.
- L'introduction progressive des variables.
- L'élimination progressive des variables.
- La régression pas à pas.

Méthode pas à pas

Cette méthode consiste à introduire ou retrancher successivement, une à la fois, les variables explicatives selon le critère de Fisher (Test sur R^2)

Étapes

(1) On fait toutes les régressions à une variables. Y/X_i

(2) On choisit le modèle qui possède le plus grand F. Soit Y/X_k

(3) On fait ensuite les régressions à deux variables
$Y/X_k, X_i$, $i \neq k$ et on choisit le plus grand F.

(4) Soit $Y/X_k, X_l$ le modèle retenu ; On fait les modèles à trois variables, e.t.c.

(5) On arrête quand on a fini tester l'entrée des toutes les variables.

N.B. Á chaque étape on vérifie tous les critères, c.à.d :
- le test de Student.
- le test Fisher.
- les valeurs des coefficients (co-linéarité possible de la dernière variable entrée).

Si un problème existe, on retranche la dernière variable ou on décide de retrancher une autre.

10.10

Validité et la qualité de la régression linéaire

Modèle : $Y = a_1X_1 + ... + a_nX_n + a_0 + \varepsilon$

(1) **Hypothèses fondamentales**

(a) Le résidu $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

(b) X_i indépendantes entre elles

(c) σ^2 petit

(d) Terme constant a_0 petit

(2) **Validité**

(a) Pour la régression simple, $m = 1$, on commence par examiner le graphique (X, Y). Ce qui n'est pas évident à faire pour $m > 1$

(b) On observe le graphe des résidus, (voir annexe)

(3) **Pour un modèle valable on doit avoir tous les critères décrits ci-dessous dans le même sens**.

(a) R^2

(b) Test sur R^2 (sur la totalité des coeffs : Fischer)

(c) Test des coeffs individuel, (signification de chaque coeff : test de Student)

(d) Variance résiduelle

(e) Terme constant

(4) **Qualité de la régression : Bonne relation linéaire entre** Y **et les** X_i

(a) $R^2 \simeq 1$

(b) Test sur R^2 : Il doit avoir au moins un coeff significatif.

i.e. $\frac{n-m-1}{m} \frac{R^2}{1-R^2} \geq F_{\alpha;m,n-m-1}$

(c) Test individuel : a_i significatif si $\frac{|\widehat{a_i}-\alpha_i|}{\sqrt{V(\widehat{a_i})}} \geq t_{\frac{\alpha}{2};n-m-1}$

(5) **Attention :** La corrélation n'implique pas une relation de cause à effet. En effet si deux variables sont en correlation forte, elles peuvent être influencées par les variations d'une cause commune extérieure du modèle.
Deux exemples :
– Un auteur américain a démontré qu'il existait une forte correlation entre les salaires versés aux ministres de l'état de Massachusetts et les prix du rhum à la Havane ! ! !
– Un autre auteur a montré que les ventes de l'huile de bronzage sont en forte correlation avec les ventes des glaces ! ! !
Mais évidemment il n'y a aucune relation de causalité, il y a une cause climatique.
Par conséquent *on doit être très prudent dans l'interprétation des résultats d'une étude*.

(6) **Remarque :** On ne doit pas se sentir toujours obligé à aboutir à un modèle de bonne relation linéaire.

10.11

Exercices

EXERCICE 11.1 *On considère les variables Y_t et X_t représentant respectivement le prix moyen annuel du kilogramme de bananes dans la region parisienne et le prix du ticket de métro à l'unité en l' été de l'année t ; l'unité est le franc.*

t	1975	1976	1977	1978	1979	1980	1981	1982	1983	
X_t	2,20	2,50	2,70	3,00	3,60	4,50	5,00	5,50	6,00	
Y_t	3,72	3,95	4,27	4,52	5,01	5,41	6,17	7,03	8,42	

Le tableau ci-dessous présente les résultats des régressions effectuées.

Modèle						
1.	$\mathbf{X_t/t}$	$\widehat{b}$	$\widehat{a}$ $(\widehat{\sigma}(\widehat{a}))$	R^2	S^2	
		1.364	*0.505* *(0.032)*	*0.973*	*0.061*	
2.	$\mathbf{Y_t/t}$	$\widehat{\beta}$	$\widehat{\alpha}$ $(\widehat{\sigma}(\widehat{a}))$	R^2	S^2	
		2.66	*0.546* *(0.063)*	*0.973* *0.916*	*0.235*	
3.	$\mathbf{Y_t/X_t}$	$\widehat{\lambda}$	$\widehat{\mu}$ $(\widehat{\sigma}(\widehat{a}))$	R^2	S^2	
		1.197	*1.078* *(0.106)*	*0.937*	*0.176*	

Étude des modèles :
- $X_t = at + b + \varepsilon_t$
- $Y_t = \alpha t + \beta + u_t$

On suppose : $E(\varepsilon_t) = 0, E(u_t) = 0, V(\varepsilon_t) = \sigma_\varepsilon^2, V(\varepsilon_t) = \sigma_u^2$.

(1) Représenter graphiquement le nuage des points (1)

(2) Tester la validité des coefficients du modèle (2).

(3) Peut-on calculer l'erreur commise par la régression pour l'année 1979 ?

(4) Il parait que le ticket de métro coûtait $6,75$ l'année 1985. Pouvez-vous le confirmer ?

(5) On effectue la régression linéaire de Y_t en X_t . $Y_t = \lambda + \mu X_t + \eta_t$ avec $E(\eta_t) = 0, V(\eta_t) = \sigma_\eta^2$

(a) Tester la validité du modèle.

(b) Conclusion ?

EXERCICE 11.2 *À partir de relevés mensuels de janvier 1980 à août 1981, on cherche à établir un modèle du commerce extérieur permettant de déterminer la liaison des exportations Y aux importations en utilisant les variables explicatives suivantes :*

- X_1 *niveau relatif des prix franpar rapport à ceux de l'Union Européenne ;*
- X_2 *niveau de consommation des ménages ;*
- X_3 *indice de la production industrielle ;*
- X_4 *indice de la demande extérieure.*

Plusieurs modèles linéaires ont été testés. Les résultats sont donnés dans le tableau suivant :

Modèle	$\widehat{b}_0$	$\widehat{b}_1$ $(\widehat{\sigma}(\widehat{b}_1))$	$\widehat{b}_2$ $(\widehat{\sigma}(\widehat{b}_2))$	$\widehat{b}_3$ $(\widehat{\sigma}(\widehat{b}_3))$	$\widehat{b}_4$ $(\widehat{\sigma}(\widehat{b}_4))$	R^2
1	*0.876*	*-0.064 (0.031)*				*0.166*
2	*0.893*	*-0.062 (0.033)*	*-0.203 (1.131)*			*0.168*
3	*1.421*	*-0.057 (0.018)*	*-1.397 (0.635)*	*0.476 (0.072)*		*0.777*
4	*2.292*	*-0.061 (0.022)*	*-1.180 (0.911)*	*0.4810 (0.076)*	*0.004 (0.011)*	*0.781*
5	*0.991*			*0.426 (0.095)*		*0.517*

(1) Tester la validité des différents modèles proposés.

(2) Indiquer le modèle qui doit être retenu si on prend un intervalle de confiance de niveau 95% pour les paramètres.

10.12

Exercices à faire avec le script `add.sci`

EXERCICE 11.3 (DONNÉES ACIDE NITRIQUE) *L'acide nitrique est industriellement le plus important des dérivés oxygénés de l'azote, (utilisé dans l'industrie des engrais, le domaine des explosifs, industrie organique).*

Notons que la production industrielle d'acide nitrique en France est de l'ordre de 4 000 000 t.

Un problème important de la production de l'acide nitrique est que l'ammoniac a tendance à s'échapper de la colonne d'absorption. Pour des raisons aussi bien d'efficacité industrielle qu'écologiques on voudrait :

- *comprendre ce phénomène d'échappement*
- *prédire*
- *agir pour réduire le taux d'échappement*

Pour la compréhension du phénomène on pense associer **le taux d'échappement** *de l'ammoniac avec d'autres variables importantes de l'unité de production afin d'obtenir un modèle qui sera, en première approximation, linéaire.*

Les variables considérées sont aux nombre de trois :

– **flux d'air** *à la colonne d'absorption*

– **temperature de l'eau** *qui absorbe le peroxyde*

– *le pourcentage de la* **concentration de l'oxyde**

On effectue 17 mesures dans une unité de production à différentes périodes du temps qui sont dans le fichier `acidnitr.dta`.

Faire une analyse complète de la dépendance linéaire du taux d'échappement de l'ammoniac en fonction des autres variables ci-dessus.

EXERCICE 11.4 (DONNÉES RAT) *On a fait une experience pour expliquer le résidu d'un médicament présent dans le foie d'un rat. On prend au hasard dix-neuf rats que l'on pèse. Ensuite leur on administre des doses de ce médicament à raison de 40mg du médicament par kilogramme. Trois jours plus-tard ont tue les rats et on a mesuré le pourcentage du médicament qui restait encore dans le foie.*

Faire une régression linéaire pour décider si le résidu dans le foie peut être expliqué par les autres variables. Les variables mesurées sont dans l'ordre : le poids du rat, le poids de son foie, la dose du médicament administrée. Les données se trouvent dans le fichier `rat.dta`.

D'après le protocole de l'experimentation il ne doit pas y avoir une quelconque relation entre le pourcentage du médicament administré et les autres variables.

EXERCICE 11.5 (DONNÉES IMPORT) *Nous avons un tableau avec des données macroéconomiques annuelles de la France entre 1949 et 1966. Ces données sont : les importations, la production indigène, la valeur des stocks et la consommation des ménages, mesurées en milliards des francs. Les données se trouvent dans le fichier* `import.dta`.

– *Faire une regression linéaire pour expliquer le volume des importations en France à l'aide des autres variables.*

– *En fonction des résultats obtenus, proposer une amélioration possible de la qualité de la regression.*

Coefficients												
Régression (modèle)	a_1 $v(a_1)$	a_2 $v(a_2)$	a_3 v(a3)	b	S^2	R^2	D.D.I m, n-m-1	T calculé	$t_{a;n-m-1}$	$F_{a;m,n-m-1}$	F calculé	Commentaires