Chapitre 7 : Régression linéaire

M1 du Master MMAS

James Ledoux

Dépt de mathématiques, Univ. Poitiers

12 juillet 2009

La fin du cours proposent des méthodes qui combinent :

- un aspect descriptif
- un aspect modèle statistique
- un aspect inférentiel
- un aspect prédictif

la modélisation statistique

Introduction à la régression linéaire

1. Quelques compléments sur la régression linéaire simple
2. Régression linéaire multiple

Quelques références spécifiques

- Azaïs, J.-M et Bardet, J.-M.
  Le modèle linéaire par l'exemple : régression, analyse de la variance et plans d'expériences illustrés avec R, SAS et Splus
  Cote BU (519.23 AZA) ou
  un poly : http ://www.lsp.ups-tlse.fr/Azais/publi/modlin.pdf
- Cornillon, P.-A. et Matzner-Lober, E. (2007)
  Régression : théorie et pratique
  Springer
- (2001) Glantz , S. A. et Slinker , B. K.
  Primer of applied regression & analysis of variance
  Cote BU (519.2 STA)

- **Données :** une variable continue explicative x
  une variable continue à expliquer y

  Les deux variables sont observées sur n individus :

$$x_1, \dots, x_n,\ y_1, \dots y_n$$

- **Objectifs :**
  - **Descriptif :** exploration d'une liaison entre y et une variable potentiellement explicative x : ici : tendance linéaire
  - **Explicatif :** liaison théorique connue **a priori**
    ↪ validation à partir des observations
  - **Prédictif :** à partir d'une liaison établie entre x et y, **prédire** la valeur de y sur la base de la connaissance de la valeur de x pour tout nouvel individu
    Dans ce contexte, analogue à une méthode de discrimination comme l'AFD lorsque la variable à expliquer est nominale (prochain chapitre)

Descriptif : déjà fait au Chapitre 2

- **Modèle :**

$$\boxed{y = ax + b + e}$$

  où a, b sont des constantes inconnues
  e variable de bruit (inconnue)

- **Travail**
  - Évaluer a, b : $\hat{a}, \hat{b}$
    (ajustement du modèle aux données)
  - Déterminer la QLT de cet ajustement

Explicatif

- Idée de base : on suppose que les données sont relatives à un échantillon prélevé de manière aléatoire sur une population
  ↪ étudier la robustesse des estimations de a, b, e obtenues sur l'échantillon
- Une modélisation probabiliste :

$$Y = ax + b + E$$

$$\left.\begin{array}{r}E \text{ est une v.a. (bruit)}\\ x \text{ est déterministe}\end{array}\right] \Longrightarrow Y \text{ est une v.a.}$$

- $(x_i, y_i)_{i=1}^{i=n}$ donne n valeurs observées $(e_i, y_i)_{i=1}^{i=n}$ du couple (E, Y) i.e. les valeurs observées d'un n-échantillon $(E_i, Y_i)_{i=1}^{n}$ du couple
- Les quantités $\widehat{a}, \widehat{b}$ deviennent des v.a., fonctions de l'échantillon : estimateurs des paramètres a, b

↪ **Inférence sur le modèle**

Modèle probabiliste :

A1) x est déterministe

A2) E_i est centrée et de variance constante : homoscédasticité

$$i = 1, \dots, n \qquad \mathbb{E}[E_i] = 0, \quad \sigma^2(E_i) = \sigma_E^2$$

A3) a et b sont des constantes (pas de rupture du modèle)

A4) les v.a. E_i sont indépendantes (ou non-corrélées) et gaussiennes $\mathcal{N}(0, \sigma_E^2)$

$\Longrightarrow$ les v.a. Y_i sont indépendantes, gaussiennes de variance σ_E^2 et

$$\mathbb{E}[Y_i] = ax_i + b$$

Notons que les constantes a, b, σ_E^2 sont a priori **inconnues** et doivent être estimées ou évaluées à partir des seules données :

$$(x_i, y_i)_{i=1}^{i=n}$$

Propriétés des estimateurs $\widehat{a}, \widehat{b}$

$$\widehat{a}(Y_1,\dots,Y_n) = \frac{\sum_i (x_i-\overline{x})(Y_i-\overline{Y})}{\sum_i (x_i-\overline{x})^2} = \frac{\sum_i (x_i-\overline{x})Y_i}{\sum_i (x_i-\overline{x})^2} = \frac{\sum_i x_i(Y_i-\overline{Y})}{\sum_i (x_i-\overline{x})^2}$$
$$\widehat{b}(Y_1,\dots,Y_n) = \overline{Y} - \widehat{a}(Y_1,\dots,Y_n)\,\overline{x}$$

Proposition 11

*Sous les hypothèses A1-A3, les estimateurs sont **sans-biais**, i.e.* $\mathbb{E}[\widehat{a}] = a$ *et* $\mathbb{E}[\widehat{b}] = b$. *De plus, leurs variances valent :*

$$\sigma^2(\widehat{a}) = \frac{\sigma_E^2}{\sum_i (x_i-\overline{x})^2} \qquad \sigma^2(\widehat{b}) = \frac{\sigma_E^2 \sum_i x_i^2}{n\sum_i (x_i-\overline{x})^2} \tag{34}$$

La variance de ces estimateurs est d'autant plus faible que la variance σ_E^2 est petite, et/ou que s_X^2 est grande, et/ou que $|\overline{x}|$ est petit.

Estimation du bruit

Les données sur la variable résiduelle, $\widehat{E} := Y - \widehat{Y}$, ne sont pas directement observées mais calculées par

$$\widehat{e}_i = y_i - (\widehat{a}x_i + \widehat{b})$$

Proposition 12

Sous les hypothèses A1-A3, on a

$$\sum_{i=1}^n \widehat{E}_i = \sum_{i=1}^n (Y_i - (\widehat{a}x_i+\widehat{b})) = 0$$

et la v.a.

$$\widehat{\sigma}_E^2 := \frac{1}{n-2}\sum_{i=1}^n (Y_i - (\widehat{a}x_i+\widehat{b}))^2 = \frac{1}{n-2}\sum_{i=1}^n \widehat{E}_i^2$$

est un estimateur sans biais de la variance σ_E^2.

Des estimateurs $\widehat{\sigma}_{\widehat{a}}^2, \widehat{\sigma}_{\widehat{b}}^2$ des variances $\sigma^2(\widehat{a}), \sigma^2(\widehat{a})$ données en (34) sont obtenus en remplaçant σ_E^2 par son estimateur $\widehat{\sigma}_E^2$

Théorème 13 (loi des estimateurs avec σ_E^2 connue)

Si l'hypothèse $A4$ *(en plus de A1-A3) est satisfaite alors on dispose des propriétés suivantes :*

1. $\widehat{a}$ *et* $\widehat{b}$ *sont v.a. gaussiennes de variance données par la formule (34).*
2. *Le vecteur* $(\widehat{a},\widehat{b})$ *est un vecteur gaussien avec* $\mathbb{E}[(\widehat{a},\widehat{b})] = (a,b)$ *et de matrice de covariance*

$$V = \frac{\sigma_E^2}{\sum_i (x_i-\overline{x})^2}\begin{pmatrix} 1 & -\overline{x} \\ -\overline{x} & \sum_i x_i^2/n \end{pmatrix}$$

3. $(n-2)\widehat{\sigma}_E^2/\sigma_E^2$ *suit une loi du* χ_{n-2}^2 *à* $n-2$ *degrés de liberté*
4. *Le vecteur* $(\widehat{a},\widehat{b})$ *et* $\widehat{\sigma}_E^2$ *sont indépendants*

Loi de Student (ou de W. Gosset)

Définition 21

La loi de Student à $n \geq 1$ *degrés de liberté, notée* t_n *est la loi du ratio* $Z_1/\sqrt{Z_2/n}$ *où*

- Z_1 *et* Z_2 *sont indépendantes*
- Z_1 *est une loi gaussienne centrée-réduite et* Z_2 *suit une loi du* χ_n^2 *à* n *degrés de liberté*

Exemple 21

Sous $A1-4$, on peut poser d'après le Théorème 13, $Z_1 := (\widehat{a}-a)/\sqrt{\sigma^2(\widehat{a})} \sim \mathcal{N}(0,1)$ et $Z_2 := (n-2)\widehat{\sigma}_E^2/\sigma_E^2 \sim \chi_{n-2}^2$, alors

$$T = \frac{Z_1}{\sqrt{\frac{Z_2}{n-2}}} = \frac{\widehat{a}-a}{\widehat{\sigma}_{\widehat{a}}} \sim t_{n-2}$$

Test de nullité de a (test de Student)

$$H_0 : a = 0 \text{ contre } H_1 : a \neq 0$$

- Sous H_0, l'Exemple 21 établit que le rapport $T = \dfrac{\widehat{a}}{\widehat{\sigma}_{\widehat{a}}} \sim t_{n-2}$

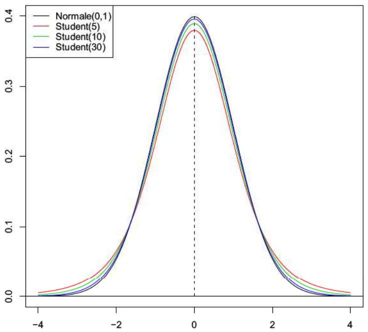
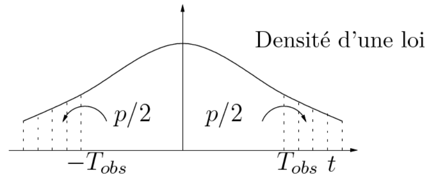


FIGURE 40: Densités de type Student/normale centrée-réduite

- **Rejet de** H_0 si la valeur observée T_{obs} de T correspond à une valeur peu vraisemblable d'une v.a. suivant une loi de Student

On calcule la **p-valeur** du test (cf Définition 18)

$$p = P\left\{|T| \geq T_{obs} = \frac{\widehat{a}}{\widehat{\sigma}_{\widehat{a}}}\right\}$$



- Si p est très faible
  $\Longrightarrow$ la valeur observée T_{obs} de T est peu vraisemblable
  $\Longrightarrow$ rejet de H_0 (le coefficient a est significativement non nul)
- Si p est significativement non nul
  $\Longrightarrow$ la valeur observée T_{obs} de T est vraisemblable
  $\Longrightarrow$ conservation de H_0 (le coefficient a n'est pas significatif pour tout risque $\alpha \leq p$)

Même chose pour le coefficient b

Tester les différentes hypothèses du modèle :

- Nuage des points $\{(x_i, y_i), i = 1, \ldots, n\}$
  ↪ détection de relation non linéaire
- Tester l'homoscédasticité sur un graphiques des résidus (x_i, e_i)
  ↪ repérer des formes suspectes de ce nuage qui devrait se répartir uniformément de part et d'autre de l'axe des abscisses.
- Tester la normalité des résidus en étudiant leur répartition (histogrammes, droite de Henry, test de Kolmogrov-Smirnov, ...) Voir le second semestre
- ...

Remarque : Le théorème 13 est un des résultats de base pour la construction d'intervalles de confiance sur les paramètres de la droite, sur la droite elle-même, ... Voir le second semestre

Prédictif

Prévoir y_{n+1} pour une nouvelle valeur x_{n+1} :

Le modèle	$Y_{n+1} = a x_{n+1} + b + E_{n+1}$
Valeur prédite	$Y^{(p)}_{n+1} = \widehat{a} x_{n+1} + \widehat{b}$

Proposition 13

Sous H1-H3 et σ^2 connue, la variance de la valeur prévue $Y^{(p)}_{n+1}$ est

$$\sigma^2(Y^{(p)}_{n+1}) = \left(\frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}\right) \sigma^2_E$$

On remarque que la prévision est d'autant plus variable que la valeur à prévoir est associée à une valeur de X éloignée du centre de gravité des données (i.e. de la moyenne $\overline{x}$) ayant servies à la construction du modèle

- **Erreur de prévision :** $E_{n+1}^{(p)} = Y_{n+1} - Y_{n+1}^{(p)}$ vérifie

$$\mathbb{E}[E_{n+1}^{(p)}] = 0 \qquad \sigma^2(E_{n+1}^{(p)}) = \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}\right)\sigma_E^2$$

- **Coefficient de PRESS d'Allen :**

$$\mathrm{PRESS} = \sum_{i=1}^{n} (y_i - \widehat{y}_{(i)})^2$$

où $\widehat{y}_{(i)}$ est la valeur estimée par le modèle où l'individu No i n'est pas pris en compte dans les calculs de la régression

(sorte de « validation croisée »)

- **Données**

  p variables continues explicatives $x_1, \dots x_p$

  une variable continue à expliquer Y

  Les p variables sont observées sur $n \geq p$ individus

$$\begin{array}{cc} y & X = \begin{bmatrix} x_1 & \cdots & x_p \end{bmatrix} \\ \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} & \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \end{array}$$

- **Objectifs**
  - **Descriptif :** recherche d'une liaison entre y et p variables potentiellement explicatives $x_1, \dots x_p$: ici, liaison linéaire
  - **Explicatif :** liaison théorique connue **a priori**
    ↪ validation à partir des observations
  - **Prédictif :** À partir d'une liaison établie entre $x_1, \dots, x_p$ et y, **prédire** la valeur de y sur la base de la connaissance de la valeur de $x_1, \dots, x_p$ pour tout nouvel individu

Exemple 22 (Relation Prix /caractéristiques de véhicules)

On dispose de 6 variables explicatives qui sont : la cylindrée, la puissance, la longueur, la largeur, le poids et la vitesse de 18 véhicules :

	Cylindré	Puissance	Vitesse	Poids	Longueur	Largeur
Honda civic	1396	90	174	850	369	166
Peugeot 205 Rallye	1294	103	189	805	370	157
Seat Ibiza SX I	1461	100	181	925	363	161
Citroën AX Sport	1294	95	184	730	350	160
Renault 19	1721	92	180	965	415	169
Fiat Tipo	1580	83	170	970	395	170
Peugeot 405	1769	90	180	1080	440	169
Renault 21	2068	88	180	1135	446	170
Citroën BX	1769	90	182	1060	424	168
Opel Omega	1998	122	190	1255	473	177
Peugeot 405 Break	1905	125	194	1120	439	171
Ford Sierra	1993	115	185	1190	451	172
Renault Espace	1995	120	177	1265	436	177
Nissan Vanette	1952	87	144	1430	436	169
VW Caravelle	2109	112	149	1320	457	184
Audi 90 Quattro	1994	160	214	1220	439	169
BMW 530i	2986	188	226	1510	472	175
Rover 827i	2675	177	222	1365	469	175
Renault 25	2548	182	226	1350	471	180
BMW 325iX	2494	171	208	1600	432	164
Ford Scorpio	2933	150	200	1345	466	176
Fiat Uno	1116	58	145	780	364	155
Peugeot 205	1580	80	159	880	370	156
Ford Fiesta	1117	50	135	810	371	162

Descriptif

- **Modèle considéré**

$$\boxed{y = \sum_{k=1}^{p} a_k x_k + e = X\underline{a} + e}$$

où $\underline{a} = (a_1, \dots, a_p)$ est un vecteur d'inconnues
e variable résiduelle (inconnue)

Remarque : en choisissant $x_{i,1} = 1$ pour tout individu i (i.e. $x_1 = 1$) revient à introduire un terme constant de la modèle

- **Travail**
  - ➡ Évaluer $\underline{a}$: $\widehat{\underline{a}}$ (ajustement du modèle aux données)
  - ➡ Déterminer la QLT de cet ajustement

Critère des moindres carrés appliquées aux résidus :

$$\widehat{a} := \arg\min_{\underline{a}} \sum_i e_i^2 \qquad e_i = y_i - \left(\sum_{k=1}^{p} a_k x_{ik}\right)$$
$$= \arg\min_{\underline{a}} (y - X\underline{a})^\top (y - X\underline{a})$$

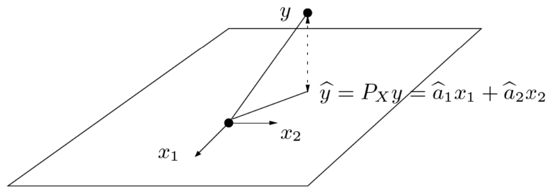


FIGURE 41: y variable à expliquer, deux x_1, x_2 variables explicatives et P_X est le projecteur orthogonal sur le plan engendré par les deux colonnes x_1, x_2 de la matrice X

On résout : $P_X y = X\widehat{a}$

Proposition 14

Si X est de rang p, alors le vecteur $\underline{\widehat{a}}$ est donné par

$$\underline{\widehat{a}} = (X^\top X)^{-1} X^\top y$$

où $P_X = X(X^\top X)^{-1} X^\top$ est la matrice du projecteur orthogonal sur le sous-espace vectoriel engendré par la famille des vecteurs colonnes de la matrice X.

Remarque : La condition de rang plein garantit la seule unicité de la solution.

Définition 22 (Hyperplan de régression et variable résiduelle)

L'hyperplan

$$\widehat{y} = \sum_{k=1}^{p} \widehat{a}_k \, x_k$$

*est appelé **l'hyperplan de régression** et $\underline{\widehat{a}}$ sont les coefficients de la régression. La variable $\widehat{e} := y - X\widehat{a}$ est appelée la **variable résiduelle**.*

Exploration des résidus

Proposition 15 (Quelques propriétés des résidus)

1. *Si le modèle contient un terme constant alors les résidus sont centrés*

$$\frac{1}{n}\sum_i \widehat{e}_i = 0$$

$\Longrightarrow$ *l'hyperplan de régression passe par le centre de gravité $G = (\overline{x}_1, \ldots, \overline{x}_p, \overline{y})$ du nuage de points*

2. *Les résidus sont non-corrélés avec chaque variable explicative*

$$\rho_{\widehat{e}, x_k} = 0, \qquad k = 1, \ldots, p$$

Le théorème de Pythagore donne :

$$\|y\|_2^2 = \|\widehat{y}\|_2^2 + \|\widehat{e}\|_2^2$$

Proposition 16

Si le modèle comporte un terme constant, alors on a toujours avec Pythagore

$$\begin{aligned}
\|y - \overline{y}\mathbf{1}_n\|_2^2 &= \|\widehat{y} - \overline{y}\mathbf{1}_n\|_2^2 + \|\widehat{e}\|_2^2 \\
\sum_i (y_i - \overline{y})^2 &= \sum_i (\widehat{y}_i - \overline{y})^2 + \sum_i \widehat{e}_i^2 \\
\textit{SCTotale} &= \textit{SCExpliquées par le modèle} + \textit{SCRésiduelle}
\end{aligned}$$

Définition 23 (Coefficient de détermination)

Le coefficient de détermination (multiple) R^2 est défini par

$$R^2 := \frac{\|\widehat{y}\|_2^2}{\|y\|_2^2} = 1 - \frac{\|\widehat{e}\|_2^2}{\|y\|_2^2} \tag{35}$$

et si la constance est intégrée au modèle, par

$$R^2 := \frac{\|\widehat{y} - \overline{y}\mathbf{1}_n\|_2^2}{\|y - \overline{y}\mathbf{1}_n\|_2^2} = 1 - \frac{\|\widehat{e}\|_2^2}{\|y - \overline{y}\mathbf{1}_n\|_2^2} = 1 - \frac{\mathrm{SCR}}{\mathrm{SCT}} \tag{36}$$

Remarque : Minimiser $\sum_i e_i^2$ est équivalent à maximiser R^2. Et on peut montrer que

$$R = \rho_{y,\widehat{y}} \quad \text{coefficient de corrélation multiple}$$

Définition 24 (Variance résiduelle et expliquée)

La fraction

$$s_e^2 := \frac{\|\widehat{e}\|_2^2}{\|y - \overline{y}\mathbf{1}_n\|_2^2} = \frac{\mathrm{SCR}}{\mathrm{SCT}}$$

est appelée la variance résiduelle du modèle et

$$s^2_{y:x_1,\dots,x_p} := \frac{\|\widehat{y} - \overline{y}\mathbf{1}_n\|_2^2}{\|y - \overline{y}\mathbf{1}_n\|_2^2} = \frac{\mathrm{SCE}}{\mathrm{SCT}}$$

Dans le cas d'un modèle avec un terme constant, R^2 représente la proportion de variance de Y expliquée par le modèle de régression :

$$R^2 = \frac{\text{Var. expliquée par le modèle}}{\text{Var. totale}} = \frac{s^2_{y:x_1,\dots,x_p}}{s_y^2}$$

- Pour un modèle avec un terme constant

$$s^2_{y:x_1,\dots,x_p} = 0 \iff R^2 = 0 \iff \mathrm{SCR} = \mathrm{SCT}$$
$$R^2 = 1 \iff \mathrm{SCE} = \mathrm{SCT} \iff \mathrm{SCR} = 0 \iff \widehat{y} = X\widehat{a}$$

Il est clair que l'ajout d'une variable explicative supplémentaire (telle que la matrice X soit de rang plein) ne peut pas augmenter la norme du vecteur des résidus.

Proposition 17

La valeur du R^2 est une fonction croissante du nombre de variables explicatives. Donc le R^2 ne peut pas être le seul indicateur d'intérêt pour juger de la qualité de la régression.

Variables	R^2
Puiss.	**0.638**
Puiss. + Poids	**0.687**
Puiss. + Poids + Cylin.	**0.699**
Puiss. + Poids + Cylin. + Larg.	**0.702**
Puiss. + Poids + Cylin. + Larg. + Vite.	**0.709**
Puiss. + Poids + Cylin. + Larg. + Vite. + Long.	**0.709**

Exemple 23 (Modèle obtenu pour l'Exemple 22)

Estimation / Coefficients (avec terme constant)
$n = 18$ individus, $p = 7$ paramètres (constante en queue)

	Coefficient	Écart-type
Cylindrée	**-3.5052**	5.551
Puissance	**282.1684**	174.883
Longueur	**-15.0376**	129.747
Largeur	**208.6932**	412.047
Poids	**12.5748**	24.622
Vitesse de pointe	**-111.1128**	222.256
Constante	**-8239.4531**	42718.363

Ajustement global	
Somme des carrés des écarts	**SCR = 213563456**
Coeff. de corrélation multiple	$\boldsymbol{R = 0.8421 \quad R^2 = 0.7091}$

- **Remarque importante :** un modèle de régression linéaire ne veut pas nécessairement dire que le lien entre la variable à expliquer et les variables explicatives est linéaire mais que le **modèle est linéaire en les paramètres** :
- Ainsi on peut considérer tout type de transformation connue et fixée à l'avance des variables explicatives :
  - mise en place d'inter-actions : $y = x_1 + 3x_2 + 6x_1x_2$ prend la forme habituelle en posant $x_3 = x_1x_2$;
  - régression polynomiale $y = 10x_1 + 8x_2 - 6x_1x_2 + 2x_1^2 + 4x_1^2$: on obtient la forme habituelle du modèle en posant $x_3 = x_1x_2$, $x_4 = x_1^2$, $x_5 = x_2^2$,
  - ...
- Par contre, toute transformation faisait intervenir des paramètres inconnus ne rentrent pas dans ce cadre.

Explicatif

Modélisation probabiliste : $i = 1, \dots, n$, $Y_i = \sum_{k=1}^{p} a_k x_{ik} + E_i$
où E_i est un bruit aléatoire et $x_{i1}, \dots, x_{ip}$ sont déterministes

$$Y = X\underline{a} + E \quad \text{(écriture vectorielle)}$$

Hypothèses fondamentales :

A1) $x_1, \dots, x_p$ sont (des vecteurs) déterministes (variables contrôlées)
E_i est une v.a. centrée et de variance constante σ_E^2 (homoscédasticité)

A2) $a_1, \dots, a_p$ sont des constantes (pas de rupture du modèle)

A3) les v.a. E_i sont gaussiennes $\mathcal{N}(0, \sigma_E^2)$ et indépendantes (non-corrélées) donc E est un vecteur gaussien de matrice des covariances $\sigma_E^2 I_n$

↪ **inférence sur le modèle**

À partir de $A1 - A3$

➜ les v.a. Y_i sont indépendantes, gaussiennes de variance constante σ_E^2 (et Y est un vecteur gaussien de matrice des covariances $\sigma_E^2 I_n$) et

$$i = 1, \dots, n \quad \mathbb{E}[Y_i] = \sum_{k=1}^{p} a_k\, x_{ik} \quad (\Longleftrightarrow \mathbb{E}[Y] = X\underline{a})$$

➜ $\widehat{\underline{a}} = (X^\top X)^{-1} X^\top Y$ est un vecteur aléatoire gaussien (comme transformation linéaire d'un vecteur gaussien) de caractéristiques données par la proposition suivante.

Proposition 18 (L'estimateur de $\underline{a}$ avec σ_E^2 connue)

L'estimateur des moindres carrés de $\underline{a}$

$$\widehat{\underline{a}} := (X^\top X)^{-1} X^\top Y$$

est sans biais et de matrice de covariance

$$V_{\widehat{\underline{a}}} = (X^\top X)^{-1} \sigma_E^2. \tag{37}$$

Variable résiduelle $\widehat{E}$

$$\widehat{E} = Y - \widehat{Y} = Y - X\widehat{\underline{a}} = Y - P_X Y = (I - P_X)Y = P_{X^\perp} Y = P_{X^\perp} E.$$

L'orthogonal $X^\perp$ de l'espace engendré par les colonnes de X est souvent appelé l'espace des résidus.

Proposition 19 (Propriétés des v.a. $\widehat{E}$ et $\widehat{Y}$)

Sous les hypothèses A1-A3, on a

1. $\mathbb{E}[\widehat{E}] = P_{X^\perp}\mathbb{E}[E] = 0$ *et la matrice de covariance de* $\widehat{E}$ *est* $V_{\widehat{E}} = \sigma_E^2\, P_{X^\perp} I P_{X^\perp}{}^\top = \sigma_E^2\, P_{X^\perp}$.
2. *Comme* $\widehat{Y} = X\widehat{\underline{a}}$, $\mathbb{E}[\widehat{Y}] = X\underline{a}$ *et* $V_{\widehat{Y}} = \sigma_E^2\, P_X$ *avec la Prop 18.*
3. $C(\widehat{E}, \widehat{Y}) = 0$.

Les résidus estimés $\widehat{E}$ de E possèdent la même espérance que E. Les composantes de $\widehat{E}$ sont généralement corrélées comme fonction de X. Donc si on veut des propriétés analogues à celle du bruit $\mathbb{E}$, il faudrait que

- les éléments non-diagonaux de P_X soient suffisamment petits ;
- les éléments diagonaux de P_X soient approximativement égaux.

Estimation du bruit

Un estimateur naturel de la variance du bruit est

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{E}_i^2$$

Mais avec Prop 19.1

$$\begin{aligned}
\mathbb{E}[\sum_{i=1}^{n} \widehat{E}_i^2] &= \sum_{i=1}^{n} \mathbb{E}[\widehat{E}_i^2] = \sum_{i=1}^{n} \sigma^2(\widehat{E}_i) = \mathrm{Trace}(V_{\widehat{E}}) \\
&= \sigma_E^2 \mathrm{Trace}(P_{X^\perp}) = \sigma_E^2 \dim(X^\perp) = \sigma_E^2 (n - p)
\end{aligned} \tag{38}$$

Proposition 20

La v.a.

$$\widehat{\sigma}_E^2 := \frac{1}{n-p}\sum_{i=1}^{n} \widehat{E}_i^2 = \frac{\|\widehat{E}\|_2^2}{n-p} = \frac{\mathrm{SCR}}{n-p}$$

est un estimateur sans biais de la variance σ_E^2.

Corollaire 1 (Estimateurs des variances des paramètres du modèles)

1. *La v.a.*
$$\widehat{V}_{\underline{\widehat{a}}} := (X^\top X)^{-1}\widehat{\sigma}^2_E = (X^\top X)^{-1}\frac{\mathrm{SCR}}{n-p}$$
est un estimateur sans biais de la matrice de covariance $V_{\underline{\widehat{a}}}$ (cf (37)).
2. *Un estimateur de la variance de $\widehat{a}_k$ du coefficient a_k de la régression :*

$$\widehat{\sigma}^2_{\widehat{a}_k} = \widehat{\sigma}^2_E \, (X^\top X)^{-1}{}_{ii} = \frac{\mathrm{SCR}}{n-p}(X^\top X)^{-1}{}_{ii} \tag{39}$$

Proposition 21 (Lois des estimateurs avec σ^2_E connue)

*Sous A1-**A3**, on a d'après le théorème de Cochran sur les vecteurs gaussiens*

1. *$\underline{\widehat{a}}$ est un vecteur gaussien de moyenne $\underline{a}$ et de variance (37)*
2. *$(n-p)\widehat{\sigma}^2_E/\sigma^2_E$ suit une loi du χ^2_{n-p}*
3. *$\underline{\widehat{a}}$ et $\widehat{\sigma}^2_E$ sont indépendantes.*

Test de nullité de a_k (test de Student)

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

- **Sous** H_0 et en utilisant la Prop 21, le rapport
$$T = \frac{\widehat{a}_k}{\sqrt{\widehat{\sigma}^2_{\widehat{a}_k}}} \sim \text{Student à } n-p \text{ degrés de liberté}$$
où $\widehat{\sigma}^2_{\widehat{a}_k}$ est l'estimateur (39) de la variance de a_k
- **Rejet de** H_0 si la valeur observée de T correspond à une valeur peu vraisemblable d'une v.a. suivant une loi de Student
- Calculer la p-valeur du test et si
$$\mathrm{p} = P\left\{|St_{n-p}| \geq T_{obs} = \frac{\widehat{a}_k}{\sqrt{\widehat{\sigma}^2_{\widehat{a}_k}}}\right\}$$
est très faible alors le coefficient a_k est significatif pour toute valeur du risque $\alpha > \mathrm{p}$.
Dans le cas contraire, a_k n'est pas significatif pour toute valeur du risque $\alpha \leq \mathrm{p}$.

Test de nullité simultanée d'un sous-ensemble des paramètres

$H_0 : r$ paramètres a_k sont nuls contre H_1 : modèle complet

- **Sous** H_0 **:** le modèle → vecteur $\widehat{Y}_0$ et $\widehat{E}_0 := Y - \widehat{Y}_0$
sous H_1 **:** le modèle complet → $\widehat{Y}$ et $\widehat{E} := Y - \widehat{Y}$.
- **La statistique de décision :**

$$\begin{aligned} F &= \frac{(\|\widehat{E}_0\|_2^2 - \|\widehat{E}\|_2^2)/\text{Espérance}}{\|\widehat{E}\|_2^2/\text{Espérance}} \\ &= \frac{\|\widehat{E}_0\|_2^2 - \|\widehat{E}\|_2^2}{\|\widehat{E}\|_2^2}\frac{n-p}{r} = \frac{\mathrm{SCR}_{H_0} - \mathrm{SCR}_{H_1}}{\mathrm{SCR}_{H_1}} \times \frac{n-p}{r} \\ &\sim \mathrm{Fisher}(r, n-p) \ \textbf{sous}\ H_0 \end{aligned} \tag{40}$$

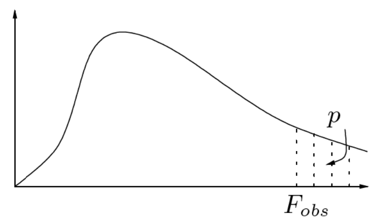


FIGURE 42: Densité d'une loi de Fisher

- **Si** F **est grand** (diff. $\mathrm{SCR}_{H_0} - \mathrm{SCR}_{H_1} >> 0$) alors l'effet des r variables est important et on **rejette** H_0
- **Si la p-valeur** $\mathrm{p} := P\{\mathrm{Fisher}(r, n-p) \geq F_{obs}\}$ est très faible alors la valeur de F est peu vraisemblable et l'ajout des r variables est significatif.
↪ Comparer des modèles emboîtes

Exemple 24 (Prix d'un véhicule en fonction de 6 caractéristiques)

Estimation / Coefficients (avec terme constant)

$n = 18$ individus, $p = 7$ paramètres (constante en queue)

	Coefficient	Ecart-type	**Test coef nul Student(11)**	**p-valeur**
Cylindrée	-3.5052	5.551	**0.631**	**0.541**
Puissance	282.1684	174.883	**1.613**	**0.135**
Longueur	-15.0376	129.747	**0.116**	**0.910**
Largeur	208.6932	412.047	**0.506**	**0.623**
Poids	12.5748	24.622	**0.511**	**0.620**
Vitesse de pointe	-111.1128	222.256	**0.500**	**0.627**
Constante	-8239.4531	42718.363	**0.193**	**0.851**

Test de nullité simultanée des coefficients des 6 variables de base

Fisher = 4.469 Deg.Lib = (6 , 11)
p-valeur = 0.0156 V.Test = 2.16

Exemple 25 (Comparaison de modèles)

Modèle à 2 caractéristiques (Poids, Puis.) et une constante

Somme des carrés des écarts : $\mathrm{SCR} = 230064624$

Coefficient de corrélation multiple : $R = 0.8286$ $R^2 = 0.6866$

Variance estimée des résidus : $\widehat{\sigma}_E^2 = 15337641$ $\widehat{\sigma}_E = 3916.3301$

Test de nullité simultanée des coefficients des 2 caractéristiques

Fisher = 16.433 Deg.Lib = (2 , 15)
p-valeur = 0.0002 V.Test = 3.60

Modèle à 6 caractéristiques et une constante

Somme des carrés des écarts : $\mathrm{SCR} = 213563456$

Coefficient de corrélation multiple : $R = 0.8421$ $R^2 = 0.7091$

Variance estimée des résidus : $\widehat{\sigma}_E^2 = 19414860$ $\widehat{\sigma}_E = 4406.2295$

Test de nullité simultanée des coefficients des 6 caractéristiques

Fisher = 4.469 Deg.Lib = (6 , 11)
p-valeur = 0.0156 V.Test = 2.16

	Prix	Prix estimé avec les variables poids et puissance	Résidu	Prix estimé avec six variables	Résidu
Alfasud-TI-1350	30570	29752,50	817,48	29616,10	953,90
AUDI-100-L	39990	34738,60	5251,40	36259,70	3730,35
Simca-1307-GLS	29600	30811,10	-1211,09	31411,10	-1811,15
Citroën GS-club	28250	27280,20	969,75	26445,70	1804,25
Fiat 132-1600GLS	34900	36904,90	-2004,92	37043,00	-2143,00
Lancia-beta-1300	35480	33726,20	1753,83	34972,80	507,16
Peugeot 504	32300	34523,40	-2223,36	33749,10	-1449,14
Renault 16-TL	32000	27904,50	4095,53	26579,90	5420,05
Renault 30-TS	47700	45630,90	2069,06	44445,60	3254,42
Toyota Corolla	26540	24696,50	1843,51	24650,20	1889,78
Alfetta	42395	38067,30	4327,74	38270,50	4124,54
Princess-1800-HL	33990	35042,30	-1052,26	34830,40	-840,42
Datsun-200L	43980	44204,90	-224,92	44872,40	-892,43
Taunus-2000-GL	35010	36493,60	-1483,64	36343,50	-1333,48
Rancho	39450	34186,30	5263,66	35638,10	3811,94
Mazda-9295	27900	34145,90	-6245,90	32233,40	-4333,43
Opel Rekord-L	32700	37497,60	-4797,63	37103,50	-4403,50
Lada 1300	22100	29248,20	-7148,23	30389,80	-8289,80

TABLE 29: Prix estimé pour chaque marques par les deux modèles

Prédictif

Prévoir y_{n+1} pour un nouveau vecteur $x_{n+1}^\top = (x_{n+1,1}, \ldots, x_{n+1,p})$:

$$\text{Le modèle} \quad Y_{n+1} = x_{n+1}^\top \underline{a} + E_{n+1}$$

$$\text{La prévision} \quad Y_{n+1}^{(p)} = x_{n+1}^\top \underline{\widehat{a}}$$

- **Erreur de prévision :** $E_{n+1}^{(p)} = Y_{n+1} - Y_{n+1}^{(p)}$ vérifie

$$\begin{aligned} \mathbb{E}[E_{n+1}^{(p)}] &= 0 \\ V_{E_{n+1}^{(p)}} &= V_{x_{n+1}^\top(\underline{\widehat{a}}-\underline{a})+E_{n+1}} = x_{n+1}^\top V_{\underline{\widehat{a}}-\underline{a}}\, x_{n+1} + \sigma_E^2 \\ &= x_{n+1}^\top (X^\top X)^{-1} x_{n+1} \sigma_E^2 \quad \text{(voir (37))} \end{aligned}$$

- **Coefficient de PRESS d'Allen**

$$\mathrm{PRESS} = \sum_{i=1}^{n} (y_i - \widehat{y}_{(i)})^2$$

où $\widehat{y}_{(i)}$ est la valeur estimée par le modèle où l'individu No i n'est pas pris en compte dans les calculs de la régression

↪ **comparer les capacités prédictives de deux modèles**

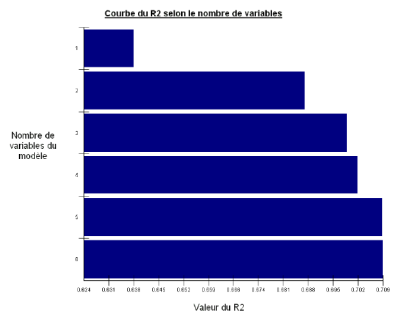
Objectif

Un bon choix de variables dépend de l'objectif poursuivi :

- **descriptif :** tendance à adjoindre trop de variables pour obtenir une QLT d'ajustement importante
- **prédictif :** le modèle qui explique le mieux les données selon un critère descriptif (au prix d'un nombre important de variables), **n'est pas celui qui conduit aux prédictions les plus fiables**
  ↪ modèles parcimonieux

1. **Critère du $R^2 = \mathrm{SCR}/\mathrm{SCT}$**
   - R^2 mesure la qualité du modèle aux données
   - **Pb :** indicateur croissant avec le nombre de variables explicatives
   - Utile pour **comparer deux modèles avec même nbre de variables**

Exemple 26 (Exemple 22)



Le meilleur ajustement suivant le R^2 pour chaque nombre de variables

1 Variable : $R^2 = 0.638$

Coefficient	Student	p-valeur	Libellé de la variable
257.5898	5.31	**0.000**	**Puissance**

2 Variables : $R^2 = 0.687$

Coefficient	Student	p-valeur	Libellé de la variable
172.9672	2.39	**0.031**	**Puissance**
16.4512	1.53	**0.148**	**Poids**

3 Variables : $R^2 = 0.699$

Coefficient	Student	p-valeur	Libellé de la variable
-3.6126	0.75	**0.465**	**Cylindrée**
203.8248	2.42	**0.030**	**Puissance**
20.7195	1.68	**0.115**	**Poids**

4 Variables : $R^2 = 0.702$

Coefficient	Student	p-valeur	Libellé de la variable
-3.9924	0.79	**0.445**	**Cylindrée**
208.8129	2.37	**0.034**	**Puissance**
99.7384	0.36	**0.723**	**Largeur**
18.1956	1.25	**0.232**	**Poids**

5 Variables : $R^2 = 0.709$

Coefficient	Student	p-valeur	Libellé de la variable
-3.4666	0.65	**0.526**	**Cylindrée**
283.8549	1.70	**0.115**	**Puissance**
180.9165	0.56	**0.584**	**Largeur**
11.0632	0.55	**0.590**	**Poids**
-113.4316	0.53	**0.602**	**Vitesse de pointe**

2 Critère du R^2 ajusté

- R^2 est calculé sur un échantillon de taille n
  On peut définir, à partir du modèle de régression multiple, un R^2 pour la population entière

$$R^2_{pop} = 1 - \frac{\sigma^2(E)}{\sigma^2(Y)}$$

- Si le modèle intègre une constante, on reporte les estimateurs de $\sigma^2(E)$ et $\sigma^2(Y)$

$$\widehat{\sigma}^2_E = \frac{\text{SCE}}{n-p} \qquad \widehat{\sigma}^2_Y = \frac{\text{SCT}}{n-1}$$

$$\boxed{R^2_{\text{ajusté}} := 1 - \frac{\text{SCE}}{\text{SCT}} \times \frac{n-1}{n-p}} \qquad \left(R^2 = 1 - \frac{\text{SCE}}{\text{SCT}}\right)$$

- On vérifie que $R^2_{\text{ajusté}} = \frac{(n-1)R^2-(p-1)}{n-p}$

$$\Longrightarrow R^2_{\text{ajusté}} < R^2 \quad \text{pour } p \geq 2$$

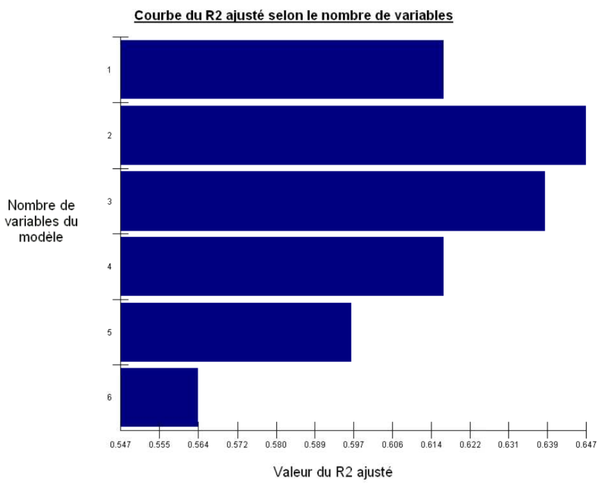
- $R^2_{\text{ajusté}}$ peut prendre des valeurs négatives, n'est plus croissant avec p

↪ **comparer deux modèles n'ayant pas le même nombre de variables**

- $\sigma^2(Y)$ ne dépend pas du modèle

$$\Longrightarrow \left| \begin{array}{l} \max R^2_{\text{ajusté}} \\ \Longleftrightarrow \min \widehat{\sigma}^2_E \\ \Longleftrightarrow \text{minimiser l'erreur quadratique moyenne} \end{array} \right.$$

Exemple 27 (Exemple 22)



Exemple 28 (Les meilleurs ajustements selon le $R^2_{\text{ajusté}}$)

1 Variable : $R^2_{\text{ajusté}}$ = 0.617
Puissance

2 Variables : $R^2_{\text{ajusté}}$0.647

Coefficient	Student	**p-valeur**	**Libellé de la variable**
172.9672	2.39	**0.031**	**Puissance**
16.4512	1.53	**0.148**	**Poids**

3 Variables : $R^2_{\text{ajusté}}$ = 0.639
Cylindrée, Puissance, Poids

4 Variables : $R^2_{\text{ajusté}}$= 0.617
Cylindrée, Puissance, Largeur, Poids

5 Variables : $R^2_{\text{ajusté}}$= 0.597
Cylindrée, Puissance, Largeur, Poids, Vitesse de pointe

6 Variables : $R^2_{\text{ajusté}}$= 0.564
Longueur, Cylindrée, Puissance, Largeur, Poids,Vitesse de pointe

3 Statistique de Fisher (cadre explicatif)

- Critère basé sur le test de nullité d'une famille de r coefficients.
- **La statistique (partielle) de Fisher** (avec terme constant)

$$\begin{aligned}
F &= \frac{\mathrm{SCE}_{H_0} - \mathrm{SCE}}{\mathrm{SCE}} \times \frac{r}{n-p} \\
&= \frac{\mathrm{SCE}_{H_0}/\mathrm{SCT} - \mathrm{SCE}/\mathrm{SCT}}{\mathrm{SCE}/\mathrm{SCT}} \times \frac{r}{n-p} \\
&= \frac{1 - R^2_{H_0} - (1-R^2)}{1-R^2} \times \frac{r}{n-p} \\
&= \frac{R^2 - R^2_{H_0}}{1-R^2} \times \frac{r}{n-p} \\
&\sim \mathrm{Fisher}(r, n-p) \ \textbf{sous} \ H_0 \text{ voir (40)}
\end{aligned}$$

- **Sous l'hypothèse d'un modèle réduit**, plus la p-valeur associée au test est grande plus le modèle réduit est significatif.

↬ **Comparer des modèles emboîtés**

4 Critère du C_p de Mallows

- Si le modèle est correct, alors

$$\mathbb{E}[\widehat{Y}_i] = \mathbb{E}[Y_i]$$

Dans le cas contraire, on définit alors **le biais de** $\widehat{Y}_i$ par

$$\mathrm{Biais}(\widehat{Y}_i) = \mathbb{E}[\widehat{Y}_i] - \mathbb{E}[Y_i]$$

- **L'erreur quadratique moyenne** de l'estimateur $\widehat{Y}_i$ est définie par

$$\begin{aligned}
\mathrm{MSE}(\widehat{Y}_i) &= \mathrm{Var}(\widehat{Y}_i) + \mathrm{Biais}(\widehat{Y}_i)^2 \\
& \widehat{y}_i \to \mathbb{E}[\widehat{y}_i] \to \mathbb{E}[Y_i]
\end{aligned}$$

- **Idée de base du critère :** pour j variables explicatives

$$\min \sum_i \mathrm{MSE}(\widehat{Y}_i) = \min \left\{ \mathbb{E}[\mathrm{SCE_j}] - \sigma_E^2(n-2j) \right\}$$

- **Définition du coefficient**

$$\Gamma_j := \frac{\sum_i \mathrm{MSE}(\widehat{Y}_i)}{\sigma_E^2} = \frac{\mathbb{E}[\mathrm{SCE}_j]}{\sigma_E^2} - n + 2j$$

Si le modèle est correct alors $\mathbb{E}[\mathrm{SCE}_j] = (n-j)\sigma_E^2$ (voir (38))

$$\Longrightarrow \Gamma_j = j$$

- On estime Γ_j par

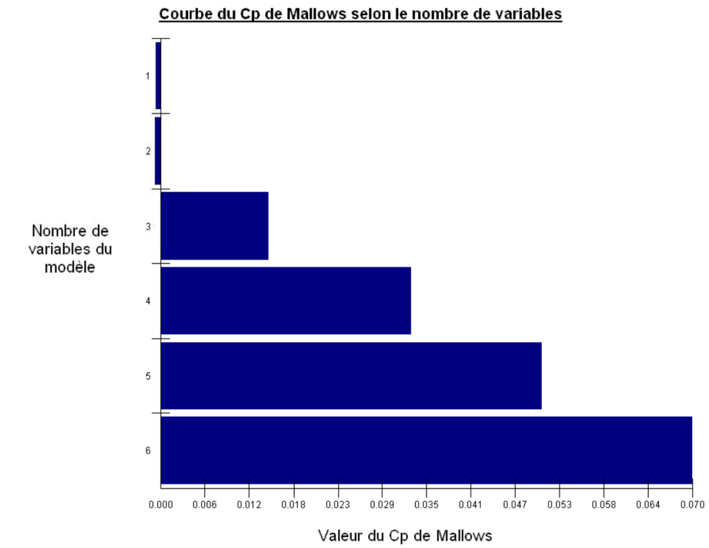
$$\boxed{C_j = \frac{\mathrm{SCE}_j}{\widehat{\sigma}_E^2} - n + 2j}$$

où, en général, l'estimateur $\widehat{\sigma}_E^2$ est obtenu pour le modèle avec toutes les variables explicatives du modèle : $\widehat{\sigma}_E^2 = \mathrm{SCE}_p/(n-p)$.

- **Rechercher un modèle tel que C_j soit proche de** j (en général $C_j \leq j$).

Revient à préférer au « vrai » modèle complet, un modèle réduit donc biaisé mais d'estimation plus précise

Exemple 29 (Exemple 22)



Les meilleurs ajustements selon le coefficient de Mallows C_p

1 Variable : C_p = -0.064
Puissance

2 Variables : C_p= -0.073

Coefficient	Student	**p-valeur**	**Libellé de la variable**
172.9672	2.39	**0.031**	**Puissance**
16.4512	1.53	**0.148**	**Poids**

3 Variables : C_p = 1.425
Cylindrée, Puissance, Poids

4 Variables : C_p= 3.301
Cylindrée, Puissance, Largeur, Poids

5 Variables : C_p= 5.015
Cylindrée, Puissance, Largeur, Poids, Vitesse de pointe

6 Variables : C_p= 7.000
Longueur, Cylindrée, Puissance, Largeur, Poids,
Vitesse de pointe

TABLE 30: Le logiciel sélectionne le modèle à deux variables, i.e. $\min_j C_p(j)$!

Recherche exhaustive : délicat pour p grand : 2^p modèles possibles

1. **Pas à pas**
   - **forward :**
     - À chaque pas, une variable est ajoutée au modèle et elle est sélectionnée de manière à présenter la plus petite p-valeur.
     - Arrêt lorsque toutes les variables ont été ajoutées ou la p-valeur $p_c >$ seuil fixé (0.5)
   - **backward :**
     - On démarre du modèle complet.
     - À chaque étape, la variable associée à la plus grande valeur de la p-valeur p_c du test de Student est éliminée du modèle.
     - Arrêt lorsque les variables restant dans le modèle ont des valeurs de $p_c <$ seuil fixé (0.1)
   - **stepwise :** introduction d'une étape d'élimination de variable après chaque étape de sélection
     But : retirer du modèle d'éventuelles variables devenues moins pertinentes du fait de la présence de celle nouvelle variable

2. **Par échange**
   - **Maximisation de R^2 :** déterminer le meilleur modèle de chaque niveau (nb de var. fixées).
     - À chaque niveau, sélection d'une variable complémentaire qui rend l'accroissement du R^2 maximum.
     - Puis on regarde tous les échanges possibles entre une variable présente dans le modèle et une extérieure, et on exécute l'échange qui fournit l'accroissement max.
     - Ceci est itéré tant que le R^2 croît.

   - **Minimisation du R^2 :**
     - idem algo précédent
       sauf que la procédure fait appel au couple de variables associées au plus petit accroissement du R^2.

     L'objectif est d'explorer plus de modèles que dans le cas précédent et donc d'obtenir éventuellement un meilleur optimum.

Pour ces algos de type pas à pas ou d'échange, il est important de compléter les comparaisons des différentes solutions à l'aide de critères globaux

3. **Global**
   - **Algorihtme de Furnival et Wilson (74)**
     Comparer tous les modèles possibles en cherchant à optimiser l'un des critères : $R^2, R^2_{\text{ajusté}}$ ou C_p de Mallows

     En général, temps d'exécution acceptable avec un nombre de variables inférieur à 40.

Colinéarité dans les variables explicatives

- On approche Y (au sens des moindres carrés) par une combinaison linéaire des variables explicatives
- Sous l'hypothèse que la matrice X est constituée d'une famille de colonnes linéairement indépendantes, **l'ajustement requiert l'inversion de la matrice $X^\top X$**

  Si certains des vecteurs $x_1, \ldots, x_p$ ont des directions voisines alors l'inversion est difficile
  et comme $\det(X^\top X) \approx 0$ **les variances (39) des estimateurs des coefficients deviennent grandes**.
  $\Longrightarrow$ la solution $\widehat{a}$ est instable

  ↪ Pb dit « de colinéarité » ou « multilinéarité »

- **Une solution possible : régression sur composantes principales issues d'une ACP du tableau X**
  - Apprécier la colinéarité
    Détecter les redondances
  - Compos. principales $\equiv$ Variables orthogonales (non-corrélées) 2 à 2
    ↪ Nouvelles variables explicatives
  - On détermine, en général, le nombre de valeurs propres à conserver en fonction d'un critère adapté à l'objectif suivi : descriptif, explicatif ou prédictif.
    ou choisir **les composantes principales les plus fortement corrélées avec la variable à expliquer Y**
  - Les composantes associées aux valeurs propres proches de 0 fournissent une image des relations linéaires existants entre les variables explicatives initiales.
- D'autres solutions : régression ridge, **régression aux moindres carrés partiels (PLS)**,. . .

Résultats de la régression sur composantes principales pour l'Exemple 22

k	RMSE	Const.	Cyl.	Puis.	Long.	Larg.	Poids	Vit.
1	4301.68	-43286.46	2.7437	49.978	46.028	175.804	7.589	71.383
2	4401.15	-34893.04	2.9482	62.544	34.555	124.103	6.498	102.827
3	4451.25	-5360.02	4.310	75.618	30.148	-39.880	11.593	45.222
4	4296.24	-5829.58	-2.621	131.959	70.751	-167.635	18.661	64.667
5	4294.23	-9856.87	-4.015	181.544	-42.917	141.908	26.310	11.216
6	4406.23	-8239.36	-3.505	282.169	-15.038	208.694	12.575	-111.114

TABLE 31: Les résultats sont ordonnés suivant le nombre k de CP retenues

```
HISTOGRAMME DES  6 PREMIERES VALEURS PROPRES
+----+--------+-------+--------+----------------------------------------------------------------------------------+
| No | VALEUR | POURC.| POURC. |                                                                                  |
|    | PROPRE |       | CUMULE |                                                                                  |
+----+--------+-------+--------+----------------------------------------------------------------------------------+
| 1  | 4.4209 | 73.68 |  73.68 | ******************************************************************************** |
| 2  | 0.8561 | 14.27 |  87.95 | ****************                                                                 |
| 3  | 0.3731 |  6.22 |  94.17 | *******                                                                          |
| 4  | 0.2139 |  3.57 |  97.73 | ****                                                                             |
| 5  | 0.0928 |  1.55 |  99.28 | **                                                                               |
| 6  | 0.0433 |  0.72 | 100.00 | *                                                                                |
+----+--------+-------+--------+----------------------------------------------------------------------------------+
     +--------------------------------+
     | CORRELATIONS VARIABLE-FACTEUR |
     +--------------------------------+
     |    1      2      3      4     5 |
-----+--------------------------------+
Prix |  0.77 -0.09  0.13  0.23  0.16 |
-----+--------------------------------+
```


Exemple 30 (Prévision de la qualité d'un programme (en termes de nombres de fautes potentielles) à partir de métriques classiques de complexité)

- Observation de 11 métriques sur n programmes
  Prévision pour le $(n+1)$ avec une régression multiple
- Problème de très forte colinéarité

	A	B	C	D	E	F	G	H	I	J	K
A) Total lines	1	**0.98**	**0.96**	0.79	0.74	**0.90**	0.89	0.86	0.86	0.83	0.42
B) Code lines		1	**0.92**	0.72	0.62	**0.93**	**0.92**	0.88	0.87	0.85	0.43
C) Total chars			1	0.79	0.80	**0.93**	**0.90**	0.87	0.87	0.80	0.41
D) Comments				1	0.78	0.65	0.64	0.65	0.65	0.62	0.32
E) Comment chars					1	0.54	0.52	0.53	0.52	0.52	0.30
F) Code chars						1	**0.96**	**0.92**	**0.92**	0.81	0.40
G) Halstead(1)							1	**0.92**	**0.92**	0.84	0.40
H) Halstead(2)								1	1	0.80	0.39
I) Jensen									1	0.79	0.38
J) McCabe										1	0.55
K) Bandwith											1

ACP préliminaire

Métrique	C_1	C_2
A	**0.965**	0.192
B	**0.964**	0.213
C	**0.941**	0.265
D	**0.932**	0.285
E	**0.917**	0.269
F	**0.916**	0.273
G	**0.909**	0.306
H	**0.899**	0.107
I	**0.833**	-0.010
J	**0.799**	0.481
K	0.145	**0.964**
Valeurs propres	8.291	1.65

Pourcentage d'inertie $(C_1, C_2) \quad = (8.291 + 1.65)/11 \approx 0.90$
↪ Régression sur les deux variables explicatives : C_1, C_2

SAS

Procédure : **REG** avec méthode de sélection de variables via l'option **SELECTION** (**STEPWISE**,**FORWARD**, **CP**, . . .)