

Le modèle de régression linéaire
Master 2 Recherche SES-IES Analyse de données

Ana Karina Fermin

Université Paris Nanterre

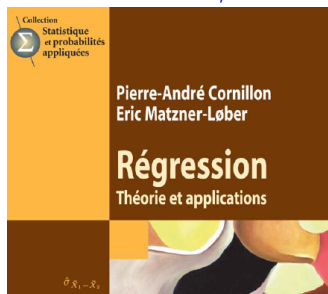
<http://fermin.perso.math.cnrs.fr/>

- ## 2 Multiple Linear Regression

D'un point de vue pratique l'objectif est double.

- Ajuster un modèle pour expliquer Y en fonction de X
- Prédire les valeurs de Y pour de nouvelles valeurs de X .

Bibliographie : Pierre-André Cornillon, Eric Matzner-Lober



On souhaite “expliquer” une variable Y à partir de X .
Nous allons chercher une fonction f tel que

$$y_i \approx f(\mathbf{x}_i).$$

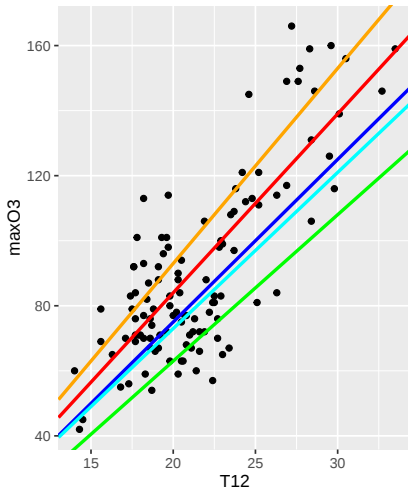
Pour définir $\approx$ il faut donner un critère quantifiant la qualité de l'ajustement de la fonction f aux données. On a besoin également d'une classe de fonctions $\mathcal{S}$ dans laquelle on choisira f .

$$\hat{f} = \arg \min_{f \in \mathcal{S}} \sum_{i=1}^n \ell(f(\mathbf{x}_i) - y_i)$$

où $\ell(\cdot)$ est appelée fonction de coût ou encore fonction de perte.

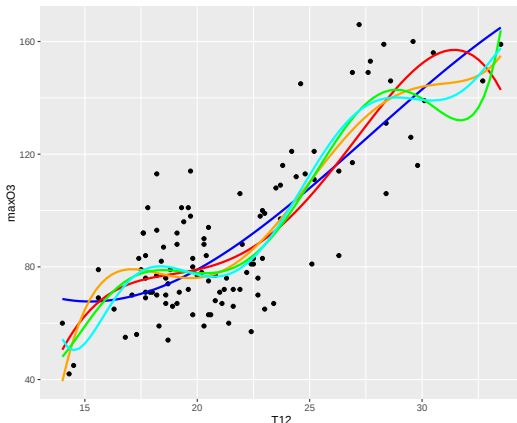
Nous considérons ici la fonction de perte quadratique ($\ell(\cdot) = (\cdot)^2$).

$\mathcal{S}$: Famille des fonctions linéaires



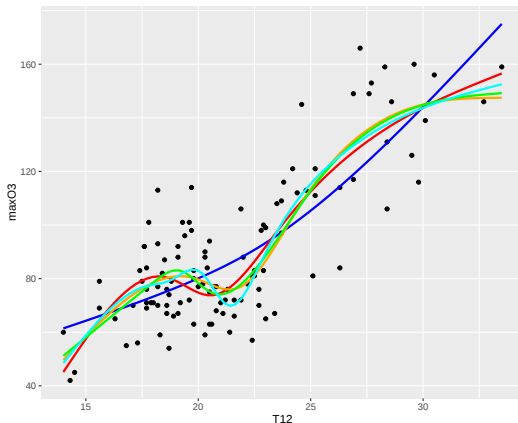
Objectif : Parmi toutes les droites possibles, déterminer la droite qui minimise la somme des écarts aux carrés.

f est choisie dans une classe des fonctions $\mathcal{S}$ polynomiales
Modèles obtenus par des polynôme du degré 3, 4, 5, 6 et 7
Pb : Choisir "le bon" degré !



Objectif : Parmi toutes les fonctions possibles, déterminer la fonction qui minimise la somme des écarts aux carrés.

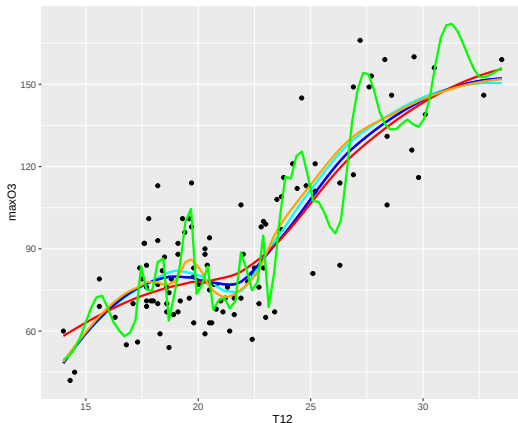
Modèles obtenus par splines



Objectif : Parmi toutes les fonctions possibles, déterminer la meilleur fonction qui minimise la somme des écarts aux carrés.

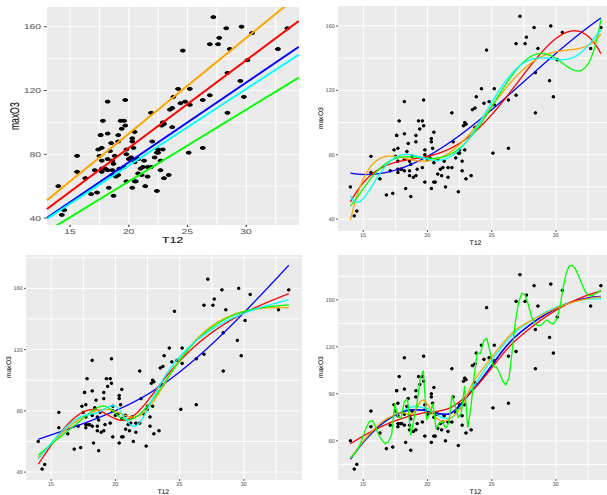
f est choisie dans une classe des fonctionnes $\mathcal{S}$ plus complexe

Modèles obtenus par estimateurs à noyau



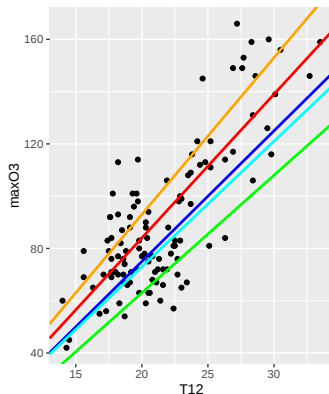
Objectif : Parmi toutes les fonctions possibles, déterminer la meilleur fonction qui minimise la somme des écarts aux carrés.

Quel modèle choisir? Linéaire, Polynomiale, Spline, Noyau?



$\mathcal{S}$: Famille des fonctions linéaires

$$\mathcal{S} = \{f : f_{\beta}(\mathbf{T12}) = \beta_1 + \beta_2 \mathbf{T12} \quad \beta_1 \in \mathbb{R}, \beta_2 \in \mathbb{R}\}$$



Objectif : Parmi toutes les droites possibles, déterminer la droite qui minimise la somme des écarts aux carrés.

Méthode des moindres carrés

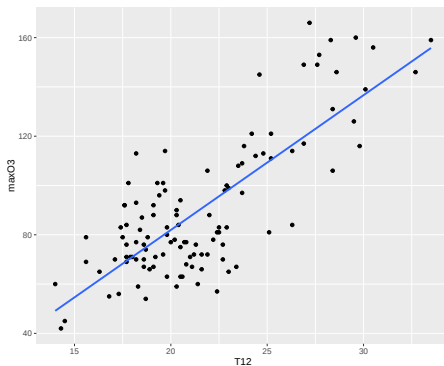
$$\begin{aligned}\sum_{i=1}^n |Y_i - f_{\beta}(\mathbf{X}_i)|^2 &= \sum_{i=1}^n |\max 03_i - f_{\beta}(\mathbf{T12}_i)|^2 \\ &= \sum_{i=1}^n |\max 03_i - (\beta_1 + \beta_2 \mathbf{T12}_i)|^2\end{aligned}$$

- Choisir β qui minimise la quantité

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^2} \sum_{i=1}^n |\max 03_i - (\beta_1 + \beta_2 \mathbf{T12}_i)|^2$$

- Minimisation facile avec solution explicite!

Prédiction



- Prédiction linéaire pour ozone :

$$\widehat{\text{maxO3}} = f_{\hat{\beta}}(\text{T12}) = \hat{\beta}_1 + \hat{\beta}_2 \text{T12}$$

Démarche à suivre :

- ➊ Voir et représenter les données.
- ➋ Choisir le type de modèle.
- ➌ Ajuster le modèle.
- ➍ Valider le modèle.
- ➎ Selon les besoins, faire de l'inférence (tests, régions de confiance...), de la prédiction etc.

Modèle de régression

On dispose de n observations $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)$ du couple $(\mathbf{X}, Y)$. On suppose que

$$y_i = f(\mathbf{x}_i) + \varepsilon_i \text{ pour tout } i = 1, \dots, n$$

- les $\mathbf{x}_i$ son des valeurs connues non aléatoires
- f est une fonction inconnue
- ε_i sont des réalisations inconnues d'une variable aléatoire.

Pour chaque individu i , la variable aléatoire ε_i représente l'erreur commise. Généralement pour étudier le modèle "le statisticien" formule des hypothèses sur la loi des erreurs ε_i .

Modèle gaussien de la régression linéaire simple

On observe des observations bruités

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n$$

avec β_0 et β_1 inconnus.

- Le premier terme correspond à l'équation d'une droite.
- Le deuxième terme correspond à l'erreur et varie de façon aléatoire d'un individu à l'autre.

Hypothèse sur les erreurs

On suppose que les ε_i sont les réalisations i.i.d. d'une variable aléatoire gaussienne centrée et de variance σ^2 inconnue. Cette hypothèse va nous permettre de calculer des régions de confiance et de proposer des tests.

Équation de régression

- $Y \in \mathbb{R}$ et $\mathbf{X} \in \mathbb{R}^p$ (ici $p = d + 1$ correspond au nombre de paramètres à estimer)
- $\beta \in \mathbb{R}^p$
- Exemple :
 - Régression linéaire simple :
 - $Y = \text{maxO3}$ et $\mathbf{X} = \begin{pmatrix} 1 \\ \mathbf{T12} \end{pmatrix}$ (variable *dummy* pour l'intercept)
 - $f_\beta(\mathbf{X}) = \langle \mathbf{X}, \beta \rangle = \beta_0 \times 1 + \beta_1 \times \mathbf{T12} = \beta_0 + \beta_1 \mathbf{T12}$
 - Régression linéaire multiple:
 - $Y = \text{maxO3}$ et $\mathbf{X} = \begin{pmatrix} 1 \\ \mathbf{T12} \\ \mathbf{Vx} \\ \mathbf{Ne12} \end{pmatrix}$
 - $f_\beta(\mathbf{X}) = \langle \mathbf{X}, \beta \rangle = \beta_0 + \beta_1 \mathbf{T12} + \beta_2 \mathbf{Vx} + \beta_3 \mathbf{Ne12}$

Modèle gaussien de la régression linéaire multiple

On observe des observations bruités

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_d x_{id} + \varepsilon_i, \quad i = 1, \dots, n$$

avec $\beta_0, \beta_1, \dots, \beta_d$ inconnus.

On suppose que les ε_i sont les réalisations i.i.d. d'une variable aléatoire gaussienne centrée et de variance σ^2 inconnue.

Supposons qu'on dispose de d -variables explicatives $X_1, X_2, \dots, X_d$.
Soit $\mathbb{X}$ la matrice augmentée (n lignes et $d + 1$ colonnes).
Soit $\beta = (\beta_0, \beta_1, \dots, \beta_d)$ le vecteur de coefficients inconnus.

Modèle Théorique

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_d X_d + \varepsilon$$

Modèle Théorique (sous forme matricielle)

$$Y = \mathbb{X}\beta + \varepsilon$$

Considérons le modèle théorique de régression linéaire multiple.

- ① Coefficients estimés (para le méthode de MC) :

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_d)$$

$$\hat{\beta} = (\mathbb{X}^t \mathbb{X})^{-1} \mathbb{X}^t Y$$

- ② Valeur prédite pour l'i-ème individu

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_d x_{id}$$

- ③ Somme des carrés des résidus

$$\text{SCR} = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

- ④ Estimateur sans biais de σ^2 est

$$\hat{\sigma}^2 = \frac{\text{SCR}}{n - p}.$$

Ici $p = d + 1$

Effet d'une variable explicative

- La variable X_j est-elle utile ?
- On a besoin de d'un test d'hypothèse pour répondre à cette question

Le Modèle

- Le modèle est raisonnable ?
- On a besoin de d'un test d'hypothèse pour répondre à cette question

Test de Student

- La variable X_j est-elle utile ?

Test sur le paramètre β_j

Nous souhaitons tester une hypothèse nulle de la forme

$$H_0 : \beta_j = 0$$

L'hypothèse alternative est

$$H_1 : \beta_j \neq 0$$

Sous H_0 , $T = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$ suit la loi de Student à $n - p$ degrés de liberté ($n - 2$ degrés de liberté dans le cas simple).

Modèle gaussien de régression linéaire simple

$O3_i = \beta_0 + \beta_1 T12_i + \varepsilon_i$, où les ε_i sont i.i.d. gaussiennes centrées.

Résultat obtenue avec logiciel R (une partie):

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -27.4196 9.0335 -3.035 0.003 **

T12 5.4687 0.4125 13.258 <2e-16 ***

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

Modèle gaussien de régression linéaire simple

$O3_i = \beta_0 + \beta_1 T12_i + \varepsilon_i$, où les ε_i sont i.i.d. gaussiennes centrées.

On obtient avec le logiciel R :

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -27.4196 9.0335 -3.035 0.003 **

T12 5.4687 0.4125 13.258 <2e-16 ***

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

Residual standard error: 17.57 on 110 degrees of freedom

Multiple R-squared: 0.6151, Adjusted R-squared: 0.6116

F-statistic: 175.8 on 1 and 110 DF, p-value: < 2.2e-16

Rappelons qu'on dispose d'un échantillon de taille $n = 112$

Test de Global du modèle (Test de Fischer)

- Supposons que le modèle est $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_d X_d + \varepsilon$,
- Soit $p = d + 1$
- $SCR = \sum(\hat{y}_i - y_i)^2$ et $SCE = \sum(\hat{y}_i - \bar{y})^2$
- Le modèle est raisonnable ?

Test Global du modèle

Nous souhaitons tester une hypothèse nulle de la forme

$$H_0 : \beta_j = 0 \text{ pour tout } j \in \{1, \dots, p\},$$

L'hypothèse alternative H_1 est qu'il existe au moins un $j \in \{1, \dots, p\}$ pour lequel $\beta_j \neq 0$.

Sous H_0 , $F = \frac{SCE/(p-1)}{SCR/(n-p)}$ suit la loi de Fisher à $p - 1$ et $n - p$ degrés de liberté.

$$\text{MLG1 } \max\text{O3}_i = \beta_0 + \beta_1 \text{T12}_i + \beta_2 \text{Vx12}_i + \varepsilon_i$$

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept)	-14.4242	9.3943	-1.535	0.12758
T12	5.0202	0.4140	12.125	< 2e-16 ***
Vx12	2.0742	0.5987	3.465	0.00076 ***

Residual standard error: 16.75 on 109 degrees of freedom

Multiple R-squared: 0.6533, Adjusted R-squared: 0.6469

F-statistic: 102.7 on 2 and 109 DF, p-value: < 2.2e-16

$$\text{MLG2 } \max\text{O3}_i = \beta_0 + \beta_1 \text{T12}_i + \beta_2 \text{Ne12}_i + \varepsilon_i$$

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept)	7.7077	15.0884	0.511	0.61050
T12	4.4649	0.5321	8.392	1.92e-13 ***
Ne12	-2.6940	0.9426	-2.858	0.00511 **

Residual standard error: 17.02 on 109 degrees of freedom

Multiple R-squared: 0.6419, Adjusted R-squared: 0.6353

F-statistic: 97.69 on 2 and 109 DF, p-value: < 2.2e-16

Comparer **MLG1** et **MLG2** : Test de Fisher, R^2 , R^2 -ajusté, ...

Attention :

- $SCR = \sum(\hat{y}_i - y_i)^2$ et $SCE = \sum(\hat{y}_i - \bar{y})^2$
- $SCT = SCE + SCR$
- $R^2 = \frac{SCE}{SCT}$
- R^2 ne s'interprète que dans les modèles comportant un intercept.
- R^2 augmente si on ajoute des variables explicatives

$$\text{MLG1 } \max O3_i = \beta_0 + \beta_1 T12_i + \beta_2 Vx12_i + \varepsilon_i$$

Estimate Std. Error t value Pr(>|t|)

(Intercept)	-14.4242	9.3943	-1.535	0.12758
T12	5.0202	0.4140	12.125	< 2e-16 ***
Vx12	2.0742	0.5987	3.465	0.00076 ***

Residual standard error: 16.75 on 109 degrees of freedom

Multiple R-squared: 0.6533, Adjusted R-squared: 0.6469

F-statistic: 102.7 on 2 and 109 DF, p-value: < 2.2e-16

$$\text{MLG3 } O3_i = \beta_0 + \beta_1 T12_i + \beta_2 Vx12_i + \beta_3 Ne12_i + \varepsilon_i$$

lm(formula = maxO3 ~ T12 + Vx12 + Ne12)

Estimate Std. Error t value Pr(>|t|)

(Intercept)	3.8958	14.8243	0.263	0.7932
T12	4.5132	0.5203	8.674	4.71e-14 ***
Vx12	1.6290	0.6571	2.479	0.0147 *
Ne12	-1.6189	1.0181	-1.590	0.1147

Residual standard error: 16.63 on 108 degrees of freedom

Multiple R-squared: 0.6612, Adjusted R-squared: 0.6518

F-statistic: 70.25 on 3 and 108 DF, p-value: < 2.2e-16

Test de Fischer

On test la nullité d'un certain nombre q de paramètres dans un modèle de p paramètres.

H_0 : modèle réduit avec $p - q$ paramètres

H_1 : modèle avec p paramètres.

Modèles Emboîtés

$$\text{MLG1 } O3_i = \beta_0 + \beta_1 T12_i + \beta_2 Vx12_i + \varepsilon_i$$

$$\text{MLG3 } O3_i = \beta_0 + \beta_1 T12_i + \beta_2 Vx12_i + \beta_3 Ne12_i + \varepsilon_i$$

Model 1: $O3 \sim T12 + Vx12$

Model 2: $O3 \sim T12 + Vx12 + Ne12$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	109	30580				
2	108	29881	1	699.61	2.5286	0.1147

Remarque : Le test F entre ces deux modèles est équivalent au test T de nullité du coefficient de la variable Ne12 dans le modèle MLG3 (les deux p-values valent 0.1147).

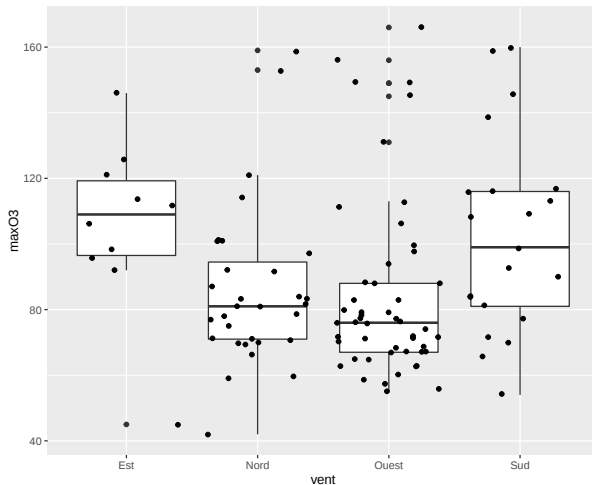
Régression sur des variables qualitatives

- X variable qualitative à k modalités $A_1, A_2, \dots, A_k$.
- Comment **coder** une variable qualitative à k modalités pour l'utiliser dans un seul modèle de régression linéaire ?
- **Codage disjonctif** : codage par $k - 1$ variables muettes ou indicatrices

$$\mathbf{X} = (\mathbf{1}_{A_1}, \dots, \mathbf{1}_{A_k})$$

Rappel : Une variable muette ou indicatrice (en anglais on parle de *variable dummy*) est une variable qualitative qui prend les valeurs 0 ou 1.

Motivation



Nous remplaçons la variable vent pour son codage disjonctif.

$$y_{ij} = \beta_0 + \beta_j + \varepsilon_{ij} \quad i = 1, \dots, n_j \quad j = A_1, \dots, A_k$$

Variable vent : A_1 : Est, A_2 : Nord, A_3 : Ouest et A_4 : Sud

Est	Nord	Ouest	Sud
10	31	50	21

$$\text{maxO3} = \beta_0 + \beta_1 \mathbf{1}_{\text{Nord}} + \beta_2 \mathbf{1}_{\text{Ouest}} + \beta_3 \mathbf{1}_{\text{ventSud}} + \varepsilon$$

- Modèle avec intercept

(Intercept)	ventNord	ventOuest	ventSud
105.60	-19.47	-20.90	-3.08

- Modèle sans intercept

ventEst	ventNord	ventOuest	ventSud
105.60	86.13	84.70	102.52

Que peut-on remarquer ?

$$\text{maxO3} = \beta_0 + \beta_1 \text{ventNord} + \beta_2 \text{ventOuest} + \beta_3 \text{ventSud} + \varepsilon$$

On obtient les résumés suivants :

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept)	105.600	8.639	12.223	<2e-16 ***
ventNord	-19.471	9.935	-1.960	0.0526 .
ventOuest	-20.900	9.464	-2.208	0.0293 *
ventSud	-3.076	10.496	-0.293	0.7700

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

Residual standard error: 27.32 on 108 degrees of freedom

Multiple R-squared: 0.08602, Adjusted R-squared: 0.06063

F-statistic: 3.388 on 3 and 108 DF, p-value: 0.02074

Rappelons qu'on dispose d'un échantillon de taille $n = 112$

- 1 Régression linéaire simple
- 2 Multiple Linear Regression
- 3 Sélection de variables**
- 4 Validation de modèle

Sélection de variables : motivations

- 1 Description : Quelles sont les variables les plus influentes?
- 2 On cherche un "bon" modèle : ni trop simple ni trop complexe
- 3 On recherche donc un modèle parcimonieux.

Recherche exhaustive

Démarche :

- On se donne un critère de qualité, qu'on calcule pour tous les sous-modèles comportant un intercept, et on retient le modèle qui optimise le critère.
- Critères :
 - R^2 , R^2 -ajusté,
 - Par pénalisation
 - C_p
 - AIC, BIC
 - Ridge, Lasso, Elastic-net
 -
 - Validation croisée
- Dans la suite verra rapidement les critères : C_p , AIC, BIC
- Validation croisée (voir le chapitre suivant)

Estimation par pénalisation

Ceux-ci se présentent comme l'ajout d'une pénalisation de la fonction objectif du problème d'optimisation : moindres carrés ou risque empirique, vraisemblance.

Estimation par pénalisation (Cp de Mallow)

- ❶ Méthode : moindres aux carrés (Least Squares en anglais)
- ❷ Risque empirique : $R_n(f) = \frac{1}{n} \sum_{i=1}^n (Y_i - f(\mathbf{X}_i))^2$
- ❸ Cp de Mallow: premier critère visant à une meilleure estimation de l'erreur de prévision que la seule considération de l'erreur d'ajustement (ou le R^2) dans le modèle linéaire.
- ❹ La méthode de Mallows consiste à chercher la fonction que minimise le risque empirique pénalisé

$$R_n(f) + \text{pen}(\mathcal{S})$$

$$\text{avec } \text{pen}(\mathcal{S}) = 2\hat{\sigma}^2 \frac{D_{\mathcal{S}}}{n} = 2\hat{\sigma}^2 \frac{p}{n}$$

Exemples de Penalités

Moindres carrés $R_n(f) = \frac{1}{n} \sum_{i=1}^n (Y_i - f(\mathbf{X}_i))^2$

- Cp/Unbiased estimation: $\text{pen}(\mathcal{S}) = 2\hat{\sigma}^2 \frac{D_{\mathcal{S}}}{n}$
- BIC (Bayesian): $\text{pen}(\mathcal{S}) = \log n \hat{\sigma}^2 \frac{D_{\mathcal{S}}}{n}$
- Generally: $\text{pen}(\mathcal{S}) \sim \lambda_n \hat{\sigma}^2 \frac{D_{\mathcal{S}}}{n}$

Maximum de Vraisemblance (Likelihood)

$R_n(f) = \frac{1}{n} \sum_{i=1}^n -\log \mathbb{P}(Y_i | X_i)$

- AIC : $\text{pen}(\mathcal{S}) = \frac{D_{\mathcal{S}}}{n}$
- BIC (Bayesian): $\text{pen}(\mathcal{S}) = \frac{\log n}{2} \frac{D_{\mathcal{S}}}{n}$
- Generally: $\text{pen}(\mathcal{S}) \sim \lambda_n \frac{D_{\mathcal{S}}}{n}$

Modèle Gaussien

- modèle Gaussien : ϵ i.i.d. $\mathcal{N}(0, \sigma^2)$ (hypothèse très forte)!
- Vraisemblance (Likelihood):

$$\mathbb{P}_{\sigma, \beta}(Y_i | X_i) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\|\mathbf{Y} - \mathbb{X}\beta\|^2 / (2\sigma^2)}$$

- opposé du log-vraisemblance:

$$-\log \mathbb{P}_{\sigma, \beta}(Y_i | X_i) = \frac{n}{2}(\log(2\pi) + \log \sigma^2) + \frac{1}{2\sigma^2} \|\mathbf{Y} - \mathbb{X}\beta\|^2$$

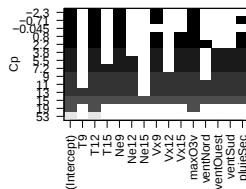
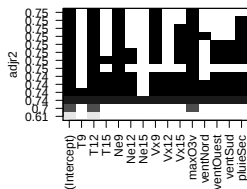
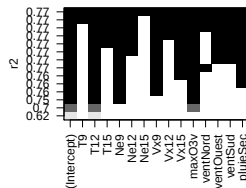
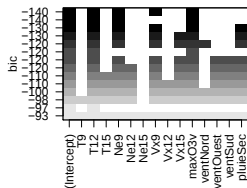
- Rem : Estimateur de β par Maximum de vraisemblance coïncide avec l'estimateur de β ! par moindres carrés. En revanche petite différence avec l'estimateur de σ^2

Recherches pas à pas

On les utilise lorsque le modèle complet est trop riche pour permettre facilement une recherche exhaustive. A chaque étape, on ajoute (ou on élimine) la variable la plus (ou la moins) significative, la variable considérée pouvant être fictive.

- ❶ Descendante (Backward) part du modèle complet et élimine les variables une à une.
- ❷ Ascendante (Forward) part du modèle constant et ajoute les variables une à une.
- ❸ Mixte (Stepwise) combine les deux précédentes.

Recherche exhaustive



Recherche pas à pas (mixte)

Stepwise (mixte):

On peut aussi donner les directions "backward" ou "forward"

Start: AIC=612.99

max03 ~ T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + Vx9 + Vx12 + Vx15 +
max03v + vent + pluie

Step: AIC=608.61

max03 ~ T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + Vx9 + Vx12 + Vx15 +
max03v + pluie

.....

Step: AIC=596.02

max03 ~ T12 + Ne9 + Vx9 + max03v

- 1 Régression linéaire simple
- 2 Multiple Linear Regression
- 3 Sélection de variables
- 4 Validation de modèle

Validation de modèle

- Qualité de l'ajustement du modèle retenu
- Graphes de résidus (simples, standardisés ou studentisés)
- QQ-plot
- Tests d'ajustement (e.g. Shapiro-Wilks, Kolmogorov-Smirnov)

Exemple Ozone : modèle retenu

$$\text{maxO3}_i = \beta_0 + \beta_1 \text{T12}_i + \beta_2 \text{Vx9}_i + \beta_3 \text{Ne9}_i + \beta_4 \text{maxO3v}_i + \varepsilon_i$$

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	12.63131	11.00088	1.148	0.253443	
T12	2.76409	0.47450	5.825	6.07e-08	***
Vx9	1.29286	0.60218	2.147	0.034055	*
Ne9	-2.51540	0.67585	-3.722	0.000317	***
maxO3v	0.35483	0.05789	6.130	1.50e-08	***

Residual standard error: 14 on 107 degrees of freedom
 Multiple R-squared: 0.7622, Adjusted R-squared: 0.7533
 F-statistic: 85.75 on 4 and 107 DF, p-value: < 2.2e-16

Test de normalité pour les résidus

Shapiro-Wilk normality test
 W = 0.9659, p-value = 0.005817

Analyse de résidus pour le modèle retenu

