Le modèle de régression linéaire

Master 2 Pyschologie

Ana Karina Fermin

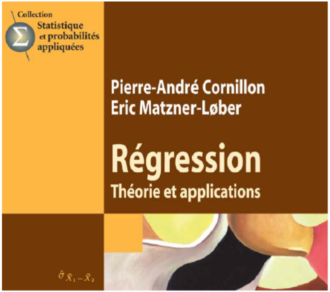
Université Paris Nanterre

aferminrodriguez@parisnanterre.fr

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

- Ajuster un "bon" modèle statistique qui décrit l’impact de plusieurs variables sur la variabilité d’une variable réponse quantitative
- Faire la prédiction

Bibliographie : Pierre-André Cornillon, Eric Matzner-Lober



Données ozone

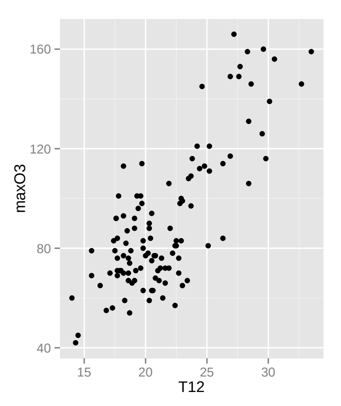
Nous commençons toujours par voir et représenter les données !

```
112 obs. of  13 variables:
maxO3 : int  87 82 92 114 94 80 79 79 101 106 ...
T9    : num  15.6 17 15.3 16.2 17.4 17.7 16.8 14.9 16.1 18.3 ...
T12   : num  18.5 18.4 17.6 19.7 20.5 19.8 15.6 17.5 19.6 21.9 ...
T15   : num  18.4 17.7 19.5 22.5 20.4 18.3 14.9 18.9 21.4 22.9 ...
Ne9   : int  4 5 2 1 8 6 7 5 2 5 ...
Ne12  : int  4 5 5 1 8 6 8 5 4 6 ...
Ne15  : int  8 7 4 0 7 7 8 4 4 8 ...
Vx9   : num  0.695 -4.33 2.954 0.985 -0.5 ...
Vx12  : num  -1.71 -4 1.879 0.347 -2.954 ...
Vx15  : num  -0.695 -3 0.521 -0.174 -4.33 ...
maxO3v: int  84 87 82 92 114 94 80 99 79 101 ...
vent  : Factor w/ 4 levels "Est","Nord","Ouest",..: 2 2 1 2 3 3 3 2 2 3 ...
pluie : Factor w/ 2 levels "Pluie","Sec": 2 2 2 2 2 1 2 2 2 2 ...
```

Exemple : Pollution l'ozone

- X : température à midi
- Y : concentration maximale en ozone

mesurés en un lieu donné et une journée donnée pendant n jours.



Objectif

On souhaite "prédire" une variable Y à partir de $\mathbf{X}$.
Nous allons chercher une fonction f tel que

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

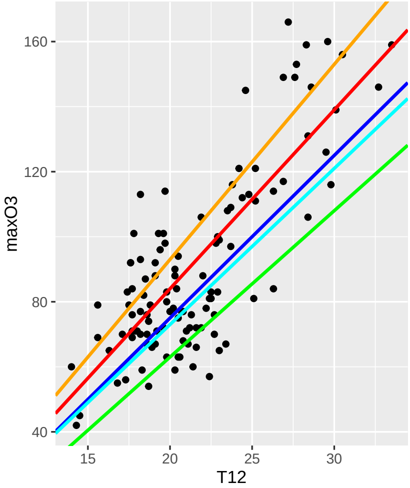
Pour définir $\approx$ il faut donner un critère quantifiant la qualité de l'ajustement de la fonction f aux données. On a besoin également d'une classe de fonctions $\mathcal{S}$ dans laquelle on choisira f.

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

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

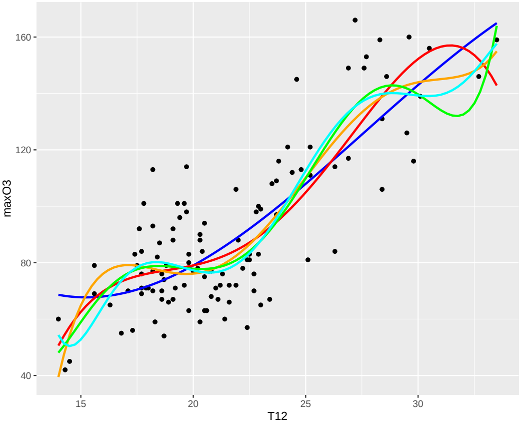
Nous considérons ici la fonction de perte quadratique ($\ell(\cdot) = (\cdot)^2$).

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



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

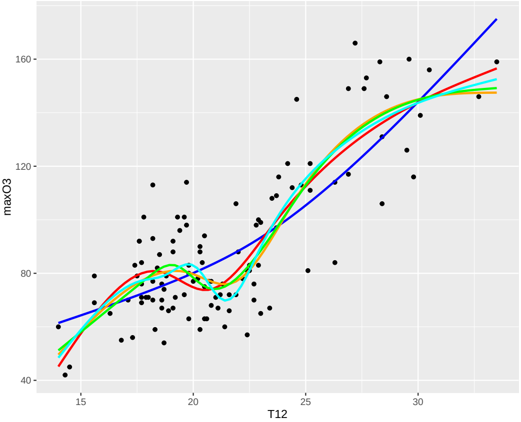
f est choisie dans une classe des fonctions $\mathcal{S}$ polynomiales
Modèles obtenus par des polynôme du degré 3, 4, 5, 6 et 7
Pb : Choisir "le bon" degré !



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

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

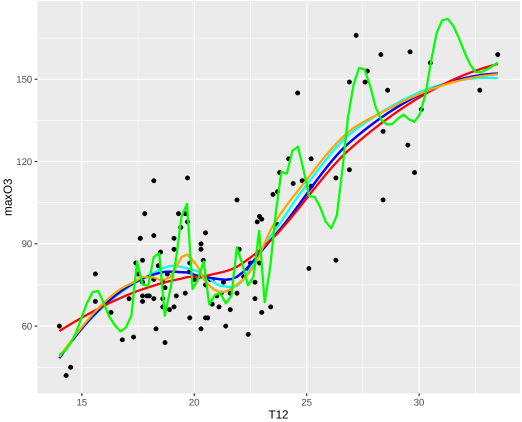
Modèles obtenus par splines



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

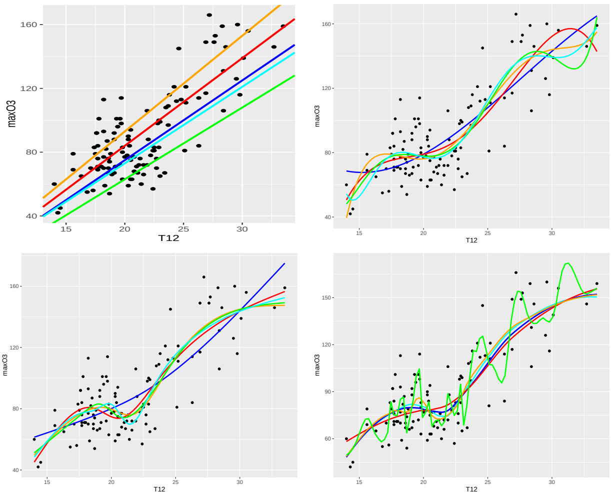
f est choisie dans une classe des fonctionnes $\mathcal{S}$ plus complexe

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?



Démarche à suivre :

1. Voir et représenter les données (si possible).
2. Choisir le type de modèle.
3. Ajuster le modèle.
4. Selon les besoins, valider le modèle, faire de l'inférence (tests, régions de confiance...), de la prédiction etc.

Dans ce cours on se concentre plus dans le problème de prédiction et moins dans l'inférence.

Questions

- Comment écrire le modèle de régression multiple?
- Comment estimer les paramètres inconnus de ce modèle?

Modèle gaussien de la régression linéaire multiple

On observe des observations bruités

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

où les ε_i sont les réalisations i.i.d. d'une variable aléatoire normale centrée et de variance σ^2 inconnue et les coefficients $\beta_0, \beta_1, \ldots, \beta_d$ sont inconnus.

Remarques :

- Les ε_i rendent compte de la variabilité individuelle, des erreurs de mesure, ...(variabilité non expliquée par X)
- Y est une variable aléatoire, X est supposée sans erreur

Hypothèses du modèle

Plusieurs hypothèses sont faites:

1. la relation x-Y est linéaire
2. les erreurs sont distribuées selon une loi normale
3. les erreurs ont même variance,
4. les erreurs sont indépendantes,
5. pas d'outliers.

Le logiciel R fournit 4 graphiques de diagnostic

- $X_1, X_2, \ldots, X_d$ d-variables explicatives
- $\mathbb{X}$ la matrice augmentée (n lignes et $d+1$ colonnes)
- $\beta = (\beta_0, \beta_1, \ldots, \beta_d)$

Modèle Théorique

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

Modèle Théorique (sous forme matricielle)

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

Inférence

A partir de l'échantillons $\{(y_i, x_i)\}_{i=1}^n$, on veut estimer

- les paramètres de moyenne : $\beta_0, \beta_1, \ldots, \beta_d$
- les paramètres de variance : σ^2

Deux méthodes:

- Méthode des moindres carrés ordinaires (MCO)
- Méthode du maximum de vraisemblance (ML)

Remarques :

- MCO est une méthode géométrique qui permet estimer les paramètres de la moyenne. MCO ne prend pas en compte la loi des erreurs et donc ne fournit pas d'estimateur de la variance
- Equivalence entre MCO et ML pour l'estimation de la moyenne (voir Annexe)

Méthode de moindres carrés ordinaires (MCO)

- Estimateur de moindres carrés

$$\widehat{f} = \arg\min_{f \in \mathcal{S}} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2$$

- Supposons $f(x) = \beta_0 + \beta_1 x_1 \dots \beta_d x_d$

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_{i1} \dots \beta_d x_{id} - y_i)^2$$

Remarque: pour la méthode ML voir l'annexe.

Considérons le modèle théorique de régression linéaire multiple.

1. Coefficients estimés (par MCO ou ML) : $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_d)$

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

2. Valeur prédite pour l'i-ème individu

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

3. Somme des carrés des résidus

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

4. Estimateur de σ^2 (par ML)

$$\widehat{\sigma}^2 = \frac{\mathrm{SCR}}{n-(d+1)}.$$

Qualité d'ajustement et prédiction

- $\mathrm{SCR} = \sum(\widehat{y}_i - y_i)^2$ et $\mathrm{SCE} = \sum(\widehat{y}_i - \bar{y})^2$
- $\mathrm{SCT} = \mathrm{SCE} + \mathrm{SCR}$

Coefficient R^2

Un critère intuitif pour mesurer l'ajustement du modèle aux données est

$$R^2 = \frac{\mathrm{SCE}}{\mathrm{SCT}} = 1 - \frac{\mathrm{SCR}}{\mathrm{SCT}}$$

- On regarde si une large part de la variabilité de Y est expliquée par le modèle
- R^2 ne s'interprète que dans les modèles comportant un intercept.
- R^2 augmente si on ajoute des variables explicatives

Effet d'une variable explicative

- La variable X_j est-elle utile ?

Le Modèle

- Le modèle est raisonnable ?

MLG1 $\text{maxO3}_i = \beta_0 + \beta_1 \text{T12}_i + \beta_2 \text{Vx12}_i + \varepsilon_i$

```
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
```

MLG2 $\text{maxO3}_i = \beta_0 + \beta_1 \text{T12}_i + \beta_2 \text{Ne12}_i + \varepsilon_i$

```
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, ...

ML du modèle Gaussien

- Densité associé à la loi normale
  $$\frac{1}{(2\pi\sigma^2)^{1/2}} e^{-(y-(\beta_0+\beta_1 x_1+\ldots+\beta_d x_d))^2/(2\sigma^2)}$$
- Vraisemblance d'un échantillon $\{(x_1, y_1), \ldots, (x_n, y_n)\}$:
  $$V = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-(y_i-(\beta_0+\beta_1 x_{i1}+\ldots+\beta_d x_{id}))^2/(2\sigma^2)}$$
- Opposé du log-vraisemblance:
  $$-\log V = \frac{n}{2}(\log(2\pi)+\log(\sigma^2)) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y-(\beta_0+\beta_1 x_1+\ldots+\beta_d x_d))^2/(2\sigma^2)$$
- Rappel : Estimateur de β par ML coïncide avec l'estimateur de β par moindres carrés.

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, $T = \frac{\widehat{\beta}_j}{\widehat{\sigma}_{\widehat{\beta}_j}}$ suit la loi de Student à $n-(d+1)$ degrés de liberté ($n-2$ degrés de liberté dans le cas simple).

Test de Global du modèle (Test de Fischer)

- Supposons que le modèle est $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_d X_d + \varepsilon$,
- $\mathrm{SCR} = \sum(\widehat{y}_i - y_i)^2$ et $\mathrm{SCE} = \sum(\widehat{y}_i - \bar{y})^2$
- Le modèle est raisonnable ?

Test Global du modèle

Nous souhaitons tester une hypothèse nulle de la forme

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

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

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

Validation de modèle : Analyse du résidus

- Qualité de l'ajustement du modèle retenu
- Graphes de résidus (simples, standardisés ou studentisés)
- QQ-plot

Analyse de résidus pour le modèle retenu