Le modèle de régression linéaire

L3 Gestion Statistiques (App)

Ana Karina Fermin

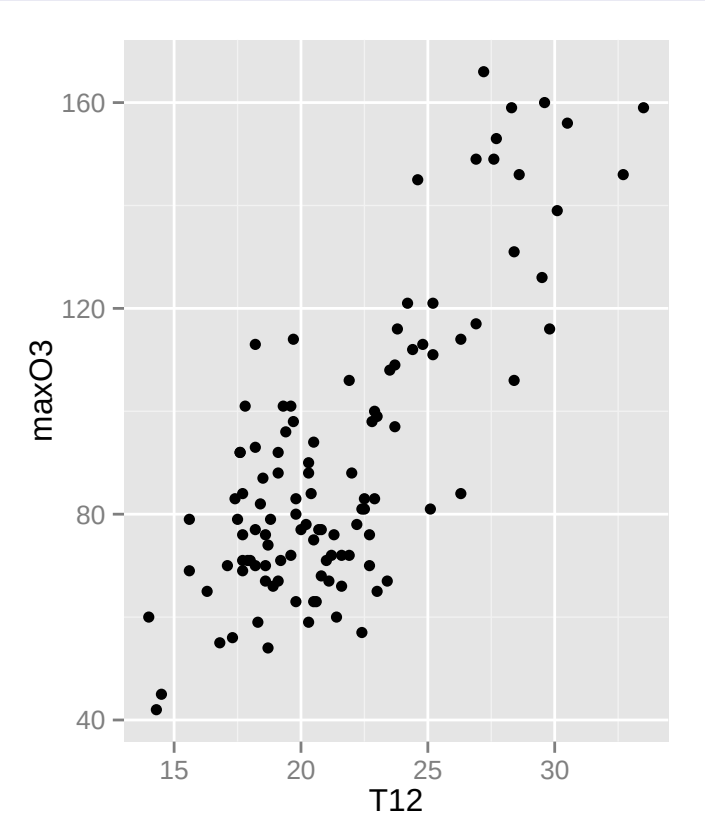
Université Paris Nanterre

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

Exemple : Pollution l'ozone

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

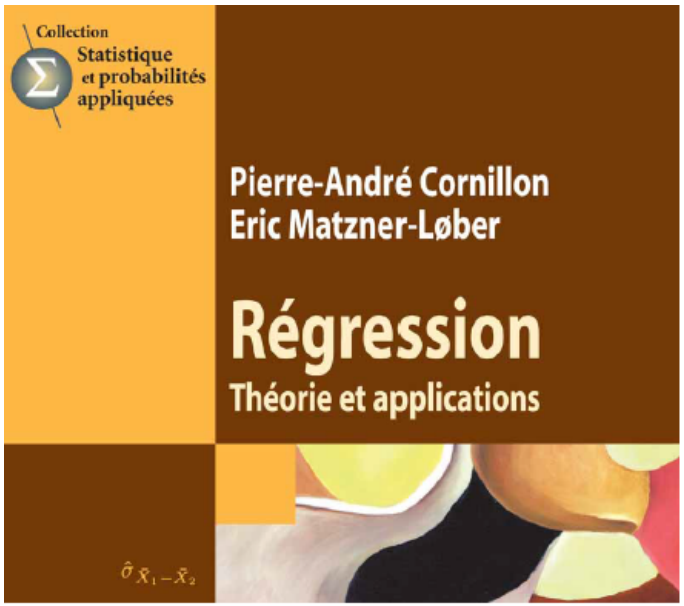
mesurés en un lieu donné et une journée donnée pendant n jours.



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

- Ajuster un modèle pour expliquer Y en fonction de $\mathbf{X}$
- Prédire les valeurs de Y pour de nouvelles valeurs de $\mathbf{X}$.

Bibliographie : Pierre-André Cornillon, Eric Matzner-Lober



Données ozone

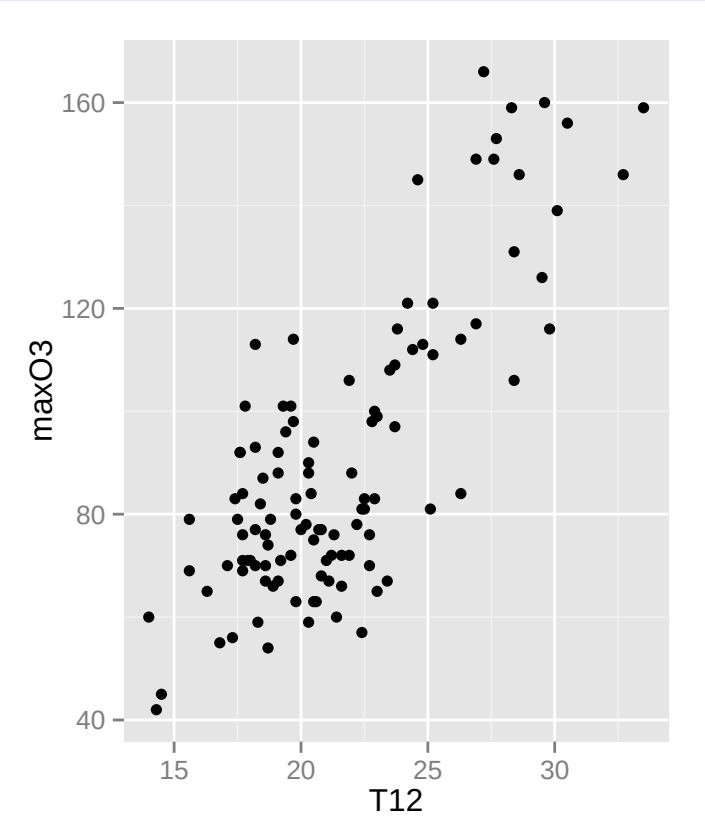
Nous commençons toujours par voir et représenter les données !

```
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

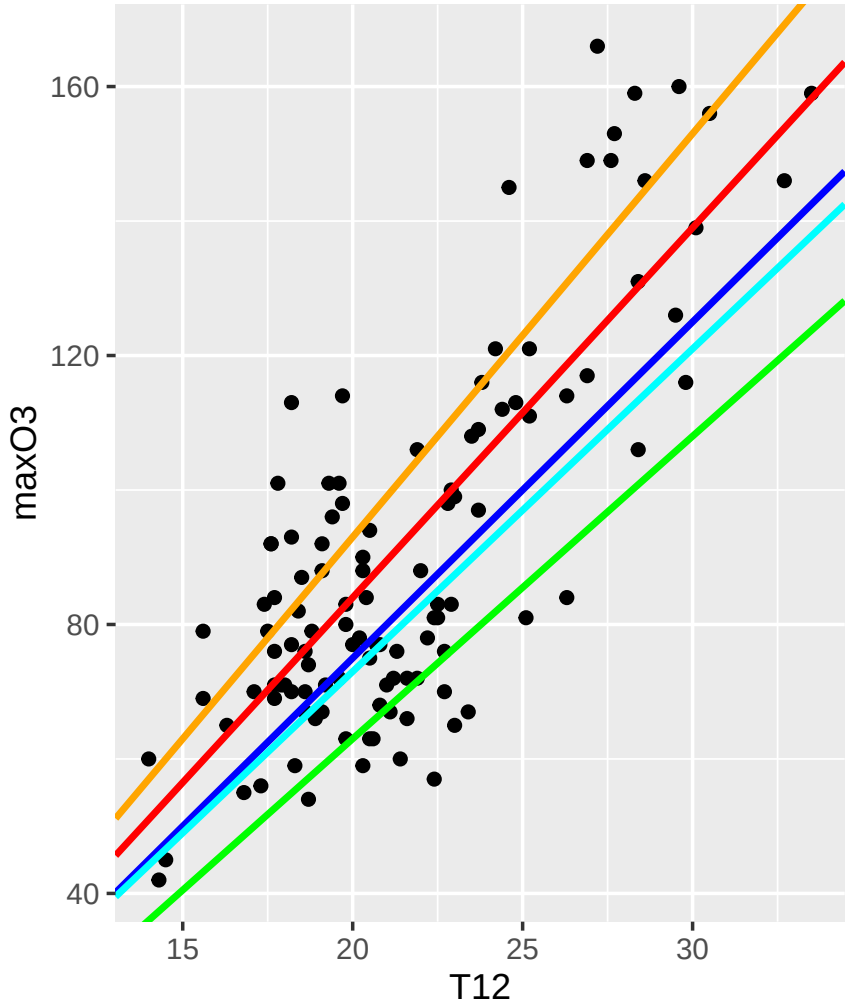
- On souhaite "expliquer" une variable Y à partir de $\mathbf{X}$.
- Nous allons chercher une fonction f tel que

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

- 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
- On choisi f dans la classe $\mathcal{S}$ qui minimise

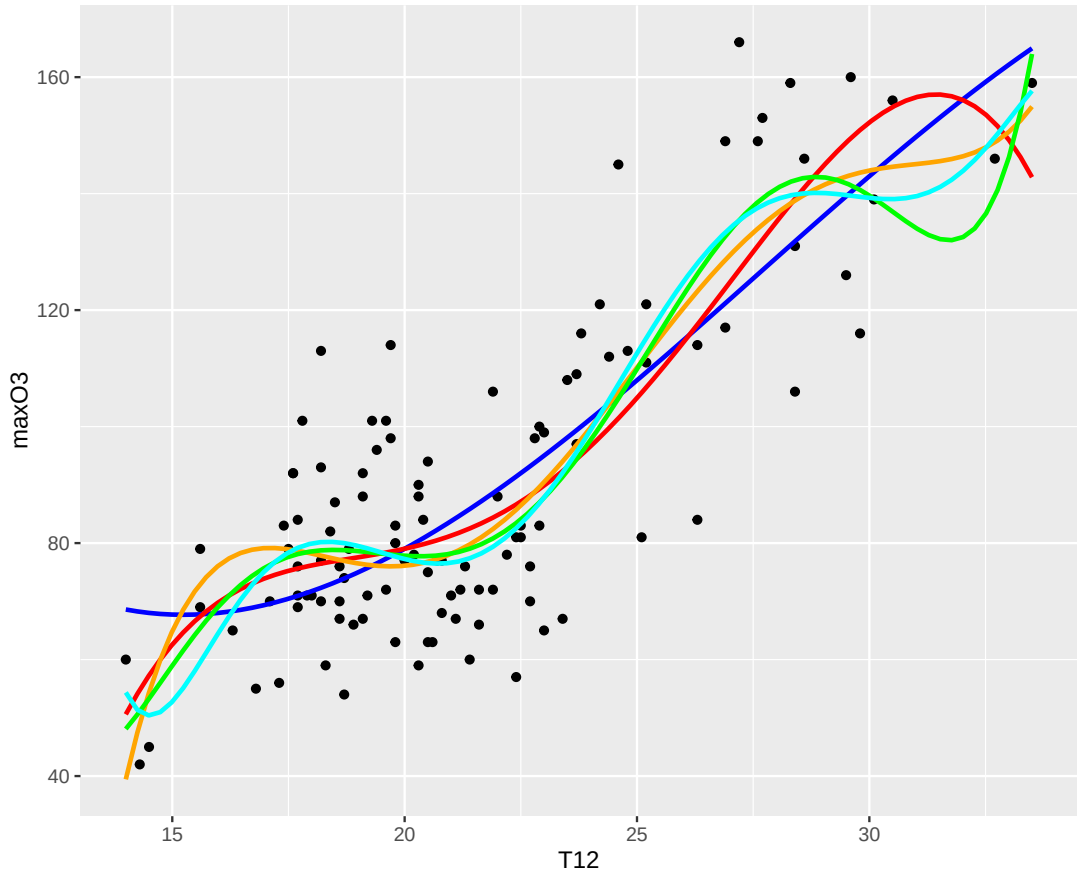
$$\sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2$$

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



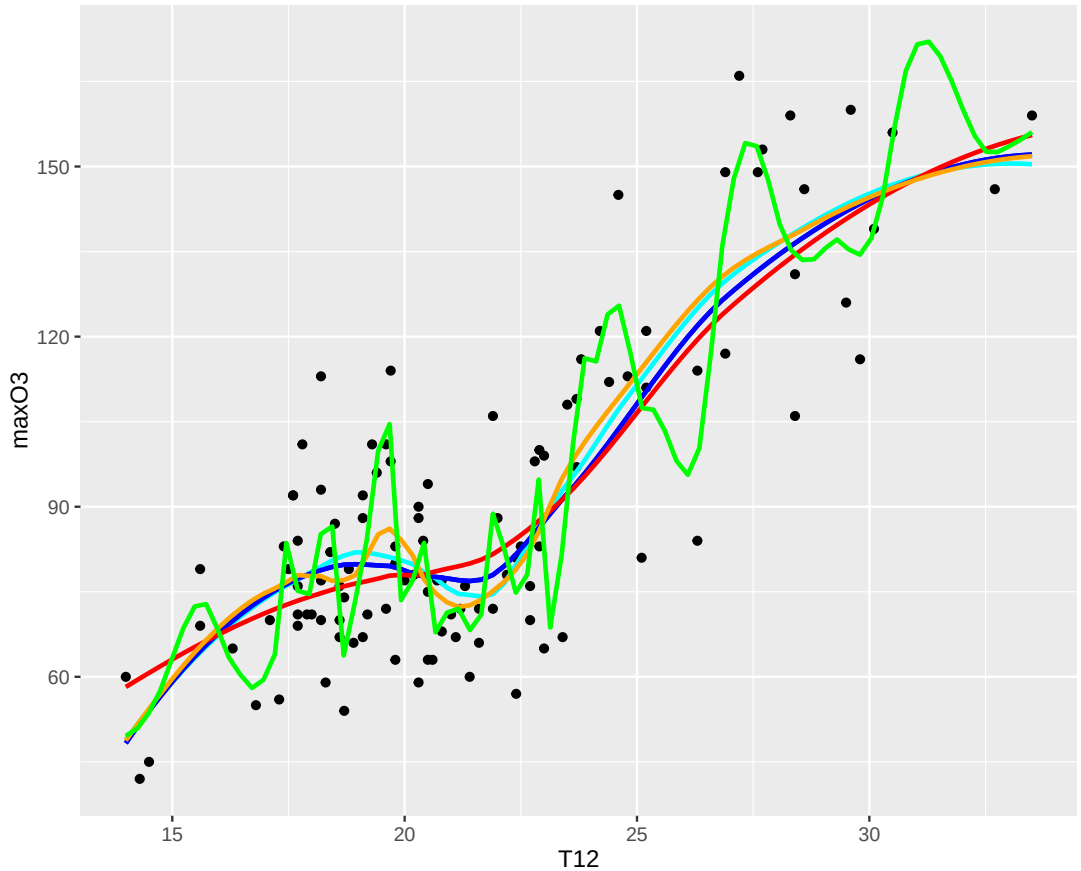
Objectif : Parmi toutes les droites possibles, déterminer la droite qui minimise la somme des écarts aux carrés.

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.

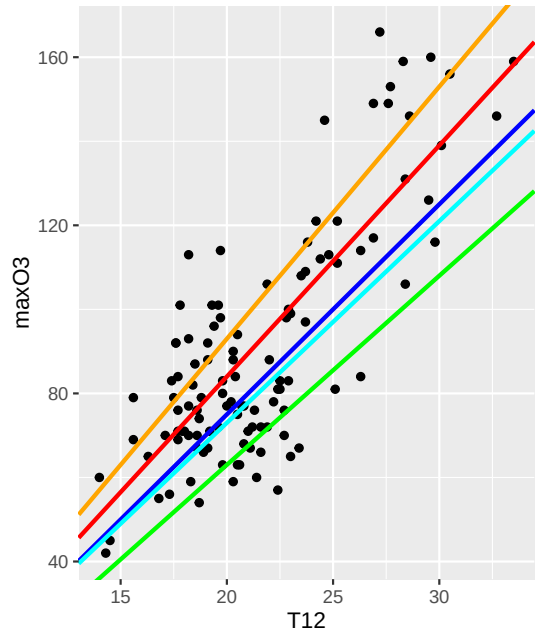
f est choisie dans une classe des fonctionnes $\mathcal{S}$ plus complexe



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

Quelle classe de fonctions $\mathcal{S}$ choisir? Linéaire, Polynomiale, ...
Dans ce cours on va travailler dans de cas simples. Par exemple,
$\mathcal{S}$: Famille des fonctions linéaires

$$\mathcal{S} = \{f : f = \beta_0 + \beta_1 \mathbf{T12}, \quad \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}\}$$



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

Méthode des moindres carrés

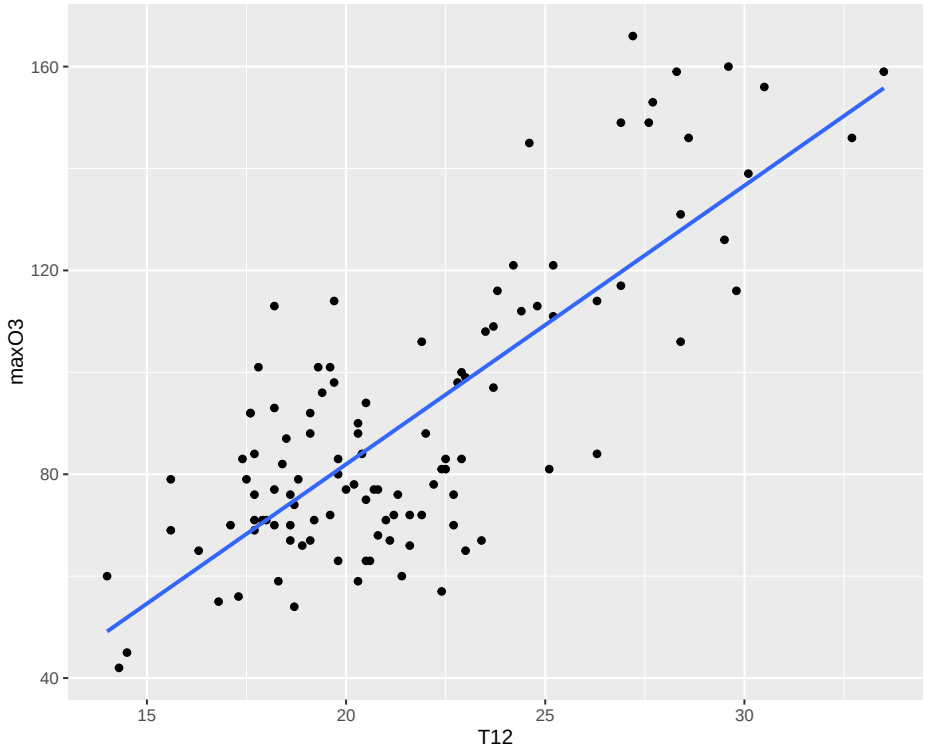
$$\sum_{i=1}^{n}(Y_i - f(\mathbf{X}_i))^2 = \sum_{i=1}^{n}(\text{maxO3}_i - f(\mathbf{T12}_i))^2$$
$$= \sum_{i=1}^{n}(\text{maxO3}_i - (\beta_0 + \beta_1 \mathbf{T12}_i))^2$$

- Choisir β qui minimise la quantité

$$\sum_{i=1}^{n}(\text{maxO3}_i - (\beta_0 + \beta_1 \mathbf{T12}_i))^2$$

- Minimisation solution explicite!

Prédiction



- Prédiction linéaire pour ozone :

$$\widehat{\text{maxO3}} = f(\mathbf{T12}) = \widehat{\beta}_0 + \widehat{\beta}_1 \mathbf{T12}$$

Démarche à suivre :

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

Modèle de régression

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

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

- 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

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n$$

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.

Modèle gaussien de la régression linéaire multiple

On observe des observations bruités

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

avec $\beta_0, \beta_1, \ldots, \beta_d$ 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 $X_1, X_2, \ldots, X_d$.
Soit $\mathbb{X}$ la matrice augmentée (n lignes et $d+1$ colonnes).
Soit $\beta = (\beta_0, \beta_1, \ldots, \beta_d)$ le vecteur de coefficients inconnus.

Modèle Théorique

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

Modèle Théorique (sous forme matricielle)

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

Considérons le modèle théorique de régression linéaire multiple.

1. Coefficients estimés (para le méthode de MC) :
$\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_d)$

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

2. Valeur prédite pour l'i-ème individu

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

3. Somme des carrés des résidus

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

4. Estimateur de σ^2 est

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

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

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 ' '
```

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 '

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 $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_d X_d + \varepsilon$,
- $\mathrm{SCR} = \sum(\widehat{y}_i - y_i)^2 \quad \text{et} \quad \mathrm{SCE} = \sum(\widehat{y}_i - \bar{y})^2$
- 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.

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, ...

Attention :

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

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

```
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
```

MLG3 $\text{O3}_i = \beta_0 + \beta_1 \text{T12}_i + \beta_2 \text{Vx12}_i + \beta_3 \text{Ne12}_i + \varepsilon_i$

```
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

MLG1 $O3_i = \beta_0 + \beta_1 T12_i + \beta_2 Vx12_i + \varepsilon_i$
MLG3 $O3_i = \beta_0 + \beta_1 T12_i + \beta_2 Vx12_i + \beta_3 Ne12_i + \varepsilon_i$

```
Model 1: O3 ~ T12 + Vx12
Model 2: O3 ~ 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).

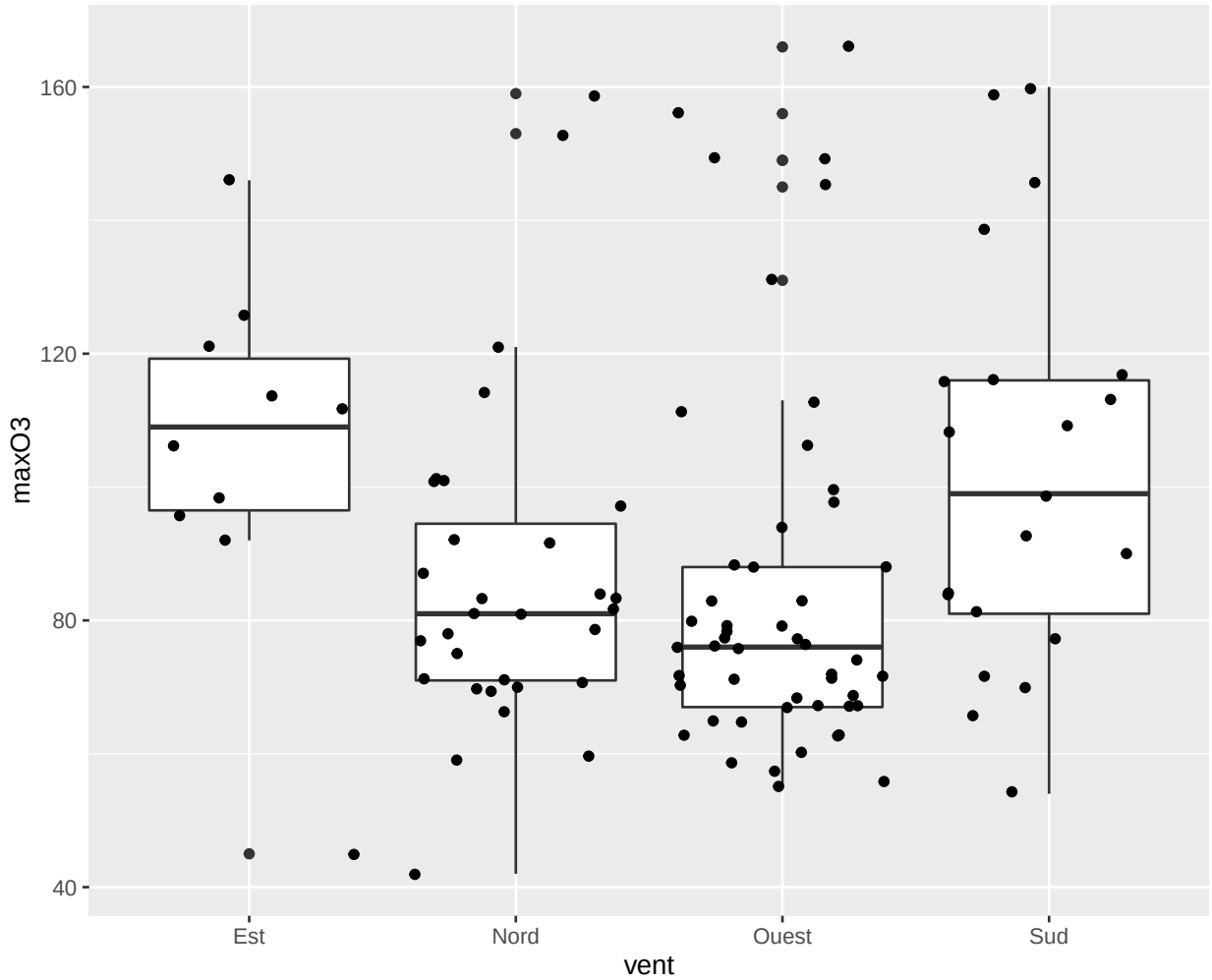
Régression sur des variables qualitatives

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

$$\mathbf{X} = (\mathbf{1}_{A_2}, \ldots, \mathbf{1}_{A_k})$$

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.

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

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 ' '
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 $n = 112$

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

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

```
---
            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