Régression Linéaire

N. Jégou

Université Rennes 2

M1 2SEP

Les données

- On a mesuré durant l'été 2001 en Bretagne :
  - Le pic d'ozone du jour : `max03`
  - La température à midi : `T12`
- Au total $n = 112$ jours d'observations

Individu	`max03`	`T12`
20010601	87	18.5
20010602	82	18.4
20010603	92	17.6
20010604	114	19.7
20010605	94	20.5
20010606	80	19.8
...	...	...

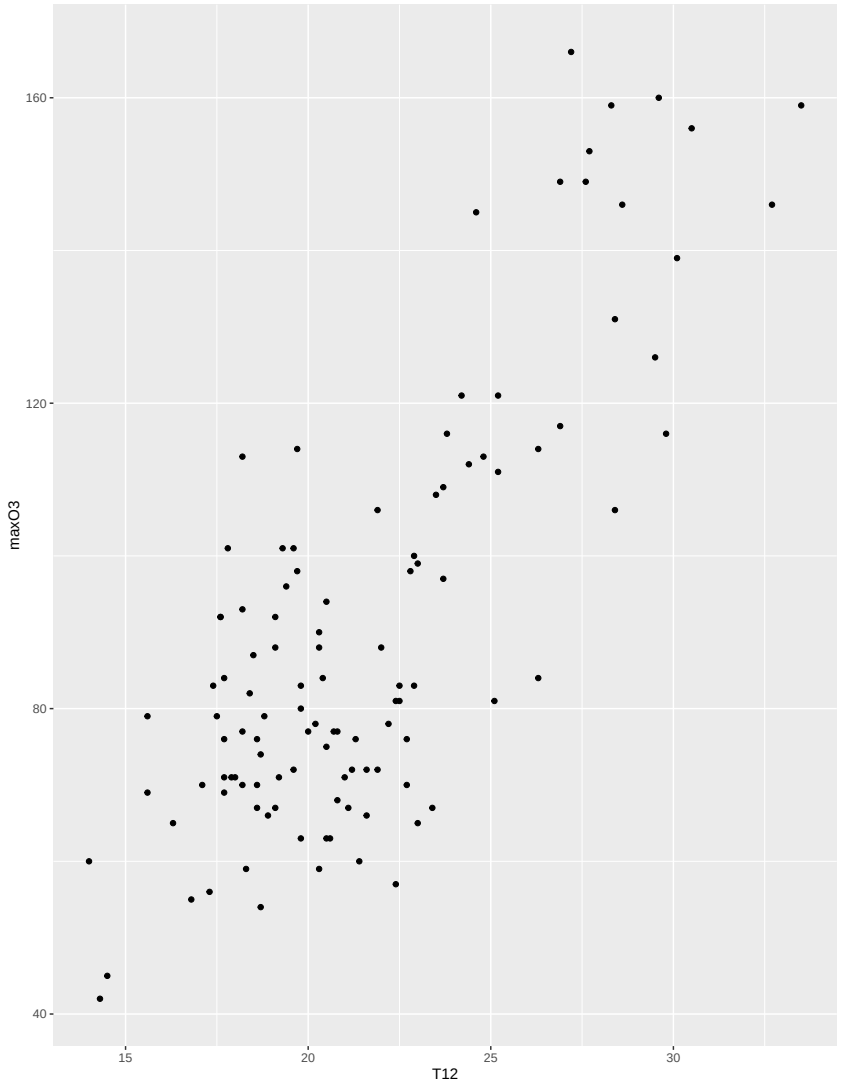
Objectifs

1. Deux variables quantitatives :
   X : `T12`
   Y : `max03`
2. Expliquer la hauteur Y par X
3. On dispose de données allant par paire :

$$(x_i, y_i)_{i=1,\dots,n} \in \mathbb{R}^2$$

4. Utiliser la température : $\rightarrow$ prévoir pics de pollution

Représentation graphique



Représentation des mesures pour les $n = 112$ jours.

Le modèle

- Idée : $\exists f : \mathbb{R} \to \mathbb{R}$ telle que $Y \approx f(X)$
- Modèle de régression : on suppose l'existence d'une fonction $f : \mathbb{R} \to \mathbb{R}$ telle que

$$Y = f(X) + \varepsilon \text{ avec } \varepsilon \text{ variable aléatoire}$$

- Modèle linéaire : f est supposée affine

$$\exists (\beta_0, \beta_1) \in \mathbb{R}^2 : Y = \beta_0 + \beta_1 X + \varepsilon$$

Le modèle

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \cdots, n$$

- Les X_i supposées fixées : on note x_i
  $\mathcal{H}_1$: on suppose que 2 valeurs de x_i au moins sont différentes
- Les Y_i sont aléatoires, ε_i sont aléatoires
- Les paramètres (β_0, β_1) sont non aléatoires

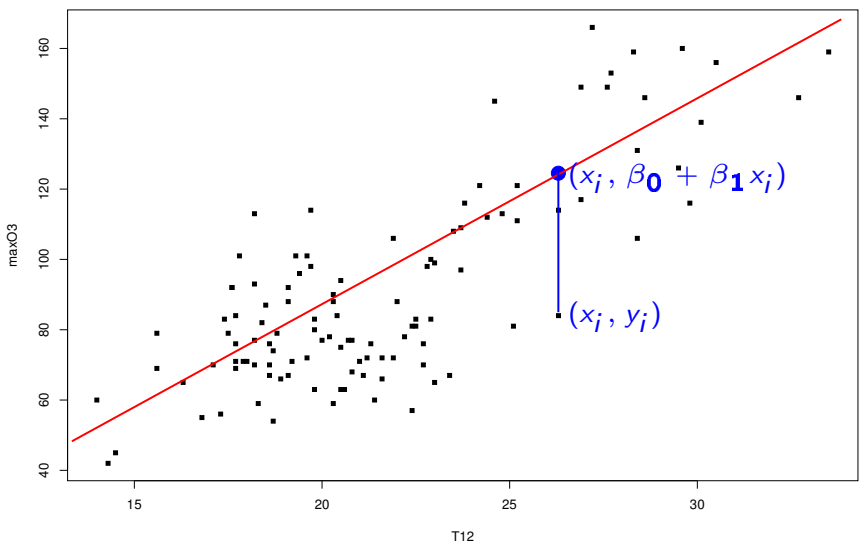
- Les données (x_i, Y_i) sont supposées centrées autour d'une droite de paramètres (β_0, β_1) fixes mais inconnus

  Les résidus sont supposés non corrélés

$$\mathcal{H}_2 : \mathbb{E}(\varepsilon_i) = 0 \ ; \ \mathbf{cov}(\varepsilon_i, \varepsilon_j) = 0$$

Estimateur des moindres carrés

- Les paramètres β_0 et β_1 du modèle sont inconnus : on utilise les données pour les estimer
- Il est naturel de chercher une droite "proche" du nuage



Estimateur des moindres carrés

- La droite la plus proche des points (au sens quadratique) s'obtient en minimisant le coût :

$$(\beta_0, \beta_1) \mapsto S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

- La solution est L'estimateur des moindres carrés :

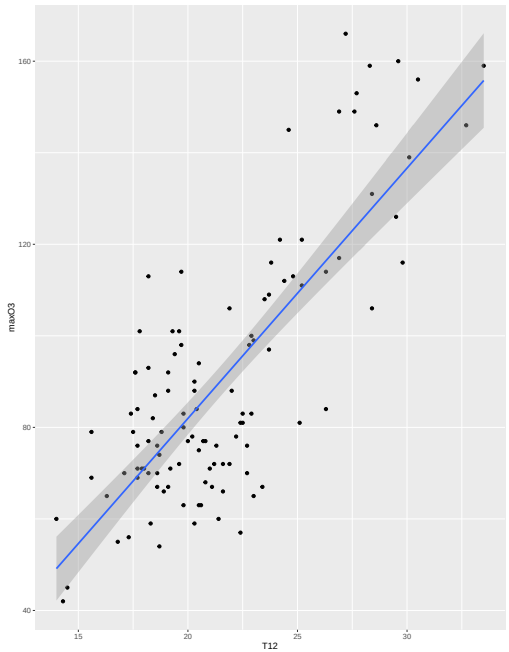
$$\hat{\beta}_1 = \frac{\mathbf{cov}(X, Y)}{\sigma_X^2} = \frac{\frac{1}{n}\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\frac{1}{n}\sum(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Estimateur des moindres carrés

- La réalisation sur les données de couple aléatoire est :

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum(x_i - \bar{x})^2} = 5,5 \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = -27,4$$

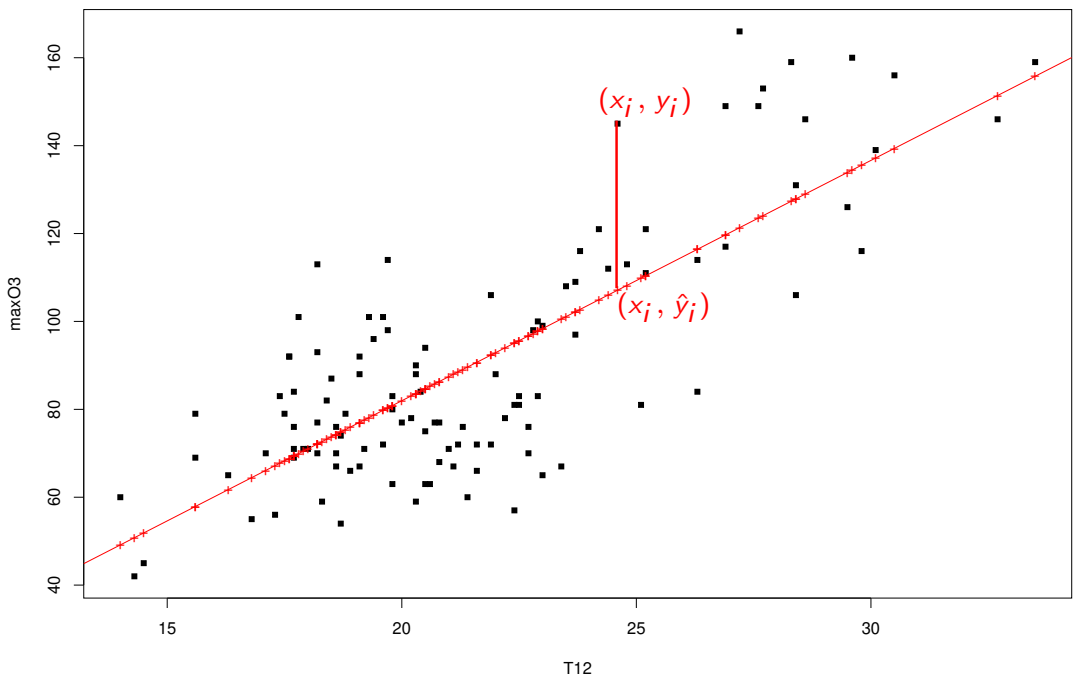
- Elle caractérise la droite des moindres carrés



Qualité de l'ajustement : le R^2

Disposant de l'estimation $\left(\hat{\beta}_0, \hat{\beta}_1\right)$, les valeurs ajustées sont :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$



Qualité de l'ajustement : le R^2

- Le coefficient de corrélation linéaire R^2 mesure la qualité de l'ajustement du modèle
- C'est le rapport de la variance des valeurs ajustées et de la variance des valeurs initiales :

$$R^2 = \frac{\text{var}(\hat{Y})}{\text{var}(Y)} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Qualité de l'ajustement : le R^2

- C'est la part de variabilité expliquée par le modèle
- $0 \leq R^2 \leq 1$ avec
  - $R^2 = 0 \Leftrightarrow \hat{\beta}_1 = 0$
    La droite de régression est horizontale
    Pas d'effet de X
  - $R^2 = 1 \Leftrightarrow$ les points (x_i, y_i) sont alignés sur la droite ajustée (de pente $\hat{\beta}_1 \neq 0$)
    Le modèle restitue la totalité de la variabilité des y_i.
- Dans l'exemple $R^2 \approx 0,6$

Prévision

1. Pour une température x^*
2. Le modèle donne la prévision $\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$
3. Une prévision n'a de sens que pour x^* dans la plage des x_i observés !

Objectifs - Hypothèse $\mathcal{H}_3$

- On souhaite, à partir de l'échantillon, tirer des conclusions valables pour la population
- Questions :
  - `T12` a-t-elle une influence significative sur `max03` ?
  - $\hat{\beta}_1 = 5,5$ significativement $\neq 0$?
  - $R^2 = 0,6$ est-il significativement $\neq 0$?
  - Intervalles de confiance sur β_1 ?

- Répondre à ces questions suppose de faire une hypothèse sur la loi des résidus :

$$\mathcal{H}_3 : \varepsilon_i \underset{\texttt{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

- La variance du bruit σ^2 est inconnue

Estimateur de la variance σ^2

- On estime la variance σ^2 du bruit en "moyennant" les résidus observés :

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

- Dans l'exemple, la réalisation de $\hat{\sigma}^2$ est

$$\hat{\sigma}^2 = 17,57^2$$

Propriétés de l'EMC

De l'hypothèse $\mathcal{H}_3$ on déduit

- des propriétés de biais-variance de $\hat{\beta}_1$:

$$\begin{aligned} \mathbb{E}(\hat{\beta}_1) &= \beta_1 \\ \text{var}(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned}$$

- un intervalle de confiance pour β_1 : (à 95%)

$$\hat{\beta}_1 - t_{0.975}(n-2)\sqrt{\hat{\text{var}}(\hat{\beta}_1)} \leq \beta_1 \leq \hat{\beta}_1 + t_{0.975}(n-2)\sqrt{\hat{\text{var}}(\hat{\beta}_1)}$$

Le coin de R

Ecriture du modèle

```
> don <- read.table("ozone.txt")
> don <- as_tibble(don)
> reg <- lm(maxO3 ~ T12, data = don)
> names(reg)
 [1]  "coefficients"  "residuals"     "effects"
 [4]  "rank"    "fitted.values"    "assign"
 [7]  "qr"       "df.residual"  "xlevels"
 [10] "call"           "terms"          "model"
```

Le coin de R

- Estimation des paramètres

```
> reg$coefficients
(Intercept)         T12
 -27.419636    5.468685
```

- Intervalles confiance sur les paramètres

```
> confint(reg, level = 0.95)
                  2.5 %    97.5 %
(Intercept) -45.321901 -9.517371
T12           4.651219  6.286151
```

- Valeurs ajustées

```
 > reg$fitted.values
       1        2        3        4        5
73.75103 73.20417 68.82922 80.31346 84.68840
```

Le coin de R

Résumé du modèle

```
> summary(reg)

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
```

Le coin de R

- On peut utiliser le modèle pour faire des prévisions
- Pour des valeurs de T12 égales à 20 et 25 degrés, le modèle prévoit

```
> aprevoir <- data.frame(T12 = c(20,25))
> predict(reg, newdata = aprevoir)
       1        2
81.95406 109.29749
```

Tests

- Dans le modèle linéaire $Y = \beta_0 + \beta_1 X + \varepsilon$, existe-t-il un lien entre `T12` et `maxO3` ?
- Si $\beta_1 = 0$, le modèle s'écrit

$$Y = \beta_0 + \varepsilon \qquad \text{Modèle 0}$$

$\Rightarrow$ pas de lien (linéaire)

- Si $\beta_1 = 0$, le modèle s'écrit

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \text{Modèle 1}$$

$\Rightarrow$ existence d'un lien (linéaire)

Tests

- Problème : β_1 est inconnu
- On dispose d'une estimation $\hat{\beta}_1 \approx 5,5$ de β_1
- Problème $\hat{\beta}_1$ est aléatoire
- Au regard des données,
  - $\hat{\beta}_1$ est-elle suffisamment éloignée de 0 que l'on privilégie le modèle 1 ?
  - $\hat{\beta}_1$ est-elle significativement différente de 0 ?

Question formalisée

Question (Modèle 0 = Modèle 1) $\Leftrightarrow$ ($\beta_1 = 0$)

ou

(Modèle 0 < Modèle 1) $\Leftrightarrow$ ($\beta_1 \neq 0$)?

Formalisation pour résoudre le pb

1. **Modèle :** Régression linéaire simple

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma) \quad i = 1, \cdots, n$$

2. **Deux hypothèses :**
   - $\mathcal{H}_0 : \beta_1 = 0$
   - $\mathcal{H}_1 : \beta_1 \neq 0$

But Choisir entre $\mathcal{H}_0$ et $\mathcal{H}_1$

Test T de nullité de β_1

1. **Modèle :** Régression linéaire simple

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma) \quad i = 1, \cdots, n$$

2. Choix d'**une erreur de première espèce** $\alpha = 5\%$
3. **Deux hypothèses :**
   - $\mathcal{H}_0 : \beta_1 = 0$
   - $\mathcal{H}_1 : \beta_1 \neq 0$
4. Statistique de test (et loi sous $\mathcal{H}_0$)

$$T = \frac{\hat{\beta}_1}{\sqrt{\hat{\text{var}}(\hat{\beta}_1)}} \overset{\mathcal{H}_0}{\sim} \mathcal{T}(n-2)$$

5. Observation de la stat. de test: T_{obs}
   $\rightarrow$ probabilité critique (p-value)
6. Conclusion

Tests T avec R

- Le résultat du test est donné dans le résumé :

```
> summary(reg)

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
```

- La probabilité critique est $< 2.10^{-16}$: on rejette $\mathcal{H}_0$