SCI03 - Analyse de données expérimentales

Régression linéaire

Thierry Denœux

Automne 2014

Plan

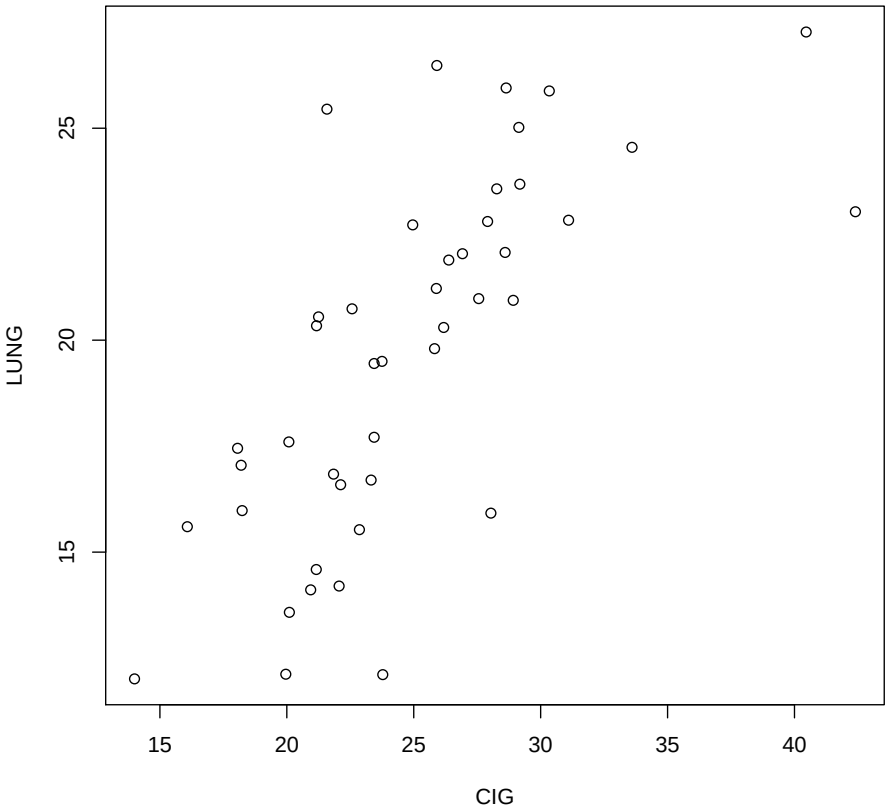
Première partie I

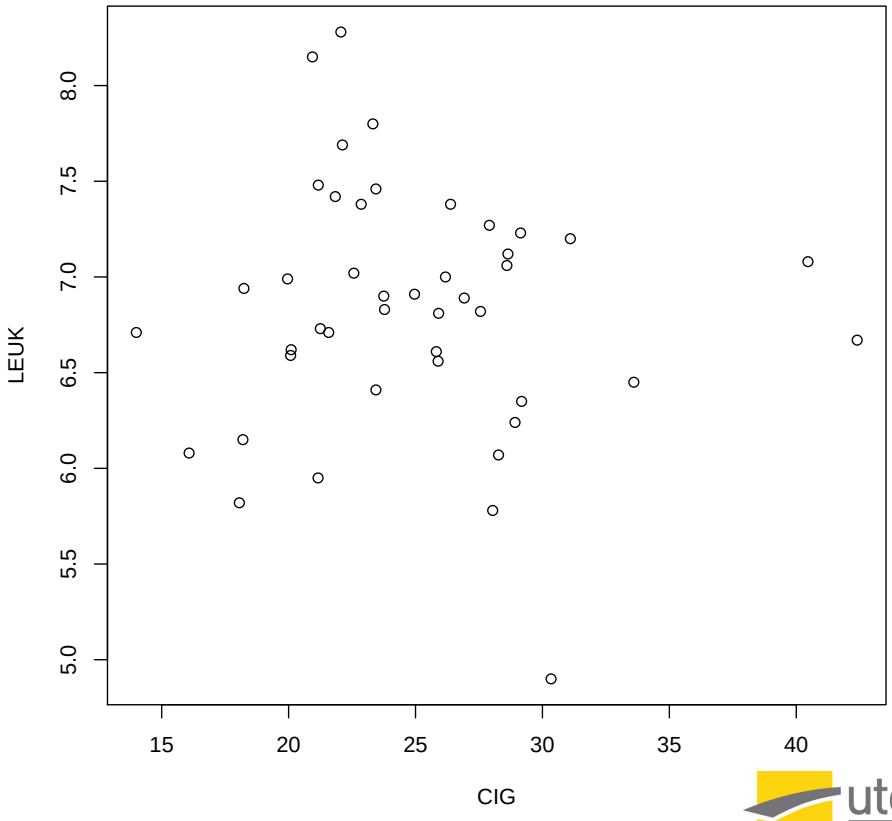
Régression linéaire simple

Relation tabac-cancer

- Données : consommation de cigarettes par habitant et nombre de décès par cent mille habitants pour différentes formes de cancer (BLAD : vessie ; LUNG : poumon ; KID : rein ; LEUK : leucémie) dans 43 états américains en 1960.
- Source : J.F. Fraumeni, "Cigarette Smoking and Cancers of the Urinary Tract : Geographic Variations in the United States," Journal of the National Cancer Institute, 41, 1205-1211.

Représentation graphique des données





Problèmes posés et démarche

- Problèmes posés :
  - Existe-t-il un lien entre la consommation de tabac et différentes formes de cancer ?
  - Prédiction de l'impact d'une modification de la consommation sur l'incidence des cancers.
- Démarche suivie
  - Spécifier le modèle
  - Estimer les paramètres du modèle (intervalles de confiance)
  - Vérifier qu'il y a bien une relation entre les deux variables (tests d'hypothèses)
  - Vérifier la validité du modèle retenu (diagnostic)
  - Prévision

Plan

Le modèle de la régression linéaire

- Données :
  - Deux variables observées sur n individus
  - X : variable explicative, indépendante
  - Y : variable à expliquer, variable dépendante
  - Données : $(x_1, y_1), \ldots, (x_n, y_n)$
- Modèle de génération des données : les x_i étant fixés, chaque y_i est supposé être une réalisation d'une v.a. Y_i :

$$Y_i = a + bx_i + \varepsilon_i \qquad \forall i$$

avec $\mathbb{E}(\varepsilon_i) = 0$, $\text{var}(\varepsilon_i) = \sigma^2$ et $\text{cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i, j$.

- La droite d'équation $y = ax + b$ est appelée droite de régression.

Estimation des paramètres

- On cherche les valeurs des coefficients a et b qui minimisent

$$E(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Solution :

$$\widehat{b} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \qquad \widehat{a} = \overline{Y} - \widehat{b}\overline{x}$$

- $\widehat{a}$ et $\widehat{b}$: estimateurs des moindres carrés.
- Droite $y = \widehat{a} + \widehat{b}x$: droite des moindres carrés de Y en x
- On note $\widehat{Y}_i = \widehat{a} + \widehat{b}x_i$ et $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$ (résidus)
- Estimateur sans biais de σ^2 :

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2$$

Coefficients de la droite des moindres carrés en R

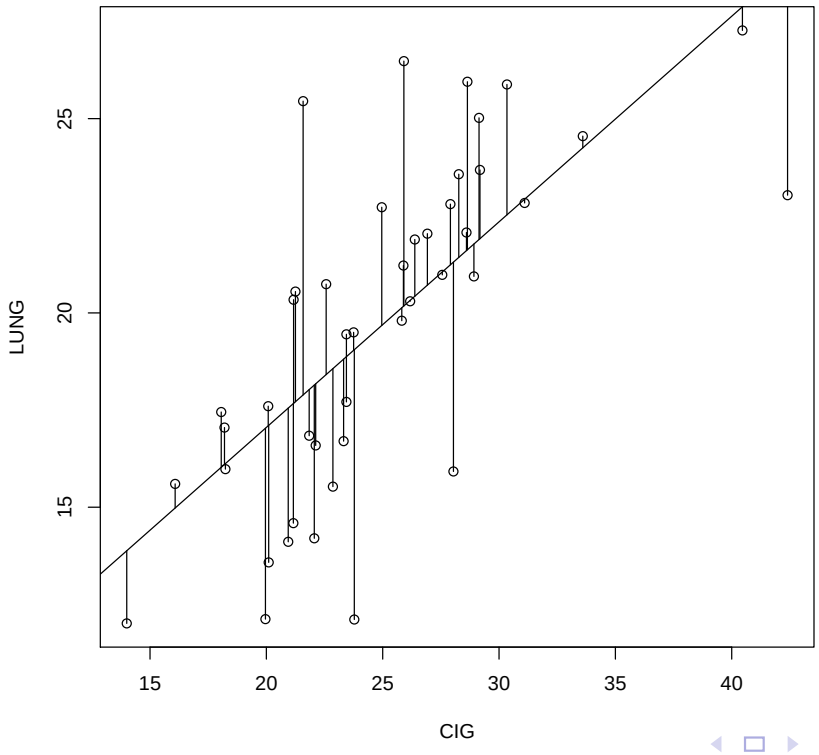
```
> reg.lung <- lm(LUNG ~ CIG)
> reg.lung

Call:
lm(formula = LUNG ~ CIG)

Coefficients:
 (Intercept)   CIG
  6.4717       0.5291
```

Droite des moindres carrés

```
> plot(CIG,LUNG)
> segments(CIG,fitted(reg.lung),CIG,LUNG)
> abline(reg.lung)
```



Plan

Équation de l'analyse de la variance

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}_{\text{Variance totale}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2}_{\text{Variance expliquée par la régression}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}_{\text{Variance résiduelle}}$$

$$S_Y^2 = S_{reg} + S_{res}$$

Coefficient de détermination

- La droite passe par tous les points :

$$\widehat{y_i} = y_i \Rightarrow s_{res} = 0 \Rightarrow s_Y^2 = s_{reg}$$

  Toute la dispersion est expliquée par la régression
- La pente de la droite est nulle :

$$\widehat{y_i} = \widehat{y} \Rightarrow s_{reg} = 0 \Rightarrow s_Y^2 = s_{res}$$

  La dispersion n'est absolument pas expliquée par la régression
- Coefficient de détermination

$$R^2 = \frac{S_{reg}}{S_Y^2}$$

- R^2 est la proportion de la dispersion des y_i expliquée par la dispersion des x_i.

Exemple en R : modèle LUNG vs CIG

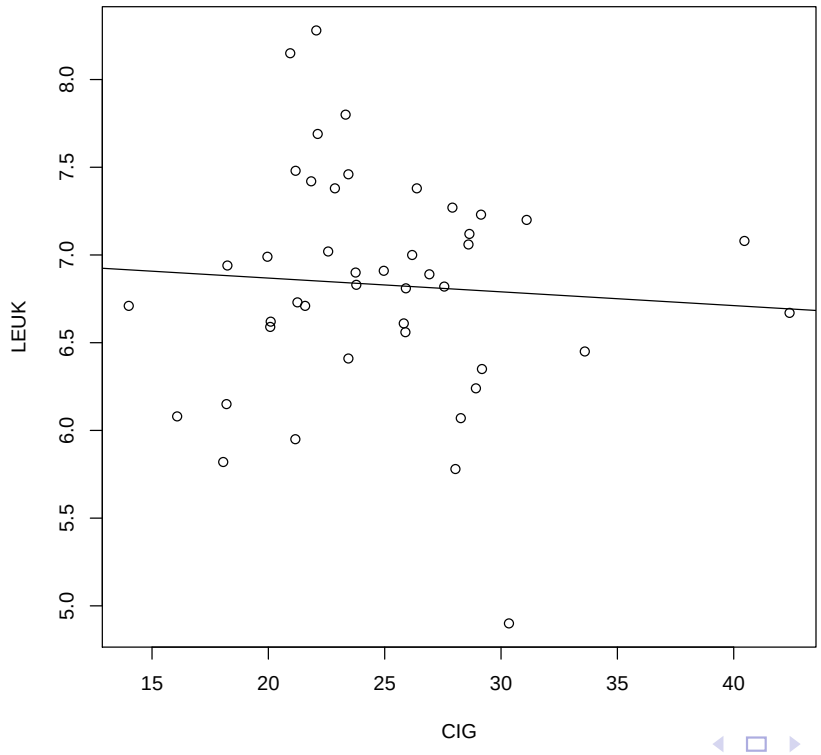
```
> summary(reg.lung)

Call:
lm(formula = LUNG ~ CIG)

Residuals:
 Min      1Q      Median   3Q      Max
 -6.943   -1.656   0.382    1.614   7.561
Coefficients:
              Estimate   Std. Error   t value   Pr(>|t|)
 (Intercept)   6.4717      2.1407      3.023    0.00425    **
 CIG           0.5291      0.0839      6.306    1.44e-07   ***
Residual standard error: 3.066 on 42 degrees of freedom
Multiple R-squared: 0.4864, Adjusted R-squared: 0.4741
F-statistic: 39.77 on 1 and 42 DF, p-value: 1.439e-07
```

Exemple en R : modèle LEUK vs CIG

```
> reg.leuk <- lm(LEUK ~ CIG)
> plot(CIG,LEUK)
> abline(reg.leuk)
```



Exemple en R : modèle LEUK vs CIG

```
> summary(reg.leuk)

Call:
lm(formula = LEUK ~ CIG)

Residuals:
     Min       1Q   Median       3Q      Max
 -1.88722 -0.28618  0.03443  0.42240  1.42784
Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
 (Intercept)  7.025163    0.449835   15.617    <2e-16   ***
 CIG         -0.007843    0.017630   -0.445     0.659
Residual standard error: 0.6443 on 42 degrees of freedom
Multiple R-squared: 0.00469, Adjusted R-squared: -0.01901
F-statistic: 0.1979 on 1 and 42 DF, p-value: 0.6587
```

Plan

Hypothèse de normalité des perturbations

- Si on inclut dans le modèle l'hypothèse supplémentaire

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n,$$

il est possible de faire différents tests de significativité de la régression.

- Deux tests :
  - significativité du R^2
  - ordonnée à l'origine.

Significativité du R^2

- Hypothèses : $H_0 : b = 0$ versus $H_1 : b \neq 0$
- Statistique de test

$$T = \frac{\widehat{b}}{\widehat{\sigma}/\sqrt{nS_X^2}} \overset{H_0}{\sim} \mathcal{T}_{n-2}.$$

- Degré de signification : $p = \mathbb{P}_{H_0}(|T| \geq |t|)$.

Test sur l'ordonnée à l'origine

- Hypothèses : $H_0 : a = 0$ versus $H_1 : a \neq 0$
- Statistique de test

$$T = \frac{\widehat{a}}{\frac{\widehat{\sigma}}{\sqrt{n}}\sqrt{1 + \frac{\overline{x}^2}{S_X^2}}} \overset{H_0}{\sim} \mathcal{T}_{n-2}$$

- Degré de signification : $p = \mathbb{P}_{H_0}(|T| \geq |t|)$.

Exemple : modèle LUNG vs CIG

```
> summary(reg.lung)

Call:
lm(formula = LUNG ~ CIG)

Residuals:
 Min      1Q      Median   3Q      Max
 -6.943   -1.656   0.382    1.614   7.561
Coefficients:
              Estimate   Std. Error   t value   Pr(>|t|)
 (Intercept)    6.4717      2.1407      3.023    0.00425    **
 CIG            0.5291      0.0839      6.306    1.44e-07   ***
Residual standard error: 3.066 on 42 degrees of freedom
Multiple R-squared: 0.4864, Adjusted R-squared: 0.4741
F-statistic: 39.77 on 1 and 42 DF, p-value: 1.439e-07
```

Exemple : modèle LEUK vs CIG

```
> summary(reg.leuk)

Call:
lm(formula = LEUK ~ CIG)

Residuals:
     Min       1Q   Median       3Q      Max
 -1.88722 -0.28618  0.03443  0.42240  1.42784
Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
 (Intercept)  7.025163    0.449835   15.617    <2e-16   ***
 CIG         -0.007843    0.017630   -0.445     0.659
Residual standard error: 0.6443 on 42 degrees of freedom
Multiple R-squared: 0.00469, Adjusted R-squared: -0.01901
F-statistic: 0.1979 on 1 and 42 DF, p-value: 0.6587
```

Plan

Principe

- Deux aspects :
  - Vérification des hypothèses et
  - étude de la stabilité des coefficients (influence d'observations aberrantes).
- Rappel des hypothèses :

$$Y_i = a + bx_i + \varepsilon_i \quad \forall i$$

avec

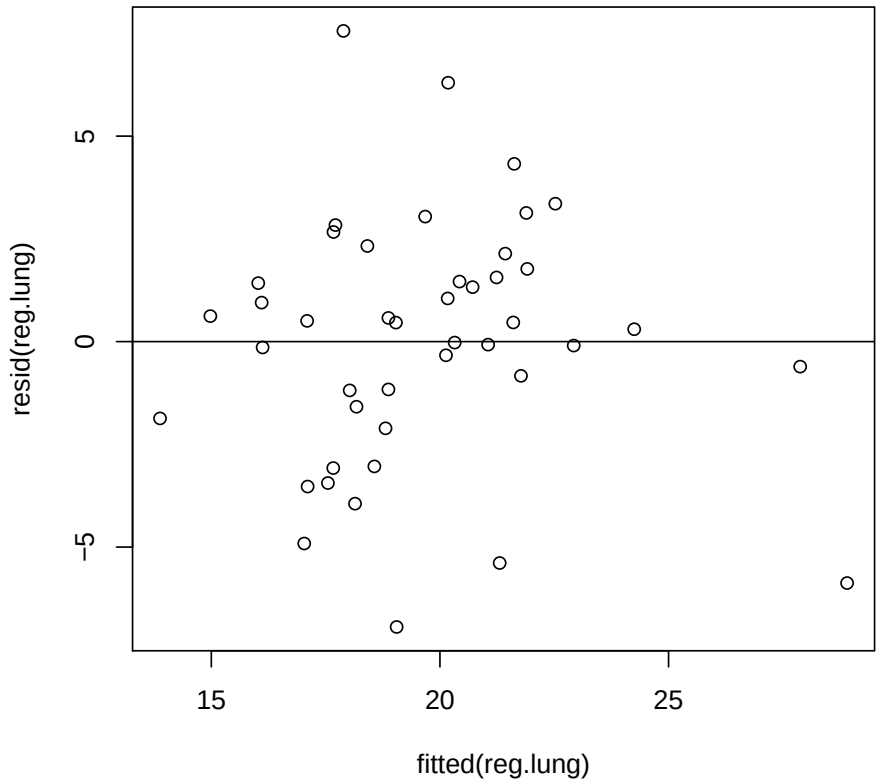
- $\mathbb{E}(\varepsilon_i) = 0$
- $\text{var}(\varepsilon_i) = \sigma^2$
- $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$
- $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Vérification des hypothèses

- Points (x_i, y_i) et droite des moindres carrés
- Analyse des résidus :
  - Résidus bruts ou standardisés (normalisés de manière à avoir une variance égale à 1)
  - Indépendance des résidus : pas de structure particulière dans le graphe des $\widehat{\varepsilon}_i$ en fonction des x_i ou des $\widehat{y}_i$.
  - Normalité des résidus : résidus standardisés entre -2 et 2, et diagramme de normalité.

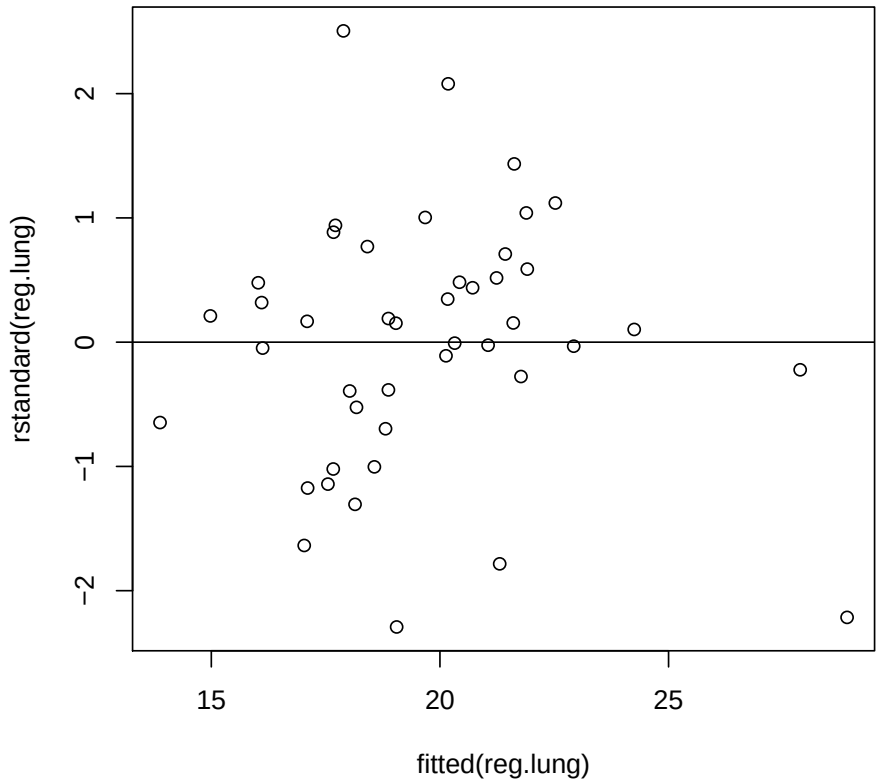
Tracé des résidus

```
> plot(fitted(reg.lung),resid(reg.lung))
```



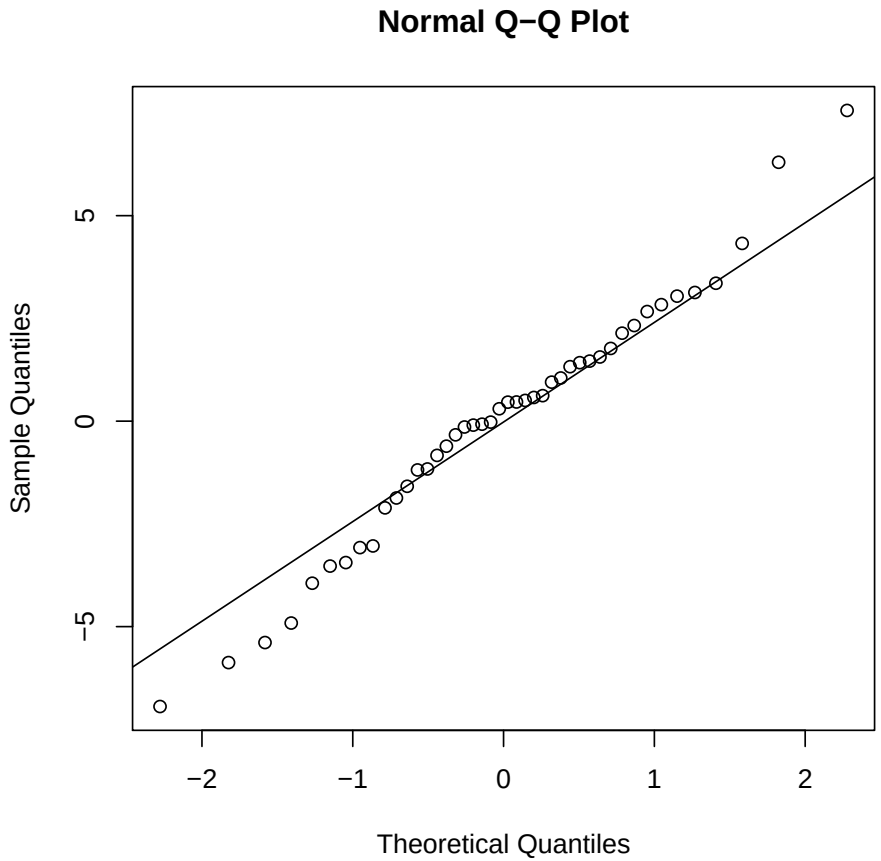
Tracé des résidus standards

```
> plot(fitted(reg.lung),rstandard(reg.lung))
```



Normalité des résidus

```
> qqnorm(resid(reg.lung))
> qqline(resid(reg.lung))
```

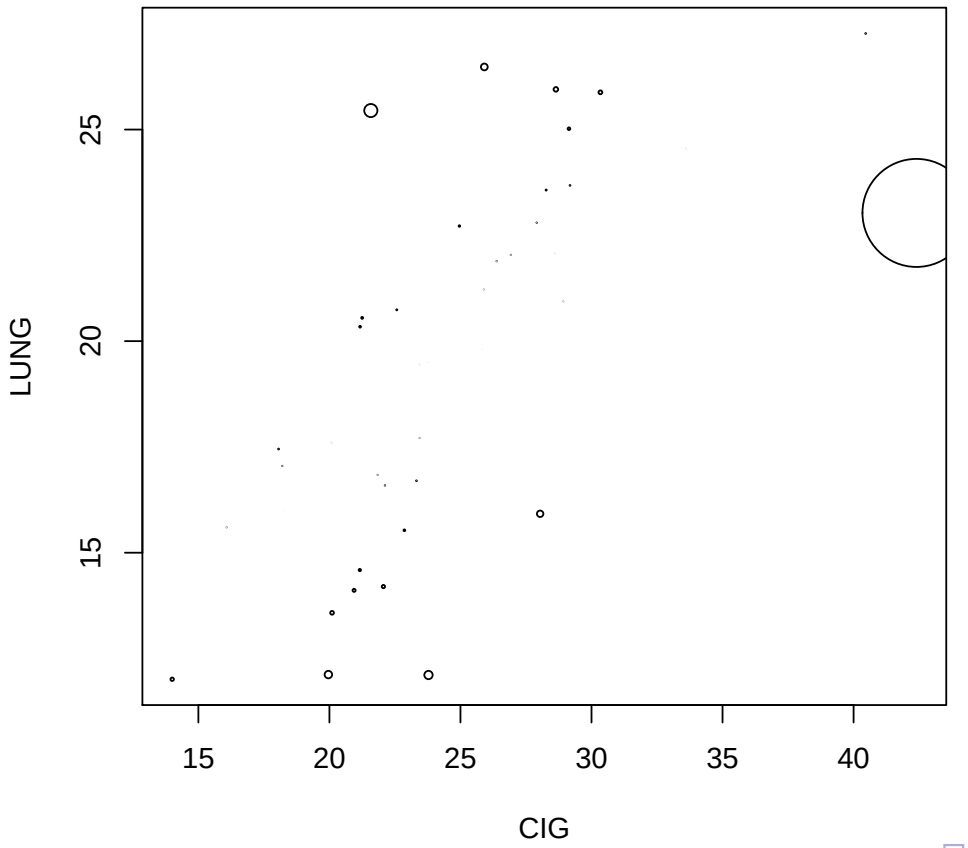


Stabilité des coefficients

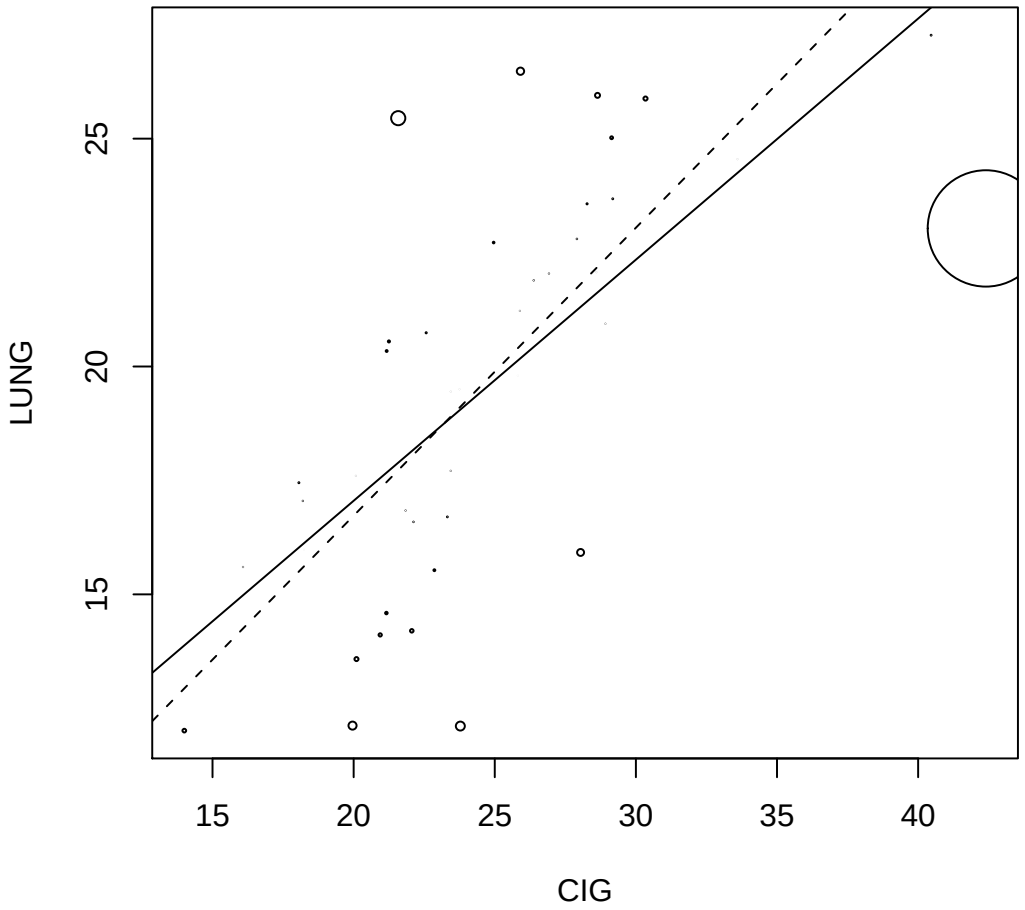
- Certaines observations peuvent avoir une grande influence sur le résultat de la régression, ce qui peut être gênant si leur validité est suspecte. Comment les repérer ?
- Distance de Cook : mesure de distance entre le vecteur $(\widehat{a}, \widehat{b})$ des coefficients de la droite des moindres carrés et le vecteur $(\widehat{a}_{(-i)}, \widehat{b}_{(-i)})$ des coefficients calculés en excluant l'observation i.
- Une distance de Cook supérieure à 1 est généralement considérée comme anormale.

Exemple en R

```
>plot(CIG,LUNG,cex=10*cooks.distance(reg.lung))
```



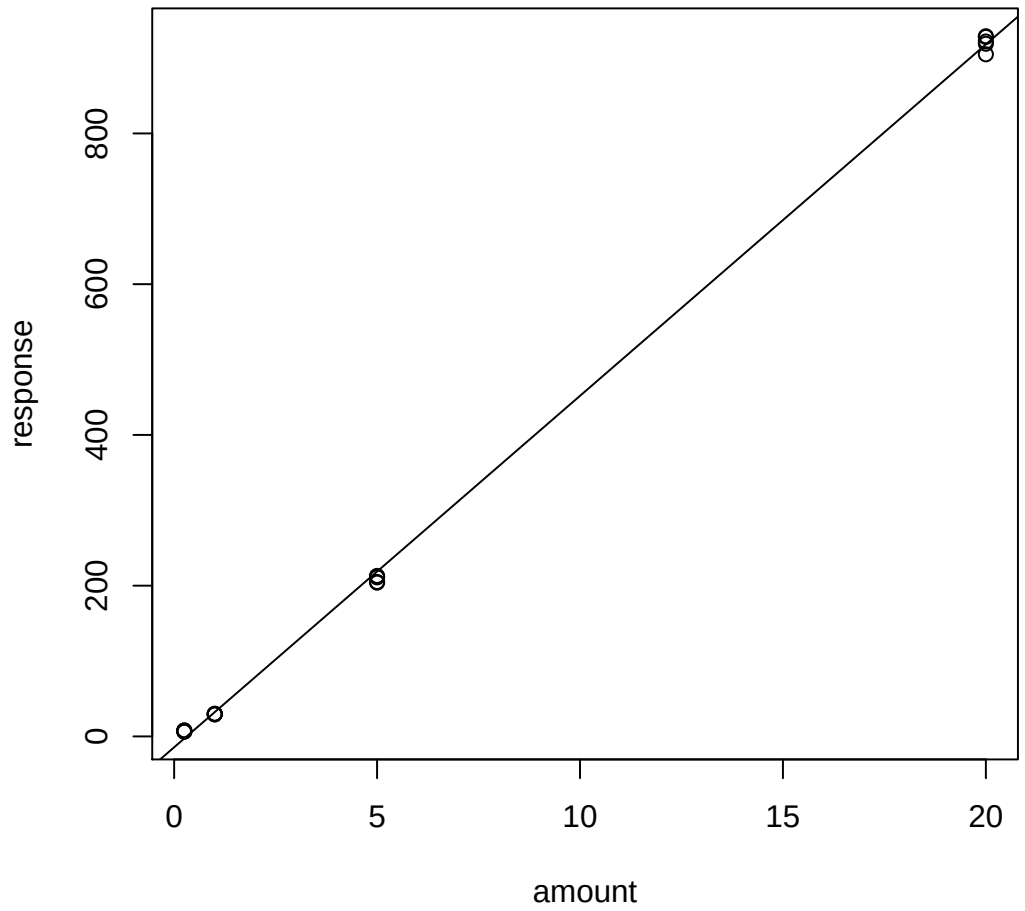
Influence du retrait de l'observation aberrante



Exemple : étalonnage d'un chromatographe

- Resultats d'une étude de chromatographie en phase gazeuse.
- Cinq mesures effectuées sur 4 échantillons contenant différentes quantité d'une certaine substance.
- Variable explicative : quantité de substance déterminée a priori.
- Variable à expliquer : sortie du chromatographe.

Données et droite des moindres carrés



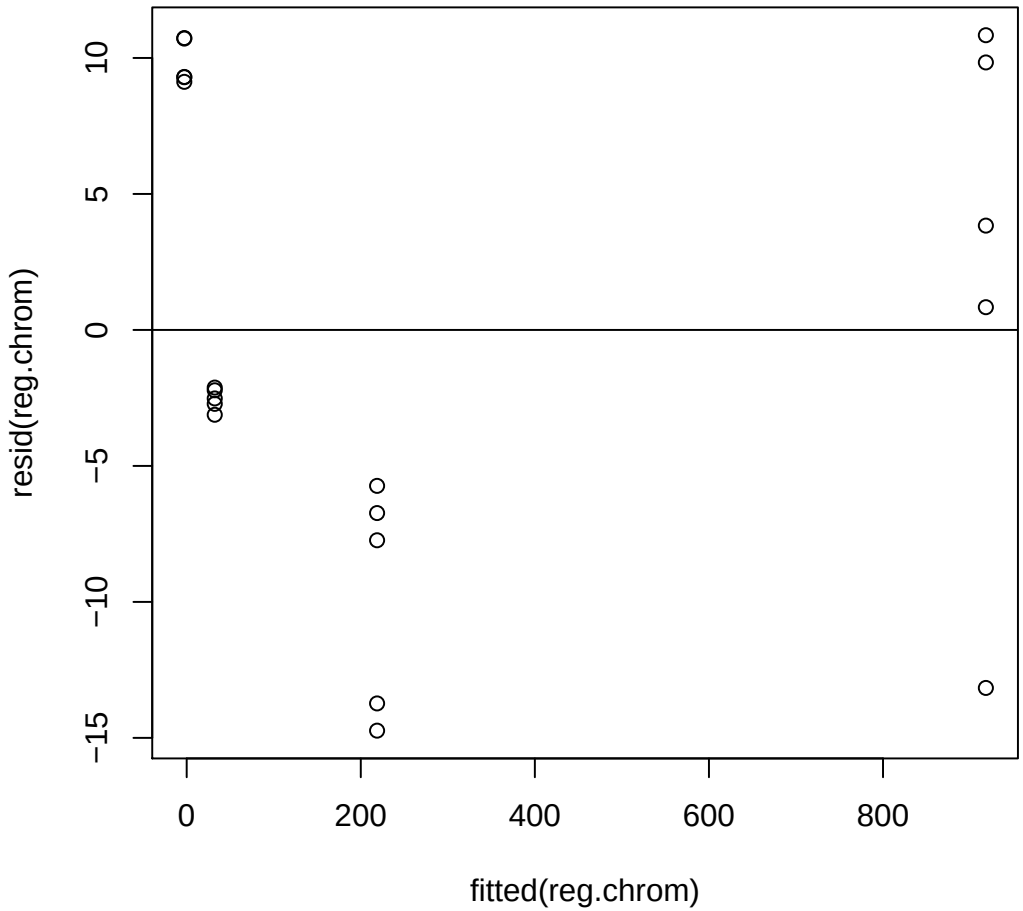
Résultats de la régression

```
> reg.chrom <- lm(response ~ amount)
> summary(reg.chrom)

Call:
lm(formula = response ~ amount)

Residuals:
 Min      1Q      Median   3Q      Max
 -14.733  -5.983  -2.168   9.296   10.837
Coefficients:
             Estimate  Std. Error  t value    Pr(>|t|)
 (Intercept) -14.4107    2.6142     -5.512    3.11e-05    ***
 amount       46.6287    0.2533    184.086   < 2e-16 ***
Residual standard error: 9.023 on 18 degrees of freedom
Multiple R-squared: 0.9995, Adjusted R-squared: 0.9994
F-statistic: 3.389e+04 on 1 and 18 DF, p-value: < 2.2e-16
```

Tracé des résidus



Intervalles de confiance et de prédiction

- x_0 nouvelle valeur, "estimation" de Y_0 par $\widehat{Y}_0 = \widehat{a} + \widehat{b}x_0$
- Intervalle de confiance (contient $\mathbb{E}(Y_0)$ avec une probabilité $1 - \alpha$)

$$\widehat{Y_0} \pm t_{n-2;1-\frac{\alpha}{2}} \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{nS_X^2}}$$

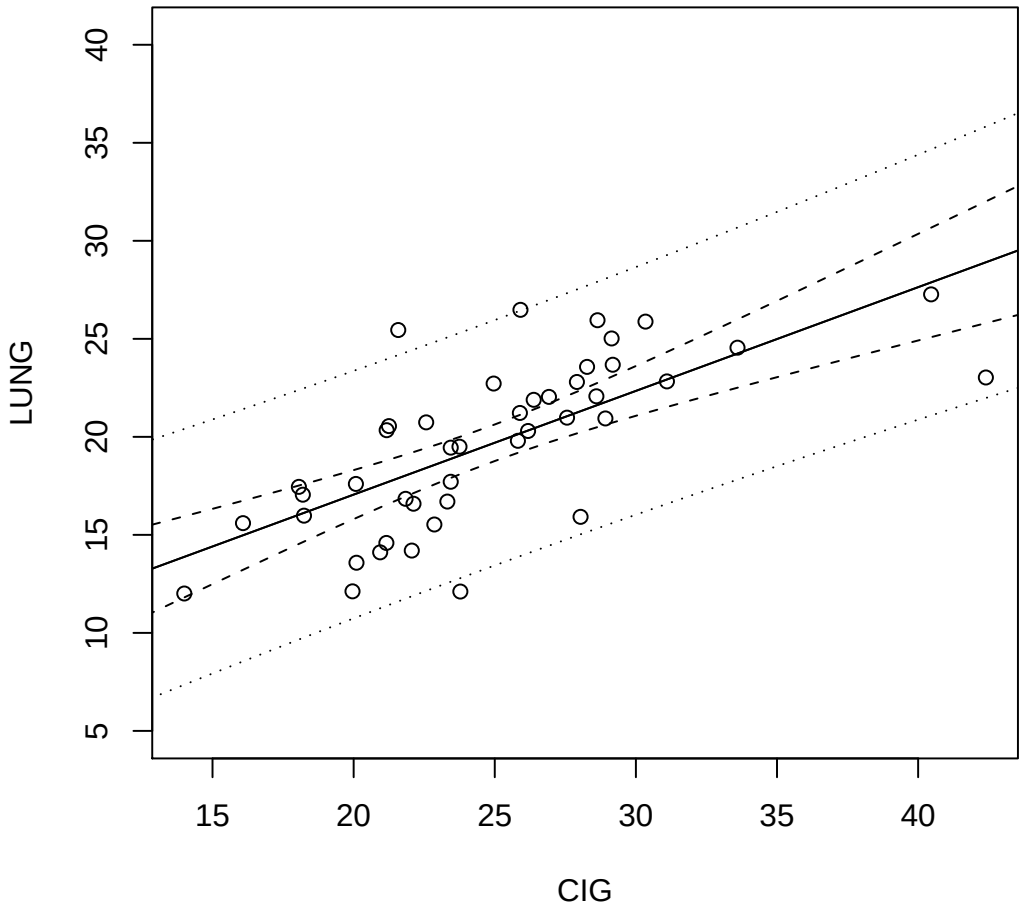
- Intervalle de prédiction (contient Y_0 avec une probabilité $1 - \alpha$)

$$\widehat{Y_0} \pm t_{n-2;1-\frac{\alpha}{2}} \widehat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{nS_X^2}}$$

Prédiction en R

```
> x0 <- data.frame(CIG=15)
> predict(reg.lung,int="c",newdata=x0)
           fit       lwr       upr
 [1,]  14.40786  12.48754  16.32817
> predict(reg.lung,int="p",newdata=x0)
           fit       lwr       upr
 [1,]  14.40786  7.92914  20.88657
```

Tracé des intervalles de confiance et de prédiction



Deuxième partie II

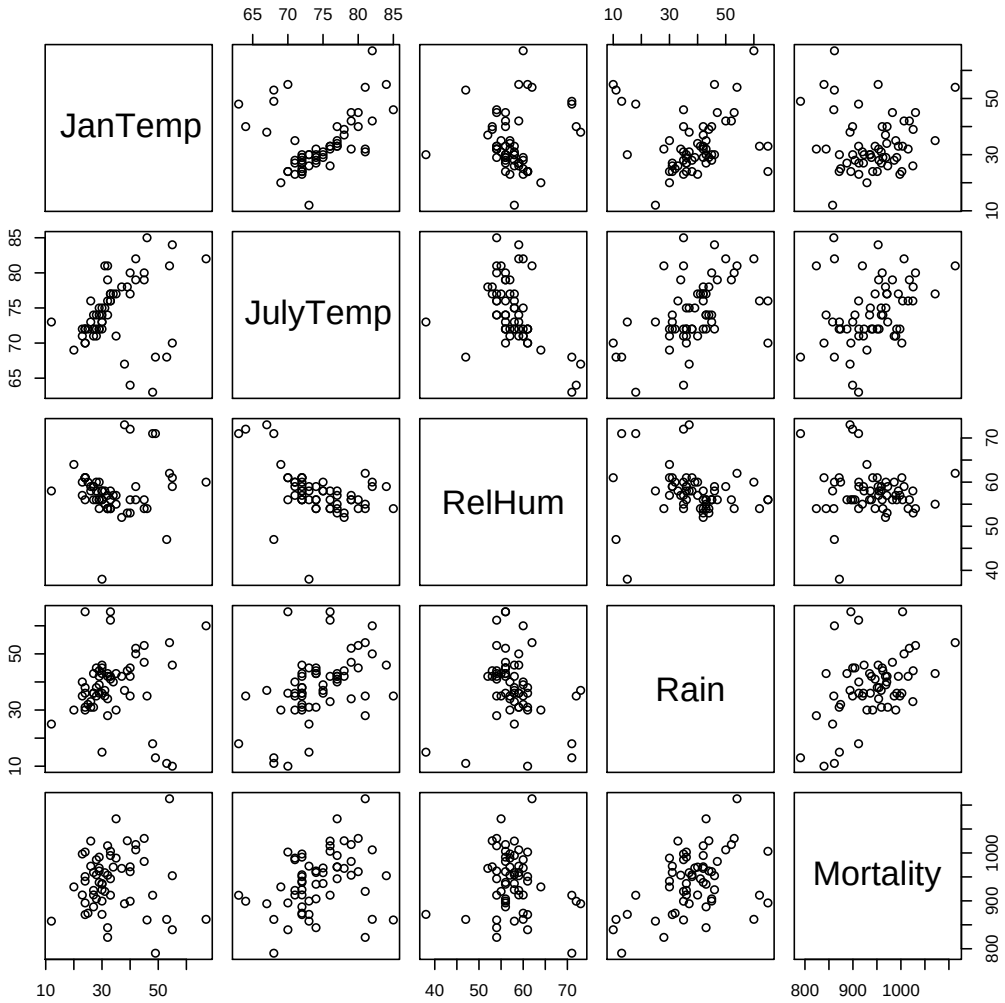
Régression linéaire multiple

Plan

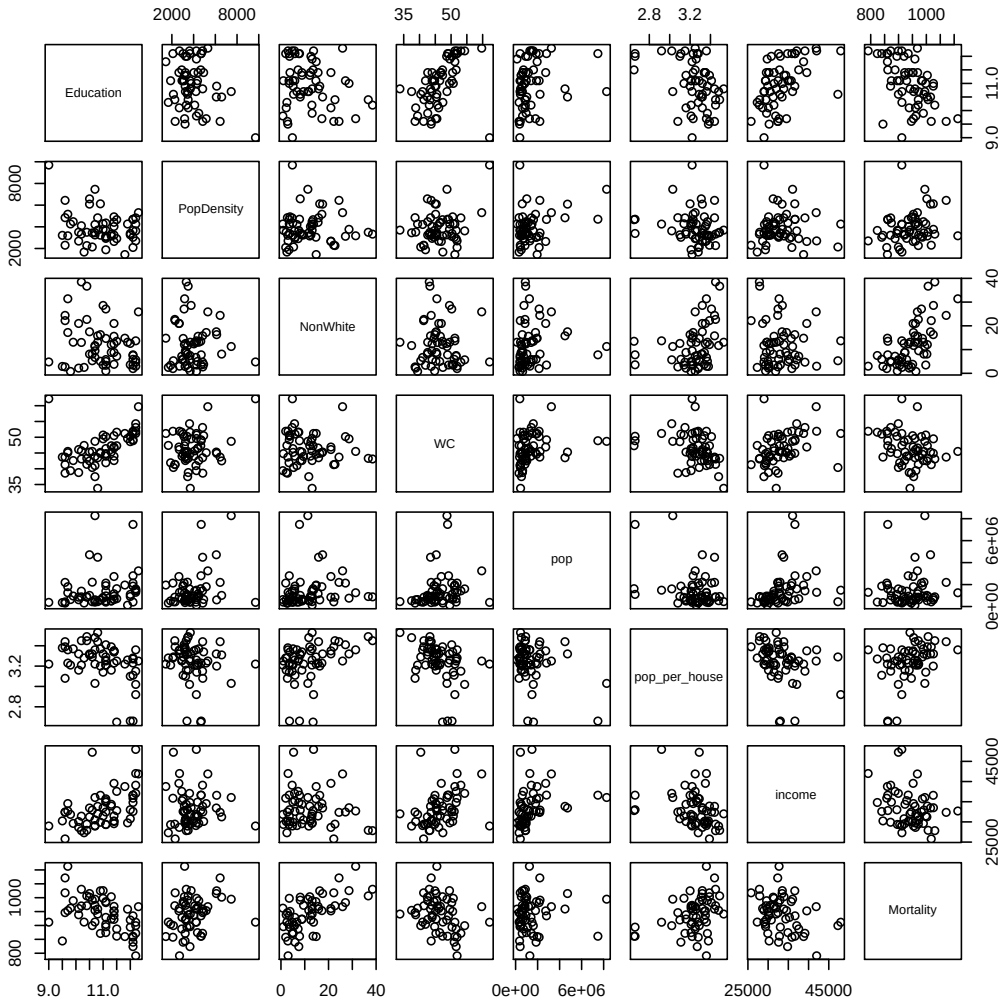
Données SMSA

- Données climatiques, sociologiques et d'environnement relatives à 60 métropoles urbaines des Etats-Unis (Standard Metropolitan Statistical Areas, SMSA)
- 15 variables :
  - climatiques : JanTemp, JulyTemp, RelHum, Rain ;
  - sociologiques : Education, PopDensity, pop, %NonWhite, %WC, pop/house, income (revenu médian)
  - pollution : HCPot (HC pollution potential), NOxPot (Nitrous Oxide pollution potential), SO2Pot (Sulfur Dioxide pollution potential)
  - Mortality (Age adjusted mortality)
- But de l'étude : étudier la relation entre la mortalité (variable à expliquer) et les 14 autres variables (variables epxlicatives).

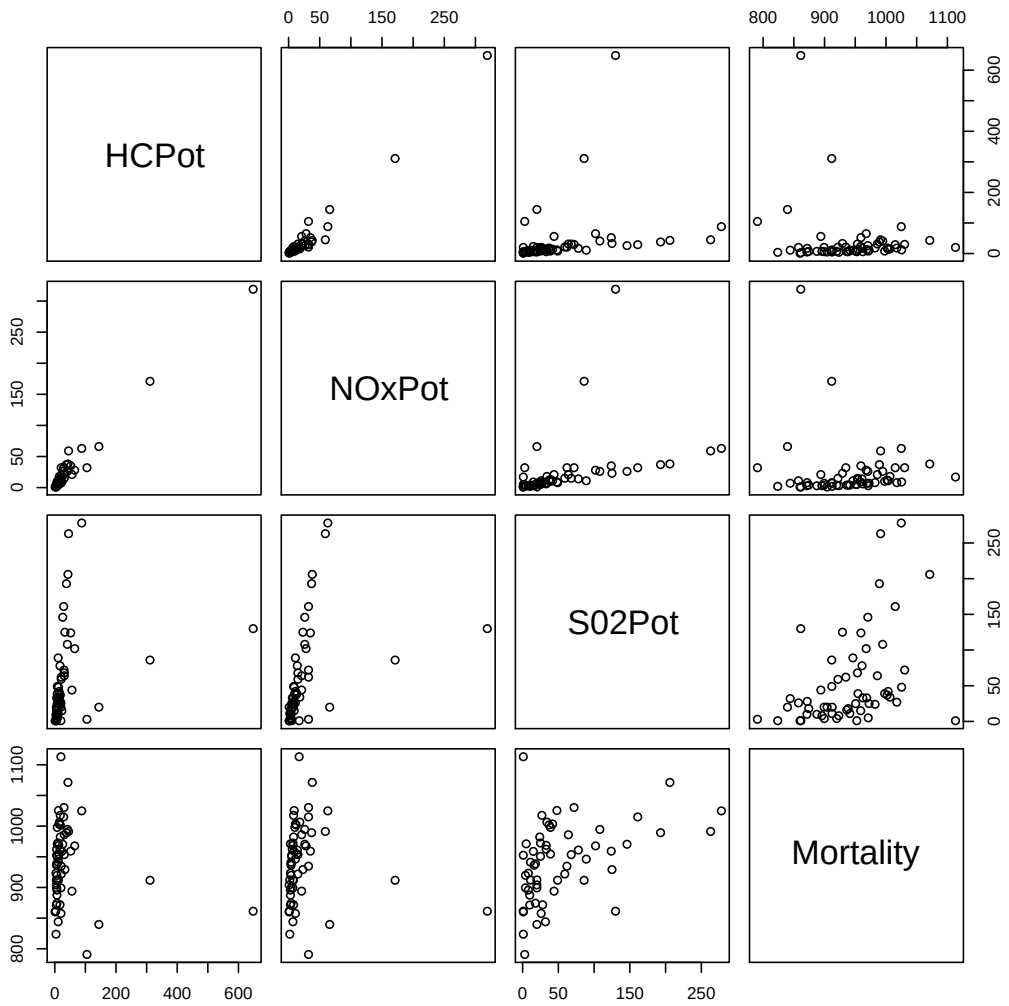
Variables climatiques et mortalité



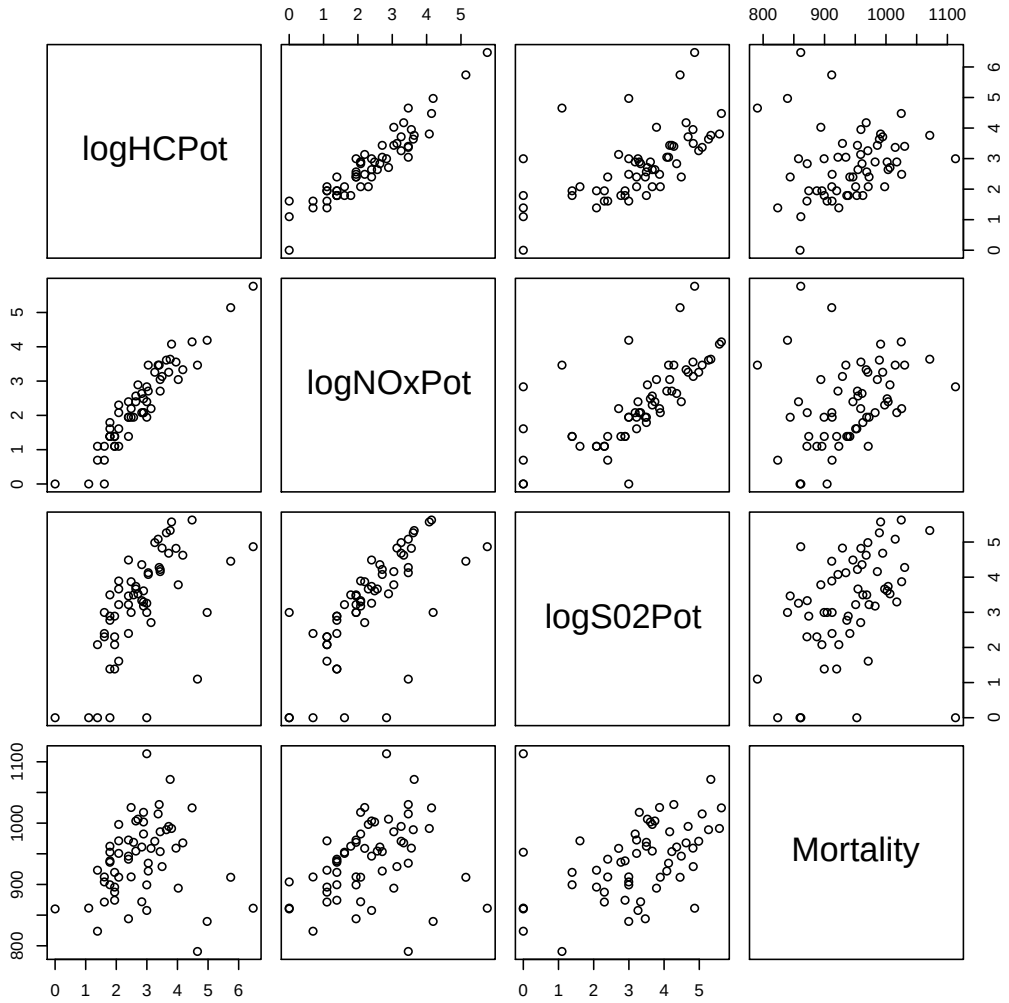
Variables sociologiques et mortalité



Variables de pollution et mortalité



Variables de pollution transformées et mortalité



Le problème

- Il s'agit d'étudier la relation entre une variable aléatoire y (variable dépendante ou à expliquer) et un ensemble de p variables $x_1, \ldots, x_p$ (variables indépendantes, explicatives), dans un but
  - descriptif : quels x_i ont une influence sur y, et comment ?
  - prédictif : prédiction de la variable y, non observée, à partir des x_i supposées connues.
- Pour cela, on dispose d'observations des x_i et de y pour n individus de la population considérée :

$$\begin{array}{ccccc} x_{11} & \cdots & x_{1p} & y_1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} & y_n \end{array}$$

Le modèle

- On suppose que chaque valeur observée y_i sur un individu i est une réalisation d'une v.a.r. Y_i de la forme :

$$Y_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip} + \varepsilon_i$$

avec $\mathbb{E}(\varepsilon_i) = 0$, $\text{var}(\varepsilon_i) = \sigma^2$ et $\text{cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i, j$.

- Matriciellement, on peut écrire

$$Y = Xb + \varepsilon$$

avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix} \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} b = \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix}$$

Plan

Critère des moindres carrés

- Les paramètres b et σ^2 sont inconnus et doivent être estimés à partir des données.
- Le principe de la méthode d'estimation utilisée (méthode des moindres carrés) consiste à minimiser la somme des écarts entre les observations y_i et les prédictions $\widehat{y}_i$ pour chaque observation i :

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

avec $\widehat{y}_i = \widehat{b}_0 + \widehat{b}_1 x_{i1} + \ldots + \widehat{b}_p x_{ip}$.

Solution

- On montre que le vecteur $\widehat{b}$ qui minimise E est :

$$\widehat{b} = (X^t X)^{-1} X^t Y.$$

  C'est un estimateur sans biais de b.
- Les erreurs de prédiction $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$ sont appelés les résidus.
- La statistique

$$\widehat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2$$

  est un estimateur sans biais de σ^2.

Application en R

```
> reg.smsa <- lm(Mortality ~ JanTemp+JulyTemp+RelHum+Rain+Education+PopDensity
+NonWhite+WC+pop+pop_per_house+income+logHCPot+logNOxPot+logSO2Pot)
> reg.smsa

Call: lm(formula = Mortality ~ JanTemp + JulyTemp + RelHum + Rain +
Education + PopDensity + NonWhite + WC + pop + pop_per_house +
income + logHCPot + logNOxPot + logSO2Pot)

Coefficients:
  (Intercept)       JanTemp     JulyTemp       RelHum         Rain    Education
    1.333e+03   -2.305e+00   -1.657e+00   4.067e-01   1.444e+00   -9.458e+00
   PopDensity      NonWhite           WC          pop  pop_per_house      income
    4.509e-03    5.194e+00   -1.852e+00   1.086e-06   -4.595e+01   -5.494e-04
     logHCPot     logNOxPot     logSO2Pot
   -2.322e+01    3.484e+01   -3.002e+00
```

Plan

Coefficient de détermination

- L'équation suivante est appelée équation d'analyse de la variance de la régression :

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

soit

$$\text{variance totale} = \text{variance expliquée} + \text{variance résiduelle}$$

- Cette équation montre que la quantité

$$R^2 = 1 - \frac{\text{variance résiduelle}}{\text{variance totale}},$$

appelée coefficient de détermination, est nécessairement comprise entre 0 et 1.

R^2 et R^2 ajusté

- Dans le meilleur des cas (prévision parfaite), $\widehat{y}_i = y_i$ pour tout i, et $R^2 = 1$.
- Dans le pire des cas, $\widehat{y}_i = \overline{y}$ pour tout i (on ne peut faire mieux que de toujours prédire la valeur moyenne), et $R^2 = 0$.
- Le R^2 peut donc être utilisé pour mesurer la qualité de l'ajustement. Cependant, on constate que la valeur du R^2 augmente artificiellement avec le nombre de variables indépendantes. On définit donc le R^2 ajusté comme :

$$\begin{aligned} \overline{R}^2 &= 1 - \frac{\frac{n}{n-p}\text{variance résiduelle}}{\frac{n}{n-1}\text{variance totale}} \\ &= \frac{n-1}{n-p}R^2 + \frac{p-1}{n-p} \end{aligned}$$

Coefficients de détermination en R

```
> summary(reg.smsa)

Call:
lm(formula = Mortality   JanTemp + JulyTemp + RelHum + Rain +
Education + PopDensity + NonWhite + WC + pop + pop_per_house +
income + logHCPot + logNOxPot + logSO2Pot)

Residuals:
Min 1Q Median 3Q Max
-70.120 -20.669 2.519 23.421 76.385

⋮

Residual standard error: 34.58 on 44 degrees of freedom
Multiple R-squared: 0.7672, Adjusted R-squared: 0.6931
F-statistic: 10.36 on 14 and 44 DF, p-value: 9.864e-10
```

Plan

Hypothèse de normalité des résidus

Si on inclut dans le modèle l'hypothèse supplémentaire $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \forall i$, il est possible de faire différents tests de significativité de la régression :

- Significativité du R^2
- Significativité des coefficients de régression.

Significativité du R^2

- Il s'agit de tester si la relation trouvée entre y et les p variables explicatives est globalement significative. Les hypothèses sont :

$$\begin{aligned} H_0 &: \quad b_1 = b_2 = \ldots = b_p = 0 \\ H_1 &: \quad \exists i, b_i \neq 0 \end{aligned}$$

- On montre que, sous H_0, $F = \frac{R^2}{1-R^2} \frac{n-p-1}{p} \sim F_{p,n-p-1}$, d'où l'on déduit le degré de signification :

$$p = \mathbb{P}_{H_0}(F > f)$$

Significativité des coefficients de régression

- Il s'agit de tester si un coefficient donné est significativement non nul (a une influence sur y) :

$$\begin{aligned} H_0 : &\quad b_j = 0 \\ H_1 : &\quad b_j \neq 0 \end{aligned}$$

- Sous H_0,

$$\frac{\widehat{b}_j}{\widehat{\sigma}\sqrt{v_j}} \sim \mathcal{T}_{n-p-1},$$

v_j étant le terme diagonal (j,j) de la matrice $(X^tX)^{-1}$.

- On en déduit le degré de signification :

$$p = \mathbb{P}_{H_0}(|T| > t).$$

Tests de significativité en R

```
> summary(reg.smsa)

⋮
Coefficients:
               Estimate     Std. Error   t value   Pr(>|t|)
  (Intercept)  1.333e+03    2.917e+02    4.569     3.94e-05   ***
  JanTemp      -2.305e+00   8.795e-01    -2.621    0.0120     *
  JulyTemp     -1.657e+00   2.051e+00    -0.808    0.4236
  RelHum       4.067e-01    1.070e+00    0.380     0.7058
  Rain         1.444e+00    5.847e-01    2.469     0.0175     *
  Education    -9.458e+00   9.080e+00    -1.042    0.3033
  PopDensity   4.509e-03    4.311e-03    1.046     0.3014
  NonWhite     5.194e+00    1.005e+00    5.167     5.55e-06   ***

⋮
Residual standard error: 34.58 on 44 degrees of freedom
Multiple R-squared: 0.7672, Adjusted R-squared: 0.6931
F-statistic: 10.36 on 14 and 44 DF, p-value: 9.864e-10
```

Plan

Principe

- C'est une étape fondamentale permettant de s'assurer de la validité des hypothèses sur lesquels se fondent les résultats précédents.
- Elle comporte 2 aspects :
  - l'analyse des résidus et
  - l'étude de la stabilité des coefficients.

Analyse des résidus

- L'étude des résidus $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$ est fondamentale. Elle permet :
  - de repérer des observations éventuellement aberrantes, ou jouant un rôle important dans la détermination de la régression ;
  - de vérifier empiriquement le bien-fondé des hypothèses du modèle (linéarité, homoscedasticité, normalité des perturbations).
- Il est intéressant de croiser les résidus avec tous les éléments qui peuvent avoir une influence (les x_i, y, etc.), afin de s'assurer de l'absence de toute structure (les résidus doivent être purement aléatoires).
- On définit différents types de résidus : bruts ($\widehat{\varepsilon}_i$), standardisés, prédits, studentisés.

Résidus standardisés

- On montre que $\widehat{\varepsilon}_i$ suit une loi normale d'espérance nulle et de variance
$$\text{var}(\widehat{\varepsilon}_i) = (1 - h_i)\sigma^2,$$
où h_i est le terme diagonal (i, i) de $H = X(X^tX)^{-1}X^t$ (*hat matrix*).
- On peut donc estimer la variance du résidu $\widehat{\varepsilon}_i$ par la quantité $(1 - h_i)\widehat{\sigma}^2$, et on définit les résidus standardisés par
$$\widehat{\varepsilon}'_i = \frac{\widehat{\varepsilon}_i}{\widehat{\sigma}\sqrt{1 - h_i}}$$
- Si l'hypothèse de normalité des perturbations est vérifiée, les $\widehat{\varepsilon}'_i$ doivent rester généralement compris entre -2 et +2.
- Remarque : si h_i est grand, une modification de y_i a une grande influence sur l'hyperplan des moindres carrés. La quantité h_i est appelée *leverage* (effet de levier) de l'observation i.

Résidus studentisés

- Une valeur aberrante ne se traduit pas nécessairement un résidu important, car une telle valeur peut exercer une forte influence sur la régression. Il est donc nécessaire d'étudier l'influence de chaque observation sur sa propre prédiction.
- On définit les résidus prédits par les quantités $\widehat{\varepsilon}_{(-i)} = y_i - \widehat{y}_{(-i)}$, où $\widehat{y}_{(-i)}$ est la prédiction obtenue avec l'échantillon de $n-1$ observations excluant l'observation i, et les résidus studentisés par

$$\widehat{\varepsilon}_i^* = \frac{\widehat{\varepsilon}_{(-i)}}{\sqrt{\text{var}(\widehat{\varepsilon}_{(-i)})}},$$

  en remplaçant $\widehat{\sigma}$ par $\widehat{\sigma}_{(-i)}$.
- La quantité PRESS $= \sum_{i=1}^{n} \widehat{\varepsilon}_{(-i)}^2$ peut être utilisée pour mesurer le pouvoir prédictif du modèle.

Distance de Cook

- On peut également étudier l'influence d'une observation sur les estimations $\widehat{b}_j$ des coefficients de régression, en définissant une distance entre $\widehat{b}$ et $\widehat{b}_{(-i)}$, par exemple la distance de Cook :

$$D_i = \frac{(\widehat{b} - \widehat{b}_{(-i)})^t X^t X (\widehat{b} - \widehat{b}_{(-i)})}{(p+1)\widehat{\sigma}^2}$$

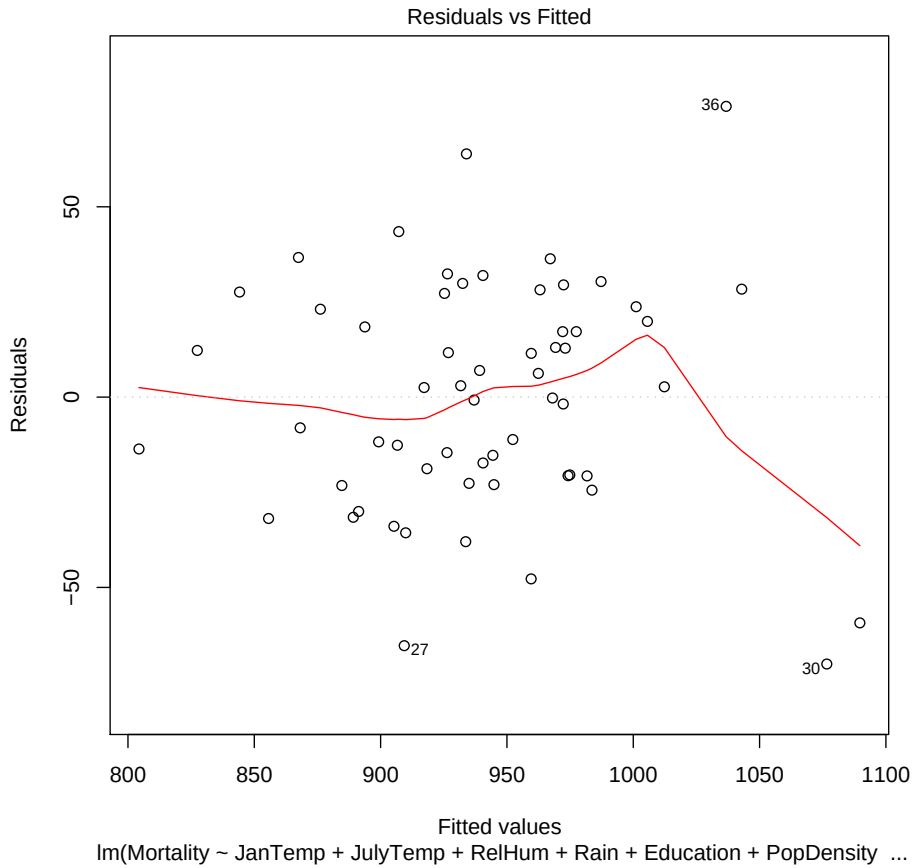
- Une distance de Cook supérieure à 1 indique en général une influence anormale.

Diagnostic de la régression en R

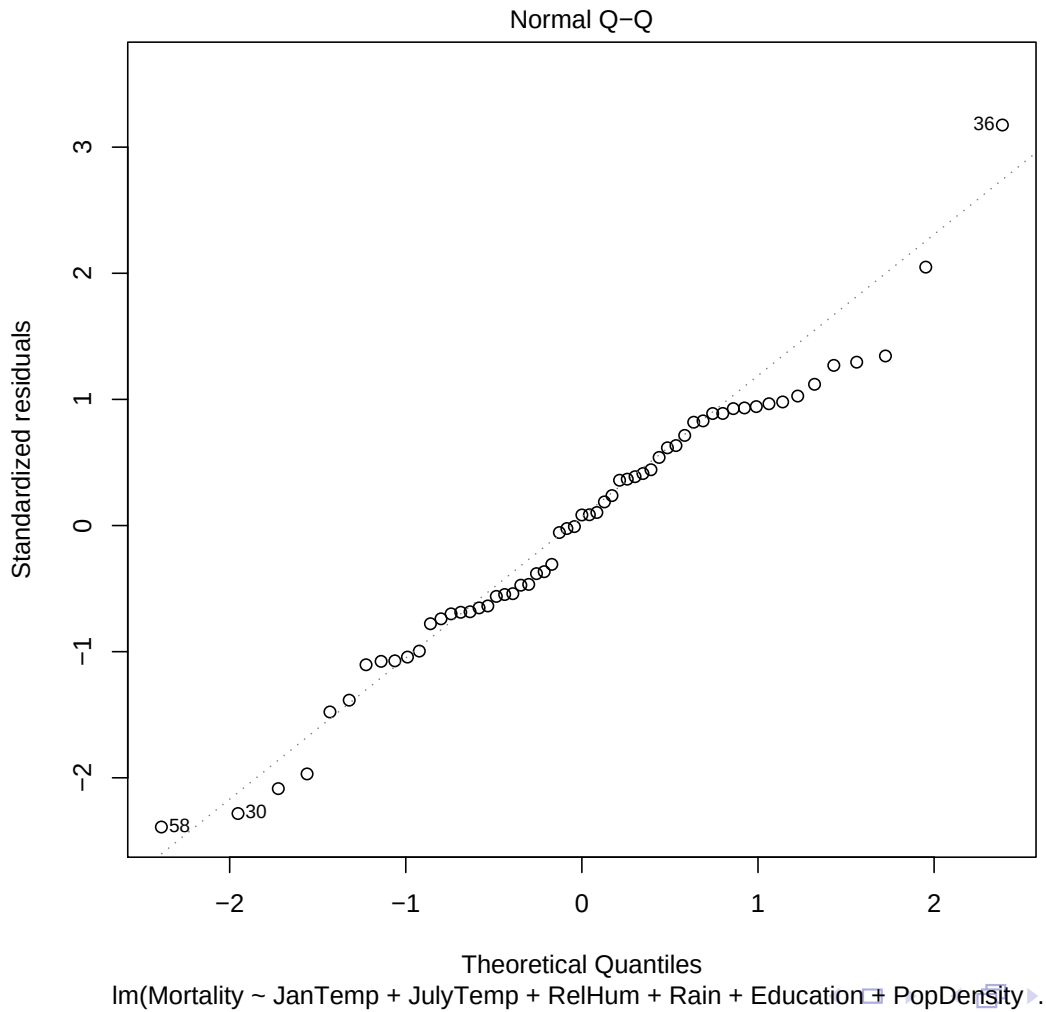
- Prédictions $\widehat{y}_i$: `fitted(reg.smsa)`
- Résidus bruts $\widehat{\varepsilon}_i$: `resid(reg.smsa)`
- Résidus standardisés $\widehat{\varepsilon}'_i$: `rstandard(reg.smsa)`
- Résidus studentisés : $\widehat{\varepsilon}^*_i$: `rstudent(reg.smsa)`
- Distances de Cook D_i : `cooks.distance(reg.smsa)`
- Leverage h_i : `hatvalues(reg.smsa)`

Application aux données SMSA (1)

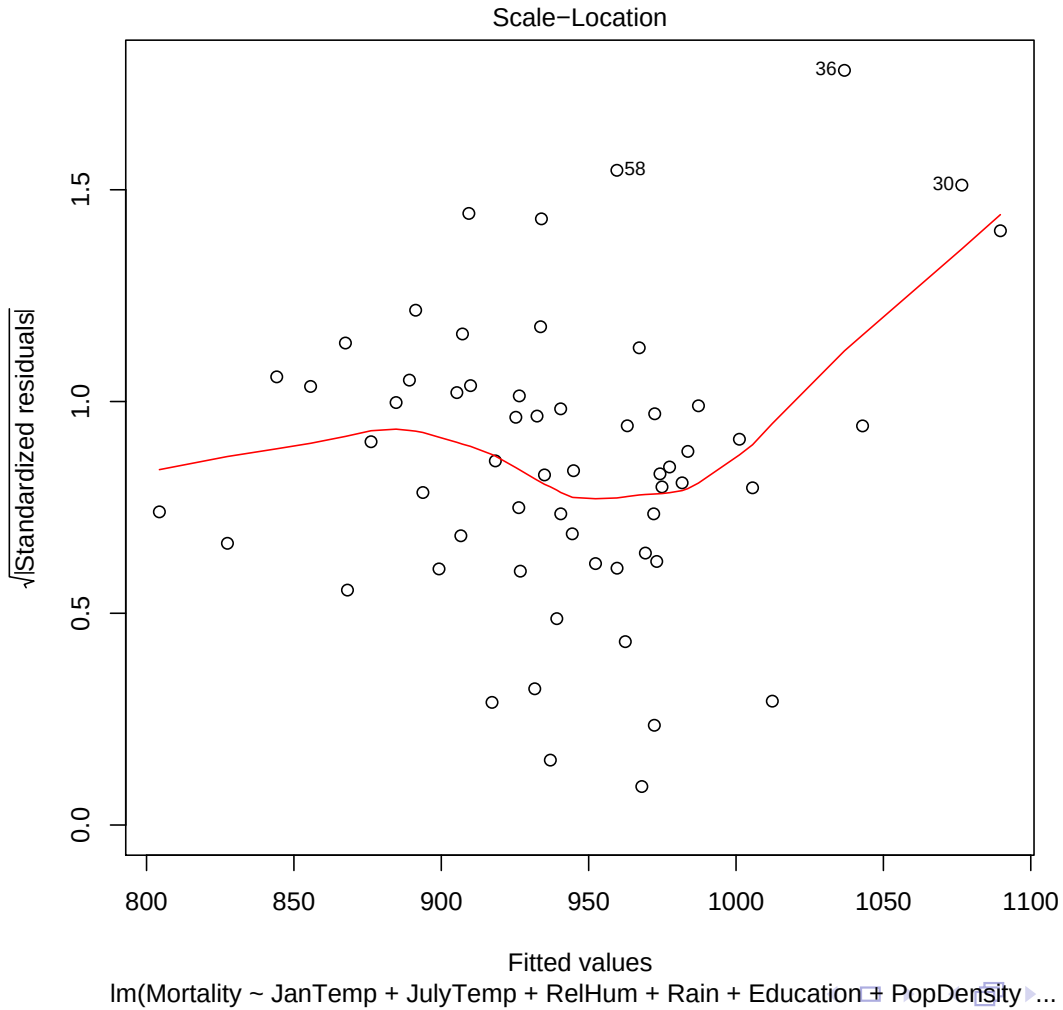
```
> plot(reg.smsa)
```



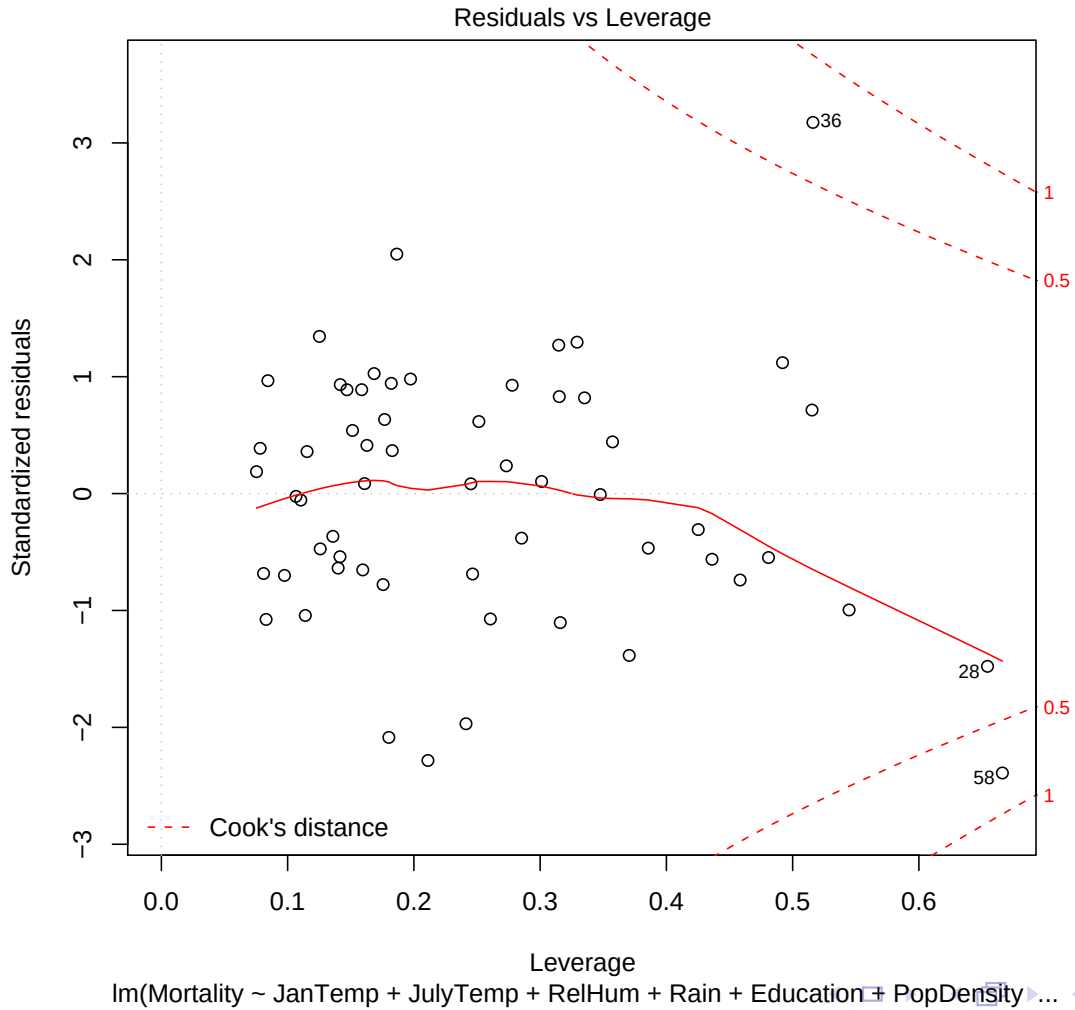
Application aux données SMSA (2)



Application aux données SMSA (3)



Application aux données SMSA (4)



Principe

- Soit $x_0 = (1, x_{10}, \ldots, x_{p0})^t$ le vecteur des variables explicatives pour un nouvel individu, et Y_0 la valeur (inconnue) correspondante de la variable à expliquer. On peut prédire Y_0, et estimer ponctuellement $\mathbb{E}(Y_0|X = x_0)$ par

$$\widehat{y}_0 = \widehat{b}_0 + \widehat{b}_1 x_{01} + \ldots + \widehat{b}_p x_{0p}.$$

- On montre que, si les hypothèses du modèle sont bien vérifiées :

$$\frac{Y_0 - \widehat{y}_0}{\widehat{\sigma}\sqrt{1 + x_0^t (X^t X)^{-1} x_0}} \sim \mathcal{T}_{n-p-1}$$

et

$$\frac{\mathbb{E}(Y_0|x_0) - \widehat{y}_0}{\widehat{\sigma}\sqrt{x_0^t (X^t X)^{-1} x_0}} \sim \mathcal{T}_{n-p-1}$$

Intervalles de prévision et de confiance

On en déduit :

- l'intervalle de prévision sur Y_0 :

$$\widehat{y}_0 \pm t_{n-p-1;1-\frac{\alpha}{2}} \sqrt{1 + x_0^t (X^t X)^{-1} x_0}$$

- l'intervalle de confiance sur $\mathbb{E}(Y_0|x_0)$:

$$\widehat{y}_0 \pm t_{n-p-1;1-\frac{\alpha}{2}} \sqrt{x_0^t (X^t X)^{-1} x_0}$$

Exemple en R

```
> x0 <- data.frame(JanTemp=27,JulyTemp=71,RelHum=59,Rain=36,Education=11.4,
+ PopDensity=3243,NonWhite=8.8,WC=42.6,pop=660328,pop_per_house=3.34,
+ income=29560,logHCPot=log(21),logNOxPot=log(15),logSO2Pot=log(59))
> predict(reg.smsa,int="c",newdata=x0)
         fit        lwr        upr
 [1,]  944.865   923.1046   966.6255
> predict(reg.smsa,int="p",newdata=x0)
         fit        lwr        upr
 [1,]  944.865   871.8582   1017.872
>
> x1 <- transform(x0,logHCPot=log(21/2),logNOxPot=log(15/2),logSO2Pot=log(59/2))

> predict(reg.smsa,int="c",newdata=x1)
         fit        lwr        upr
 [1,]  938.8903   917.0814   960.6992
> predict(reg.smsa,int="p",newdata=x1)
         fit        lwr        upr
 [1,]  938.8903   865.869   1011.912
```