Modèle linéaire

IUT de Vannes

STID 2

2021 – 2022

Responsable du cours :
Anne Cuzol - anne.cuzol@univ-ubs.fr

Responsables des TD :
Anne Cuzol & Audrey Poterie - audrey.poterie@univ-ubs.fr

Espace moodle :
/IUT Vannes / Statistique et Informatique Décisionnelle / STID2 / Modèle linéaire AC

1. Régression linéaire simple

1.1. Introduction

Le but de la régression linéaire simple est d'expliquer/prédire une variable quantitative à partir d'une autre variable quantitative.

Exemples :

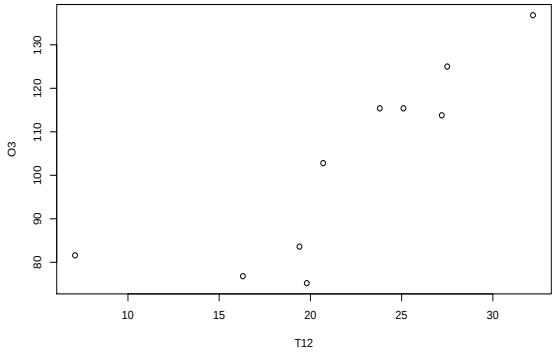
- Expliquer le prix d'un appartement en fonction de sa superficie
- Prédire la hauteur d'un arbre par rapport à sa circonférence
- Prédire les variations des ventes d'un produit par rapport à l'investissement publicitaire
- Expliquer/prédire un taux de pollution par rapport à une variable météorologique

Exemple de la pollution : on s'intéresse à la concentration d'ozone (O3) dans l'air. On souhaite expliquer le taux maximum d'ozone de la journée par la température à 12h (T12).

Extrait des données :

Température à 12h	23,8	16,3	27,2	7,1	25,1	27,5	19,4	...
O3 max	115,4	76,8	113,8	81,6	115,4	125	83,6	...

Observation graphique de la relation entre O3 et T12 :



Les points sont alignés $\Rightarrow$ on peut supposer l'existence d'une relation linéaire entre O3 et T12.

Calcul du coefficient de corrélation empirique entre deux variables x et y :

$$\rho(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$\Rightarrow$ on trouve $\rho(\text{O3}, \text{T12}) \approx 0,84$

Donc :

- La corrélation est positive et proche de 1 : plus la température est élevée, plus la pollution est élevée ;
- Le lien linéaire entre les deux variables est confirmé.

L'ajustement d'un modèle de régression simple va permettre de :

- comprendre/expliquer comment la température influe sur l'ozone ;
- prédire des taux de pollution pour de nouvelles valeurs de températures.

1.2. Ecriture du modèle

Si on suppose une relation linéaire entre x et y, on pose le modèle suivant :

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad \forall i = 1, \ldots, n$$

où :

- les y_i sont des observations de la variable à expliquer ;
- les x_i sont des observations de la variable explicative ;
- les e_i sont les erreurs ;
- β_0 (ordonnée à l'origine) et β_1 (pente) sont des paramètres inconnus du modèle.

On fait des hypothèses sur ce modèle. Ces hypothèses portent sur les erreurs aléatoires :

- Les erreurs sont centrées : $\mathbb{E}(e_i) = 0 \quad \forall i$
- Les erreurs sont toutes de même variance σ^2 (autre paramètre inconnu du modèle) : $\mathbb{V}(e_i) = \sigma^2 \quad \forall i$
- Les erreurs sont non corrélées entre elles : $Cov(e_i, e_j) = 0$ si $i \neq j$

Ecriture vectorielle du modèle (sera utile en régression multiple) :

$$Y = \beta_0 \mathbb{1} + \beta_1 X + e$$

avec :

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbb{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad e = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

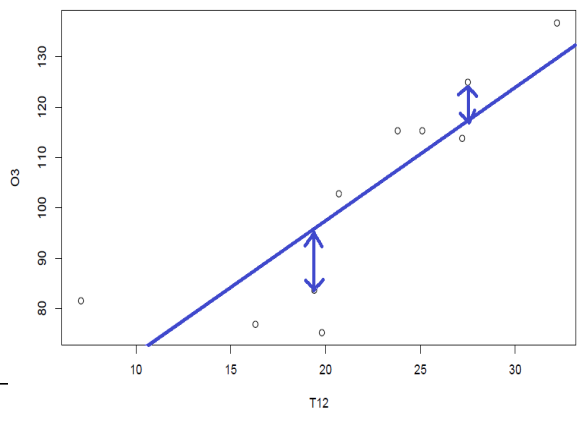
et $\beta_0, \beta_1 \in \mathbb{R}$

1.3. Estimation des paramètres β_0 et β_1

Les paramètres β_0 et β_1 sont estimés par la méthode des moindres carrés.

Les estimateurs des MCO (moindres carrés ordinaires) $\hat{\beta}_0$ et $\hat{\beta}_1$ minimisent :

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



Interprétation graphique : on cherche à minimiser la somme des carrés des distances verticales entre les points (x_i, y_i) du nuage et la droite ajustée d'équation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Calcul de $\hat{\beta}_0$ et $\hat{\beta}_1$:
$S(\beta_0, \beta_1)$ est une fonction quadratique de 2 variables, strictement convexe donc le minimum unique $\hat{\beta}_0, \hat{\beta}_1$ est obtenu en annulant les dérivées partielles de S :

$$\frac{\delta S}{\delta \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (1)$$

$$\frac{\delta S}{\delta \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (2)$$

On a alors :

$$(1) \quad \Rightarrow \quad n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

donc :

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

avec $\bar{x}$ et $\bar{y}$ moyennes empiriques des $(x_i)_{i=1,\ldots,n}$ et $(y_i)_{i=1,\ldots,n}$.

De plus :

$$(2) \quad \Rightarrow \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\iff \quad (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\iff \quad \hat{\beta}_1 (\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i) = \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i$$

$$\iff \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{y}}{\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \bar{x}}$$

$$\iff \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Finalement :

- Estimateur de la pente :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- Estimateur de l'ordonnée à l'origine :

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Rappels :

- Variance estimée de la variable x :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

(idem pour $y \Rightarrow s_y^2$)

- Covariance estimée entre x et y :

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Coefficient de corrélation entre x et y :

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y}$$

Donc :

- L'estimateur de la pente s'écrit encore de la façon suivante :

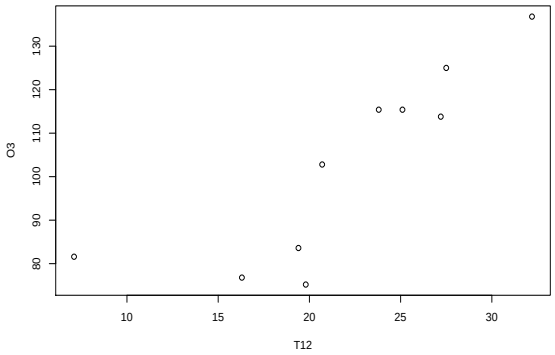
$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

- On peut aussi écrire :

$$\hat{\beta}_1 = \frac{\rho_{xy} s_x s_y}{s_x^2} = \rho_{xy}\frac{s_y}{s_x}$$

$\Rightarrow$ On voit que la pente estimée est nulle si $\rho_{xy} = 0$.

Application



On suppose une relation linéaire entre O3 et T12.
On pose donc le modèle :

$$O3_i = \beta_0 + \beta_1 T12_i + e_i \quad i = 1, \ldots, n$$

Ajustement du modèle avec R :

```
> reg <- lm(O3~T12)
> summary(reg)
```

On obtient :

```
Call:
lm(formula = O3 ~ T12)

Residuals:
    Min      1Q  Median      3Q     Max
-21.890  -9.001   3.856   7.514  17.919

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.0044    13.8050   3.260   0.0115 *
T12           2.6306     0.6029   4.363   0.0024 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.67 on 8 degrees of freedom
Multiple R-squared: 0.7041,     Adjusted R-squared: 0.6671
F-statistic: 19.03 on 1 and 8 DF,  p-value: 0.002403
```

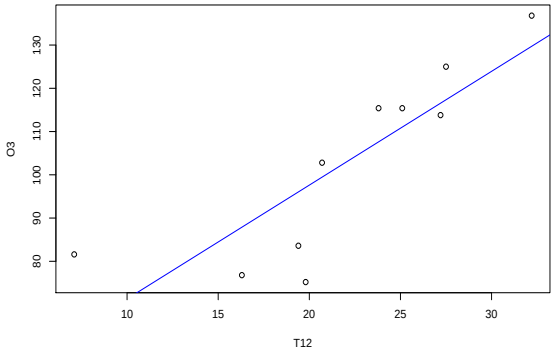
On a donc le modèle ajusté :

$$\hat{\text{O3}} = \hat{\beta}_0 + \hat{\beta}_1 \text{T12}$$

avec les paramètres estimés :

$$\begin{aligned} \hat{\beta}_0 &\approx 45 \\ \hat{\beta}_1 &\approx 2,6 \end{aligned}$$

Représentation graphique du modèle ajusté :



1.4. Propriétés des estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$

☐ Biais des estimateurs :

Rappel : On appelle biais d'un estimateur $\hat{\theta}$ de θ la quantité
$\text{biais}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$.
L'estimateur $\hat{\theta}$ est sans biais quand $\text{biais}(\hat{\theta}) = 0$ donc quand $\mathbb{E}(\hat{\theta}) = \theta$.

Propriété : $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs sans biais de β_0 et β_1.

☐ Variances des estimateurs :

$$V(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}$$
$$V(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Remarque : c'est une variance théorique, on pourra l'estimer quand on aura estimé σ^2.

☐ Théorème de Gauss-Markov :

Parmi les estimateurs sans biais linéaires en Y (c'est le cas ici), les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ sont de variance minimale.

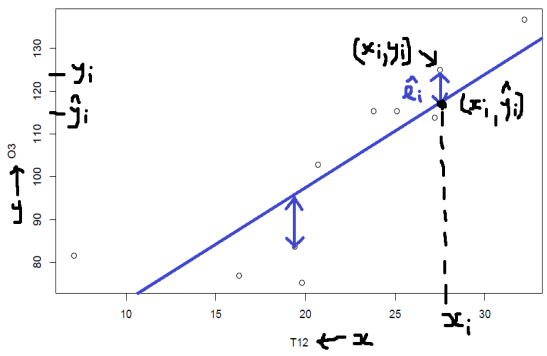
1.5. Calcul des résidus et estimation de σ^2

On a montré comment estimer β_0 et β_1. Il reste un paramètre du modèle à estimer : σ^2 (la variance des erreurs e_i).

Pour cela on utilise les résidus $\hat{e}_i$, estimateurs des erreurs inconnues e_i :

$$\hat{e}_i = y_i - \hat{y}_i$$

où les $\hat{y}_i$ sont les valeurs ajustées par le modèle : $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.



Estimation de la variance des erreurs :

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2$$

Propriété : $\hat{\sigma}^2$ est un estimateur sans biais de σ^2.

Remarque : on utilise l'estimation $\hat{\sigma}^2$ pour estimer les variances des estimateurs MCO

$$\begin{aligned}\hat{V}(\hat{\beta}_0) &= \frac{\hat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}\\ \hat{V}(\hat{\beta}_1) &= \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\end{aligned}$$

On notera dans la suite ces deux variances estimées $\hat{\sigma}^2_{\hat{\beta}_0}$ et $\hat{\sigma}^2_{\hat{\beta}_1}$.

Application

On reprend la sortie du modèle de régression ajusté :

```
Call:
lm(formula = O3 ~ T12)

Residuals:
    Min      1Q  Median      3Q     Max
-21.890  -9.001   3.856   7.514  17.919

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.0044    13.8050   3.260   0.0115 *
T12           2.6306     0.6029   4.363   0.0024 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.67 on 8 degrees of freedom
Multiple R-squared: 0.7041,     Adjusted R-squared: 0.6671
F-statistic: 19.03 on 1 and 8 DF,  p-value: 0.002403
```

On lit :

$$\begin{aligned}\hat{\sigma}^2 &\approx 12,7^2\\ \hat{\sigma}^2_{\hat{\beta}_0} &\approx 13,8^2\\ \hat{\sigma}^2_{\hat{\beta}_1} &\approx 0,6^2\end{aligned}$$

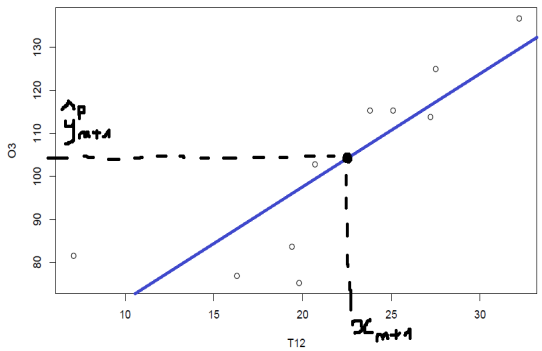
1.6. Calcul de prévisions

Jusqu'à présent on travaillait avec les valeurs $x_1, \ldots, x_n$ de la variable explicative x.
Soit x_{n+1} une nouvelle valeur observée. On veut prédire la valeur y_{n+1}.

D'après le modèle : $y_{n+1} = \beta_0 + \beta_1 x_{n+1} + e_{n+1}$

$\Rightarrow$ On prédit grâce au modèle estimé (équation de la droite ajustée) :

$$\hat{y}^P_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$$



L'erreur de prévision est définie par :

$$\hat{e}^P_{n+1} = y_{n+1} - \hat{y}^P_{n+1}.$$

Elle satisfait :

- $\mathbb{E}(\hat{e}^P_{n+1}) = 0$
- $V(\hat{e}^P_{n+1}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$

Donc :

- En moyenne, l'erreur de prévision est nulle.
- La variance augmente quand x_{n+1} s'éloigne du centre de gravité du nuage : plus une observation x_{n+1} est éloignée de la moyenne, moins on a d'information sur elle.

Exemple : prédire le taux de pollution pour une journée avec une température à 12h de 22 degrés.

On a vu que $\hat{O3} = 45 + 2,6T12$

Donc $\hat{O3}^P = 45 + 2,6 * 22 = 102,2$.

Remarque : on peut calculer les prévisions (et les variances d'erreurs associées) avec la fonction "predict" de R.

1.7. Qualité du modèle

Un "bon" modèle aura des estimations $\hat{y}_i$ proches des vraies valeurs y_i. Un indicateur de cette qualité est le coefficient de détermination R^2 :

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= \frac{SCE}{SCT} \end{aligned}$$

avec :

- SCE = Somme des carrés expliquée par le modèle
- SCT = Somme des carrés totale

$\Rightarrow$ Part de la variabilité expliquée par le modèle / variabilité totale

Le R^2 varie entre 0 ("pas bon") et 1 ("bon").

Remarque (dans le cas de la régression simple uniquement) :

$$\begin{aligned} R^2 &= \rho_{xy}^2 \\ &= \text{carré du coefficient de corrélation empirique} \end{aligned}$$

Preuve en exercice :

- remplacer d'abord $\hat{y}_i$ par $\hat{\beta}_0 + \hat{\beta}_1 x_i$
- puis $\hat{\beta}_0$ par $\bar{y} - \hat{\beta}_1 \bar{x}$

Application

On reprend la sortie du modèle de régression ajusté :

```
Call:
lm(formula = O3 ~ T12)

Residuals:
    Min      1Q  Median      3Q     Max
-21.890  -9.001   3.856   7.514  17.919

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.0044    13.8050   3.260   0.0115 *
T12           2.6306     0.6029   4.363   0.0024 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.67 on 8 degrees of freedom
Multiple R-squared: 0.7041,     Adjusted R-squared: 0.6671
F-statistic: 19.03 on 1 and 8 DF,  p-value: 0.002403
```

On lit : $R^2 \approx 0,7$

Rappel : on avait vu que $\rho(\mathsf{O3}, \mathsf{T12}) = 0,84$ donc on a bien $R^2 = \rho^2$.

1.8. Intervalles de confiance pour les paramètres β_0 et β_1

On veut connaitre la loi des estimateurs pour calculer des intervalles de confiance et faire des tests. Pour cela il faut ajouter une hypothèse au modèle : l'hypothèse gaussienne sur les erreurs.

On avait :

- $\mathbb{E}(e_i) = 0$ et $V(e_i) = \sigma^2 \quad \forall i$
- e_i non corrélées

On suppose maintenant :

- $e_i \sim \mathcal{N}(0, \sigma^2) \quad \forall i$
- e_i indépendants

On a donc le modèle :

$$y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2) \quad \forall i$$

Loi des estimateurs MCO (sous l'hypothèse gaussienne) :

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim \mathcal{T}(n-2)$$
$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim \mathcal{T}(n-2)$$

où $\mathcal{T}(n-2)$ est une loi de Student à $n-2$ degrés de liberté.

On peut en déduire des intervalles de confiance pour β_0 et β_1 :

$$IC(\beta_0) : [\hat{\beta}_0 - t_{1-\frac{\alpha}{2}}^{n-2}\hat{\sigma}_{\hat{\beta}_0}, \hat{\beta}_0 + t_{1-\frac{\alpha}{2}}^{n-2}\hat{\sigma}_{\hat{\beta}_0}]$$

$$IC(\beta_1) : [\hat{\beta}_1 - t_{1-\frac{\alpha}{2}}^{n-2}\hat{\sigma}_{\hat{\beta}_1}, \hat{\beta}_1 + t_{1-\frac{\alpha}{2}}^{n-2}\hat{\sigma}_{\hat{\beta}_1}]$$

où $t_{1-\frac{\alpha}{2}}^{n-2}$ est le quantile de niveau $1-\frac{\alpha}{2}$ d'une loi de Student $\mathcal{T}(n-2)$.

1.9. Test de Student de significativité de β_1 et β_0

☐ Significativité du paramètre de pente :

Rappel : le modèle s'écrit $y_i = \beta_0 + \beta_1 x_i + e_i \quad \forall i = 1, \ldots, n.$

On veut tester si la valeur de y est influencée par la variable x (donc si le modèle a un intérêt).

On traduit cela sous la forme d'un test d'hypothèses sur le paramètre de pente β_1 :

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

La statistique de test est $T = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$

On connait sa loi sous H_0 : $\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim \mathcal{T}(n-2)$

$\Rightarrow$ On rejette H_0 au niveau α si l'observation de T est telle que $|T| > t_{n-2}^{1-\frac{\alpha}{2}}$.

Remarque : en pratique on regarde la p-value (probabilité pour la statistique de test sous H_0 de dépasser la valeur estimée).

$\Rightarrow$ On décide de rejeter H_0 si la statistique de test correspond à une p-value très faible.

☐ Significativité de l'ordonnée à l'origine :

$$\begin{cases} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{cases}$$

La statistique de test est $T = \frac{\hat{\beta}_0}{\hat{\sigma}_{\hat{\beta}_0}}$

On connait sa loi sous H_0 : $\frac{\hat{\beta}_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim \mathcal{T}(n-2)$

La démarche du test est ensuite la même que pour β_1.

Remarques :

- Le test le plus intéressant est celui sur β_1. Tester la nullité de β_0 est seulement utile pour aboutir à un modèle plus compact (privé de β_0).
- On peut conclure sur la significativité de β_0 et β_1 à partir des intervalles de confiance : si l'intervalle de confiance contient zéro, on conclut que le paramètre n'est pas significativement différent de zéro (au niveau considéré).

Application

On reprend la sortie du modèle de régression ajusté :

```
Call:
lm(formula = O3 ~ T12)

Residuals:
    Min      1Q  Median      3Q     Max
-21.890  -9.001   3.856   7.514  17.919

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.0044    13.8050   3.260   0.0115 *
T12           2.6306     0.6029   4.363   0.0024 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.67 on 8 degrees of freedom
Multiple R-squared: 0.7041,     Adjusted R-squared: 0.6671
F-statistic: 19.03 on 1 and 8 DF,  p-value: 0.002403
```

Conclusions au niveau $\alpha = 5\%$:

Les p-values (lues dans la colonne "Pr($>|t|$") des deux tests de Student sont inférieures à 5% donc :

- la pente est significativement différente de zéro $\Rightarrow$ la température a bien une influence sur le niveau de pollution ;
- la constante du modèle est significativement différente de zéro.

De plus, la pente estimée est positive donc plus il fait chaud, plus le niveau de pollution est élevé.

2. Régression linéaire multiple

2.1. Introduction

Expliquer une variable quantitative par une seule variable quantitative peut être trop simpliste.

Dans notre exemple de pollution atmosphérique, d'autres variables météorologiques que la température peuvent expliquer la concentration en ozone : la nébulosité (indice de couverture nuageuse), la précipitation, le vent, etc.

Exemple de tableau de données avec plusieurs variables quantitatives :

Individu	O3	T12	Vx12	Ne12
1	87	18.5	-1.7101	4
2	82	18.4	-4.0000	5
3	92	17.6	1.8794	5
4	114	19.7	0.3473	1
⋮	⋮	⋮	⋮	⋮

On va donc chercher à :

- ajuster un modèle pour expliquer O3 en fonction de T12, Vx12, Ne12, etc.
- prédire les valeurs d'O3 pour de nouvelles valeurs de T12, Vx12, Ne12, etc.

On s'intéresse donc au lien entre la variable à expliquer y et p variables explicatives notées $x_1, \ldots, x_p$.

2.2. Ecriture du modèle

C'est une généralisation du modèle de régression linéaire simple pour p variables explicatives $x_1, \ldots, x_p$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i \quad \forall i = 1, \ldots, n$$

avec x_{ij} valeur de l'individu i pour la variable explicative x_j.

On a :

- $\beta_0, \ldots, \beta_p$ paramètres inconnus du modèle (non aléatoires)
- e_i erreurs inconnues aléatoires :
  - centrées : $\mathbb{E}(e_i) = 0 \quad \forall i$
  - de même variance σ^2 (paramètre inconnu du modèle)
  - non corrélées entre elles
- y_i aléatoires, $x_{i1}, \ldots, x_{ip}$ considérés non aléatoires, observés sur n individus

Ecriture matricielle du modèle :

$$Y = X\beta + e$$

avec :

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \in \mathbb{R}^n$$

X matrice de taille (n, p), concaténation d'une colonne contenant des "1" (si le modèle contient une constante) et des p variables x_j :

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Dans l'exemple :

- 1ère colonne $= (1, \dots, 1)^T$
- 2ème colonne $= n$ observations de T12
- 3ème colonne $= n$ observations de Vx12
- etc.

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \quad \in \mathbb{R}^{p+1}$$

vecteur des $p + 1$ paramètres inconnus du modèle

et :

$$e = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \quad \in \mathbb{R}^n$$

Finalement :

$$Y_{(n,1)} = X_{(n,p+1)}\beta_{(p+1,1)} + e_{(n,1)}$$

Réécriture des hypothèses sur les erreurs (sous forme matricielle) :

- centrées : $\mathbb{E}(e) = 0$
- de même variance et non corrélées : $\Sigma_e = \sigma^2 \mathbb{I}_n = \begin{pmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{pmatrix}$

Nouvelle hypothèse pour la régression multiple : pas de corrélation linéaire entre les variables explicatives.

2.3. Estimation du vecteur β aux moindres carrés

Le vecteur $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T \quad \in \mathbb{R}^{p+1}$ est l'estimateur des moindres carrés de $\beta = (\beta_0, \ldots, \beta_p)^T$.

$\hat{\beta}$ minimise :

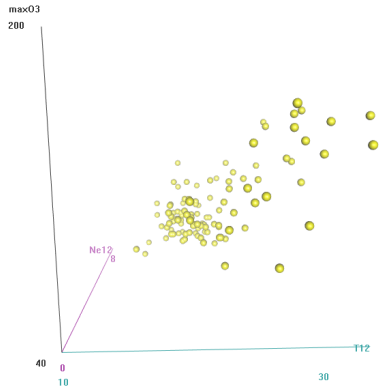
$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_p x_{ip})^2 \\ &= (Y - X\beta)^T (Y - X\beta) \end{aligned}$$

On obtient alors l'expression de l'estimateur des moindres carrés :

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

Application

On veut expliquer le taux maximum d'ozone par deux variables explicatives : la température à 12h (T12) et la nébulosité à 12h (Ne12).



On pose donc le modèle :

$$\mathsf{O3}_i = \beta_0 + \beta_1 \mathsf{T12}_i + \beta_2 \mathsf{Ne12}_i + e_i \quad i = 1, \ldots, n$$

Ajustement du modèle avec R :

```
> reg <- lm(O3~T12+Ne12)
> summary(reg)
```

On obtient :

```
Call:
lm(formula = maxO3 ~ T12 + Ne12, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-41.133 -11.799  -0.029   9.148  45.584

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

Residual standard error: 17.02 on 109 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared: 0.6419,     Adjusted R-squared: 0.6353
F-statistic: 97.69 on 2 and 109 DF,  p-value: < 2.2e-16
```

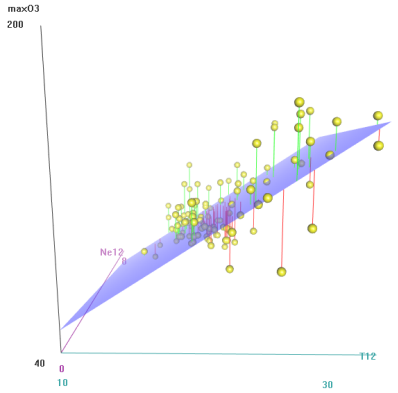
On a donc le modèle ajusté :

$$\hat{O3} = 7.7 + 4.5\text{T12} - 2.7\text{Ne12}$$

avec les paramètres estimés :

$$\begin{aligned} \hat{\beta}_0 &\approx 7.7 \\ \hat{\beta}_1 &\approx 4.5 \\ \hat{\beta}_2 &\approx -2.7 \end{aligned}$$

Représentation graphique du modèle ajusté :



2.4. Propriétés des estimateurs $\hat{\beta}$

☐ Biais des estimateurs :

Propriété : $\hat{\beta}$ est un estimateur sans biais du vecteur de paramètres β.

Preuve :

$$\begin{aligned} \mathbb{E}(\hat{\beta}) &= \mathbb{E}((X^TX)^{-1}X^TY) = (X^TX)^{-1}X'\mathbb{E}(Y) \\ &= (X^TX)^{-1}X^TX\beta \\ &= \beta \end{aligned}$$

☐ Matrice de variance-covariance des estimateurs :

$$V(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Preuve :

$$\begin{aligned} V(\hat{\beta}) &= V((X^T X)^{-1} X^T Y) \\ &= (X^T X)^{-1} X^T V(Y) X (X^T X)^{-1} \end{aligned}$$

or $V(Y) = \sigma^2 \mathbb{I}_n$ car $V(X\beta + e) = V(e) = \sigma^2 \mathbb{I}_n$

donc $V(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$.

☐ Théorème de Gauss-Markov : Parmi les estimateurs sans biais linéaires en Y, $\hat{\beta}$ est celui qui admet la plus petite variance.

2.5. Calcul des résidus et estimation de σ^2

Les résidus $\hat{e}$ sont des estimateurs des erreurs inconnues e :

$$\hat{e} = Y - \hat{Y}$$

où $\hat{Y} = X\hat{\beta}$ est l'ajustement par le modèle.

Estimation de la variance des erreurs :

$$\hat{\sigma}^2 = \frac{||\hat{e}||^2}{n-(p+1)} = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-(p+1)} = \frac{SCR}{n-(p+1)}$$

où SCR est la somme des carrés résiduelle.

Propriété : La statistique $\hat{\sigma}^2$ est un estimateur sans biais de σ^2.

Remarque : on utilise l'estimation $\hat{\sigma}^2$ pour estimer la variance de l'estimateur $\hat{\beta}$:

$$\hat{V}(\hat{\beta}) = \hat{\sigma}^2_{\hat{\beta}} = \hat{\sigma}^2 (X^T X)^{-1} = \frac{SCR}{n-(p+1)} (X^T X)^{-1}$$

Donc, pour chaque coefficient de la régression multiple :

$$\hat{V}(\hat{\beta}_j) = \hat{\sigma}^2_{\hat{\beta}_j} = \hat{\sigma}^2 (X^T X)^{-1}_{jj}$$

Application

```
Call:
lm(formula = maxO3 ~ T12 + Ne12, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-41.133 -11.799  -0.029   9.148  45.584

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

Residual standard error: 17.02 on 109 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared: 0.6419,     Adjusted R-squared: 0.6353
F-statistic: 97.69 on 2 and 109 DF,  p-value: < 2.2e-16
```

On lit :

$$\begin{aligned}
\hat{\sigma}^2 &\approx 17.02^2 \\
\hat{\sigma}^2_{\hat{\beta}_0} &\approx 15.09^2 \\
\hat{\sigma}^2_{\hat{\beta}_1} &\approx 0.53^2 \\
\hat{\sigma}^2_{\hat{\beta}_2} &\approx 0.94^2
\end{aligned}$$

2.6. Calcul d'une prévision

Soit une nouvelle valeur $x_{n+1}^T = (1, x_{n+1,1}, \ldots, x_{n+1,p})$ (n+1ème individu observé pour toutes les variables).
On veut prédire y_{n+1} par $\hat{y}_{n+1}^P$.

D'après le modèle : $y_{n+1} = x_{n+1}^T \beta + e_{n+1}$

$\Rightarrow$ On prédit grâce au modèle ajusté : $\hat{y}_{n+1}^P = x_{n+1}^T \hat{\beta}$

L'erreur de prévision est définie par :

$$\hat{e}^P_{n+1} = y_{n+1} - \hat{y}^P_{n+1}.$$

Elle satisfait :

- $\mathbb{E}(\hat{e}^P_{n+1}) = 0$
- $V(\hat{e}^P_{n+1}) = \sigma^2(1 + x^T_{n+1}(X^T X)^{-1} x_{n+1})$

Exemple : à quel taux de pollution peut-on s'attendre pour une température à 12h de 25 degrés et une nébulosité de 0 à 12h (aucun nuage) ?

On utilise l'équation du modèle ajusté :

$$\hat{\text{O3}} = 7.7 + 4.5\text{T12} - 2.7\text{Ne12}$$

Donc $\hat{\text{O3}}^P = 7.7 + 4.5 * 25 - 2.7 * 0 = 120.2.$

Remarque : comme pour la régression simple, on peut calculer les prévisions (et les variances d'erreurs associées) avec la fonction "predict" de R.

2.7. Qualité du modèle

Le coefficient de détermination multiple R^2 est défini par :

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{SCE}{SCT} \end{aligned}$$

Remarques :

- Pour un modèle sans constante on utilise :

$$R^2 = \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2}$$

- Attention : le R^2 d'un modèle sans constante est toujours plus grand que celui du même modèle avec constante.

- Attention : le R^2 augmente avec le nombre de variables explicatives. On ne peut donc utiliser le R^2 que pour comparer deux modèles ayant le même nombre de variables explicatives

  Sinon on utilise le R^2 ajusté :

$$R_a^2 = 1 - \frac{n-1}{n-(p+1)} \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

  et pour un modèle sans constante :

$$R_a^2 = 1 - \frac{n}{n-(p+1)} \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n y_i^2}$$

Application

On reprend la sortie du modèle de régression ajusté :

```
Call:
lm(formula = maxO3 ~ T12 + Ne12, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-41.133 -11.799  -0.029   9.148  45.584

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

Residual standard error : 17.02 on 109 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared: 0.6419,     Adjusted R-squared: 0.6353
F-statistic: 97.69 on 2 and 109 DF,  p-value: < 2.2e-16
```

On lit : $R^2 \approx 0.64$ et $R_a^2 \approx 0.63$.

Application

On réestime le modèle sans la constante :

```
Call:
lm(formula = maxO3 ~ T12 + Ne12 - 1, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-40.331 -11.798   0.091   9.644  45.375

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
T12    4.7274     0.1371  34.492  < 2e-16 ***
Ne12  -2.3018     0.5449  -4.224 4.97e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.96 on 110 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared : 0.9684,     Adjusted R-squared: 0.9678
F-statistic:  1685 on 2 and 110 DF,  p-value: < 2.2e-16
```

On lit : $R^2 \approx 0.97$ et $R_a^2 \approx 0.97$.

$\Rightarrow$ L'augmentation artificielle du R^2 (et R_a^2) est due au retrait de la constante.

2.8. Intervalles de confiance pour les paramètres

Rappel : le modèle s'écrit $Y = X\beta + e$

avec les hypothèses :

- $\mathbb{E}(e) = 0$ et $V(e) = \Sigma_e = \sigma^2 \mathbb{I}_n$ (erreurs centrées, non corrélées et de même variance)
- Non corrélation des variables explicatives entre elles

Pour calculer des intervalles de confiance et faire des tests, on ajoute une hypothèse gaussienne sur les erreurs :

$$e \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n)$$

On peut aussi écrire :

$$e_1, \ldots, e_n \quad \text{iid et} \quad e_i \sim \mathcal{N}(0, \sigma^2)$$

Loi des estimateurs des moindres carrés $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T$ (sous l'hypothèse gaussienne) :

$$\forall j = 1, \ldots, p \quad T_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim \mathcal{T}(n - (p+1))$$

où $\mathcal{T}(n-(p+1))$ est une loi de Student à $n-(p+1)$ degrés de liberté.

Rappels :

- $\hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma}\sqrt{(X^T X)^{-1}_{jj}}$
- $\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{e}_i^2}{n-(p+1)}$

Intervalle de confiance de niveau $1-\alpha$ pour β_j :

$$IC(\beta_j) : [\hat{\beta}_j - t_{1-\frac{\alpha}{2}}^{n-(p+1)} \hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + t_{1-\frac{\alpha}{2}}^{n-(p+1)} \hat{\sigma}_{\hat{\beta}_j}]$$

où $t_{1-\frac{\alpha}{2}}^{n-(p+1)}$ est le quantile de niveau $1-\frac{\alpha}{2}$ d'une loi de Student $\mathcal{T}(n-(p+1))$.

2.9. Tests de Student de significativité des coefficients

On met en place le test pour chaque coefficient.

Hypothèses de test :

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

Statistique de test : $T = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$

Loi sous H_0 : $\frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim \mathcal{T}(n-(p+1))$

$\Rightarrow$ On rejette H_0 au niveau α si l'observation de T est telle que $|T| > t_{1-\frac{\alpha}{2}}^{n-(p+1)}$.

Application

On reprend la sortie du modèle de régression multiple avec constante :

```
Call:
lm(formula = maxO3 ~ T12 + Ne12, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-41.133 -11.799  -0.029   9.148  45.584

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

Residual standard error : 17.02 on 109 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared: 0.6419,     Adjusted R-squared: 0.6353
F-statistic: 97.69 on 2 and 109 DF,  p-value: < 2.2e-16
```

On regarde la p-value associée à chaque coefficient : le coefficient est significativement différent de 0 au niveau 5% si la p-value est plus petite que 5%.

$\Rightarrow$ Ici on pourrait retirer la constante et relancer l'estimation d'un modèle sans constante.

2.10. Tests de Fisher entre modèles emboîtés

Soit un modèle de régression $Y = X\beta + e$ avec p variables explicatives. On veut tester la validité d'un sous-modèle (modèle "emboîté") avec certains coefficients nuls.

Par exemple : on veut tester la nullité des coefficients $\beta_1, \ldots, \beta_q$ (associés aux q premières variables explicatives).

On pose les hypothèses de test :

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0 \\ H_1 : \exists j \in \{1, \ldots, q\} : \beta_j \neq 0 \end{cases}$$

Idée du test : Si on conclut le non rejet de l'hypothèse H_0 c'est que le sous-modèle (plus simple, avec q coefficients nuls) est préférable au modèle complet.

Démarche du test :
Sous l'hypothèse H_0 on a le sous-modèle suivant :

$$Y = X_0\beta_0 + e_0$$

où X_0 est la matrice X privée de q variables.

Statistique de test :

$$F = \frac{n-(p+1)}{p-p_0}\frac{SCR_0 - SCR}{SCR}$$

avec :

- n : nombre d'observations
- $p+1$: taille du modèle complet (p variables explicatives + constante)
- $p-p_0$: différence de taille entre le modèle complet et le sous-modèle ($= q$, nombre de coefficients dont on teste la nullité)
- SCR : somme des carrés des résidus du modèle complet
- SCR_0 : somme des carrés des résidus du sous-modèle

Loi de la statistique de test sous H_0 : $F \sim \mathcal{F}(p-p_0, n-(p+1))$

$\Rightarrow$ On rejette H_0 au niveau α si $F > f^{1-\alpha}_{p-p_0,n-(p+1)}$
où $f^{1-\alpha}_{p-p_0,n-(p+1)}$ est le quantile de niveau $1-\alpha$ de la loi de Fisher $\mathcal{F}(p-p_0, n-(p+1))$.

Remarque : En pratique, on peut regarder la p-value. On rejette H_0 si la p-value est très faible ($< 5\%$ par exemple).

Première application :

On considère le modèle de régression multiple expliquant l'ozone par 3 variables explicatives : le vent à 12h (Vx12), la température à 12h (T12) et la nébulosité à 12h (Ne12).

Question : peut-on considérer que $\beta_{\text{T12}} = 0$ et $\beta_{\text{Ne12}} = 0$? On compare donc un modèle à 3 variables explicatives avec un sous-modèle à 1 variable explicative.

Hypothèses du test de Fisher emboîté :

$$\begin{cases} H_0 : \beta_{\text{T12}} = \beta_{\text{Ne12}} = 0 \\ H_1 : \text{Au moins un des deux coefficients est non nul} \end{cases}$$

- Ajustement du "sous-modèle" :

```
reg_1 <- lm(maxO3~Vx12)
```

```
Call:
lm(formula = maxO3 ~ Vx12, data = ozone)

Residuals:
   Min     1Q Median     3Q    Max
-69.67 -14.61  -6.49  10.22  72.46

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  97.3009     2.7898  34.877  < 2e-16 ***
Vx12          4.3435     0.8675   5.007 2.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.55 on 110 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.1856,Adjusted R-squared:  0.1782
F-statistic: 25.07 on 1 and 110 DF,  p-value: 2.123e-06
```

- Ajustement du modèle "complet" :

`reg_2 <- lm(maxO3~Vx12+T12+Ne12)`

```
Call:
lm(formula = maxO3 ~ Vx12 + T12 + Ne12, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-37.462 -11.448  -0.722   8.908  46.331

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.8958    14.8243   0.263   0.7932
Vx12          1.6290     0.6571   2.479   0.0147 *
T12           4.5132     0.5203   8.674 4.71e-14 ***
Ne12         -1.6189     1.0181  -1.590   0.1147
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.63 on 108 degrees of freedom
  (10 observations deleted due to missingness)
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 Fisher de modèles emboîtés :

`anova(reg_1,reg_2)`

```
Analysis of Variance Table

Model 1: maxO3 ~ Vx12
Model 2: maxO3 ~ Vx12 + T12 + Ne12
  Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1    110 71825
2    108 29881  2     41944 75.8 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On a :

- Taille du modèle complet : $p+1=4$.
- Taille du sous-modèle : $p_0+1=2$.
- $n=112$ observations.
- $n-(p+1)=108$
- $p-p_0=3-1=2$
- $SCR=29881$
- $SCR_0=71825$

$\Rightarrow F=75,8$

On observe que la p-value du test est inférieure à $5\% \Rightarrow$ on rejette H_0.

On ne peut donc pas conclure à la nullité simultanée de $\beta_{\text{T12}} = 0$ et $\beta_{\text{Ne12}} = 0$: au moins une de ces deux variables influence le taux de pollution.

Deuxième application :

On considère le modèle de régression multiple expliquant l'ozone par 3 variables explicatives : le vent à 12h (Vx12), la température à 12h (T12) et la nébulosité à 12h (Ne12).

Question : peut-on considérer que $\beta_{\text{Vx12}} = 0$? On compare donc un modèle à 3 variables explicatives avec un sous-modèle à 2 variables explicatives.

Hypothèses du test de Fisher emboîté :

$$\begin{cases} H_0 : \beta_{\text{Vx12}} = 0 \\ H_1 : \beta_{\text{Vx12}} \neq 0 \end{cases}$$

- Ajustement du "sous-modèle" :

  `reg_1 <- lm(maxO3~T12+Ne12)`

```
Call:
lm(formula = maxO3 ~ T12 + Ne12, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-41.133 -11.799  -0.029   9.148  45.584

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

Residual standard error: 17.02 on 109 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.6419,Adjusted R-squared:  0.6353
F-statistic: 97.69 on 2 and 109 DF,  p-value: < 2.2e-16
```


- Ajustement du modèle "complet" :

  `reg_2 <- lm(maxO3~Vx12+T12+Ne12)`

```
Call:
lm(formula = maxO3 ~ Vx12 + T12 + Ne12, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-37.462 -11.448  -0.722   8.908  46.331

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.8958    14.8243   0.263   0.7932
Vx12          1.6290     0.6571   2.479   0.0147 *
T12           4.5132     0.5203   8.674 4.71e-14 ***
Ne12         -1.6189     1.0181  -1.590   0.1147
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.63 on 108 degrees of freedom
  (10 observations deleted due to missingness)
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 Fisher de modèles emboîtés :
  `anova(reg_1,reg_2)`

```
Analysis of Variance Table

Model 1: maxO3 ~ T12 + Ne12
Model 2: maxO3 ~ Vx12 + T12 + Ne12
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1    109 31581
2    108 29881  1    1700.3 6.1454 0.01472 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On a :

- Taille du modèle complet : $p + 1 = 4$.
- Taille du sous-modèle : $p_0 + 1 = 3$.
- $n = 112$ observations.
- $n - (p + 1) = 108$
- $p - p_0 = 3 - 2 = 1$
- $SCR = 29881$
- $SCR_0 = 31581$

$\Rightarrow F = 6,14$

On observe que la p-value du test est inférieure à 5% $\Rightarrow$ on rejette H_0.

On conserve donc le modèle complet : la variable Vx12 est utile (en plus de T12 et Ne12) pour expliquer l'ozone.

Remarque : Dans cet exemple les 2 modèles ne diffèrent que d'une variable (Vx12), donc le test de Fisher de modèles emboîtés est équivalent au test de Student de significativité du coefficient de la variable Vx12 dans le modèle 2 (cf la p-value).

2.11. Test de Fisher global

On veut tester la validité globale du modèle de régression multiple.

On pose les hypothèses de test :

$$\begin{cases} H_0 : \text{Tous les coefficients sont nuls sauf la constante} \\ H_1 : \text{Il existe au moins un coefficient non nul} \end{cases}$$

Statistique de test :

$$F = \frac{n-(p+1)}{p} \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Loi de la statistique de test sous H_0 : $F \sim \mathcal{F}(p, n-(p+1))$

$\Rightarrow$ On rejette H_0 au niveau α si $F > f^{1-\alpha}_{p,n-(p+1)}$
où $f^{1-\alpha}_{p,n-(p+1)}$ est le quantile de niveau $1-\alpha$ de la loi de Fisher $(p, n-(p+1))$.

Application :

```
Call:
lm(formula = maxO3 ~ T12 + Ne12 - 1, data = donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-40.331 -11.798   0.091   9.644  45.375

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
T12    4.7274     0.1371  34.492  < 2e-16 ***
Ne12  -2.3018     0.5449  -4.224 4.97e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.96 on 110 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared: 0.9684,     Adjusted R-squared: 0.9678
F-statistic:  1685 on 2 and 110 DF,  p-value: < 2.2e-16
```

$\Rightarrow$ La p-value est inférieure à 5%, on peut conclure au niveau 5% que le modèle contient au moins un coefficient significativement différent de 0.

Remarque : Dans le cas de la régression simple, le test de Fisher global est inutile puisqu'il revient à tester la significativité du paramètre de pente β_1 (déjà fait avec le test de Student).

3. Validation du modèle de régression

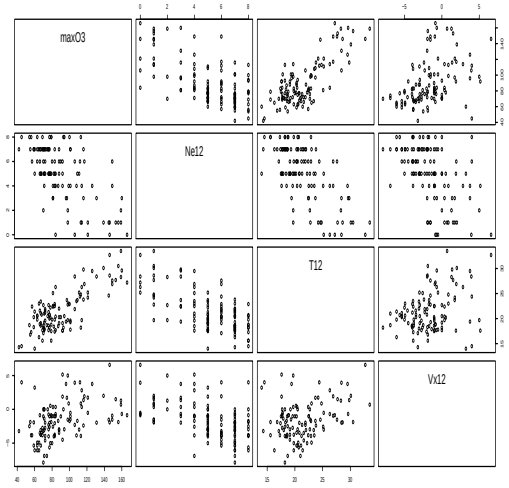
On doit faire plusieurs vérifications avant de conclure qu'un modèle est valide :

- Lien linéaire entre la variable à expliquer et les variables explicatives ;
- Non colinéarité des variables explicatives ;
- Erreurs centrées, non corrélées, de même variance ;
- Loi gaussienne sur les erreurs ;
- Étude des individus atypiques/influents.

3.1. Validation de l'hypothèse de lien linéaire

- Avant d'ajuster le modèle : on regarde les nuages de points entre la variable à expliquer et les variables explicatives.

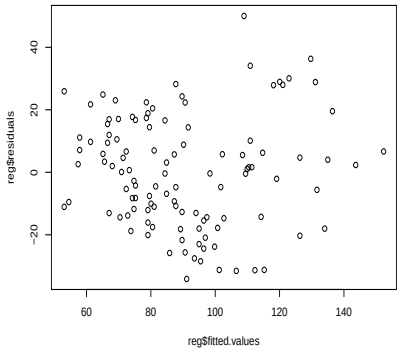
  Exemple pour l'ozone :



- Après avoir ajusté le modèle : on trace le nuage des points $(\hat{y}_i, \hat{e}_i)$ (résidus en fonction des valeurs ajustées)

  $\Rightarrow$ On ne doit pas observer de structure.

  Exemple :



  Si une structure est présente : essayer de transformer les variables explicatives pour rendre la relation linéaire ($\log$, $\sqrt{\ }$, etc).

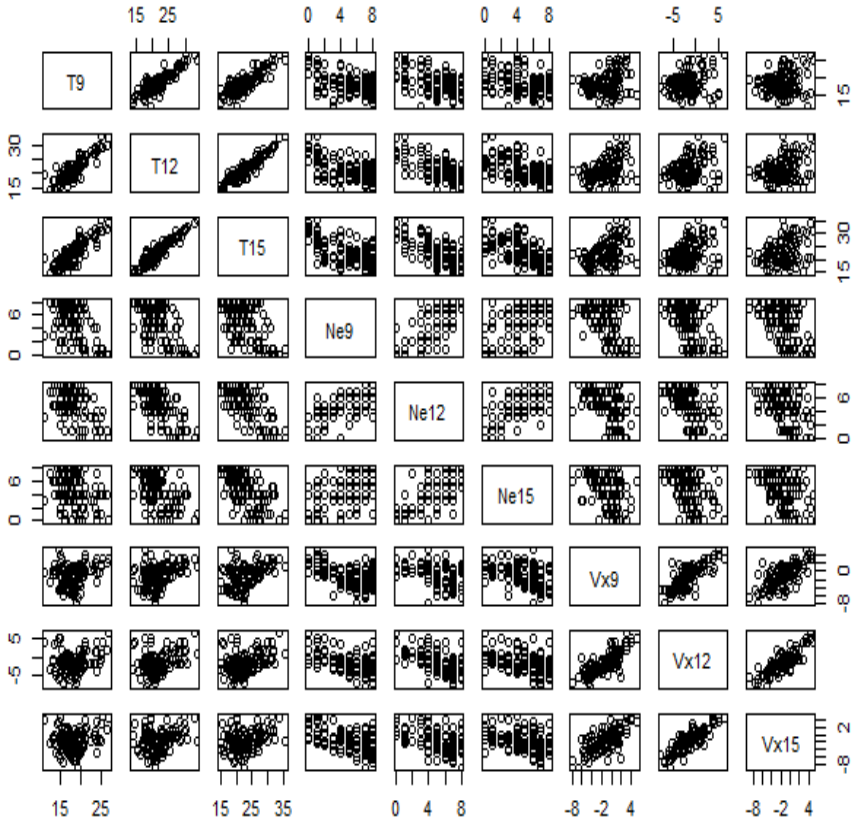
3.2. Validation de l'hypothèse de non colinéarité

Rappel : Les variables explicatives ne doivent pas être trop corrélées entre elles.

Pour le détecter :

- On regarde les nuages de points entre les variables explicatives.

Exemple pour l'ozone :



- On regarde le VIF ("variance inflation factor") : mesure de la dépendance linéaire d'une variable explicative par rapport aux autres.

  Soit x_j la $j$ème variable explicative. On fait la régression de x_j sur les autres variables explicatives (sauf x_j) $\Rightarrow$ on obtient R_j^2.

  On a alors :

$$VIF_j = \frac{1}{1 - R_j^2}$$

  et :

  - VIF_j toujours ≥ 1
  - Si $R_j^2 = 0$ alors $VIF_j = 1$ (la variable x_j n'est pas du tout colinéaire aux autres)
  - Si $R_j^2 \to 1$ alors $VIF_j \to \infty$. Un VIF élevé est donc signe de colinéarité (on peut prendre 10 par exemple comme valeur seuil, mais ce seuil est à adapter au cas par cas).

Exemple :

On ajuste le modèle expliquant l'ozone par deux variables : la température à 12h (T12) et la température à 15h (T15). On calcule les VIF avec R :

```
> reg<- lm (maxO3 ~ T12 + T15, data=donnees)
> library(car)
> vif(reg)
     T12      T15
9.549391 9.549391
```

$\Rightarrow$ les VIF sont élevés car la température à 12h est très corrélée à la température à 15h.

Pour résoudre le problème de colinéarité : si deux variables sont très corrélées on retire une des deux du modèle.

3.3. Validation des hypothèses sur les erreurs

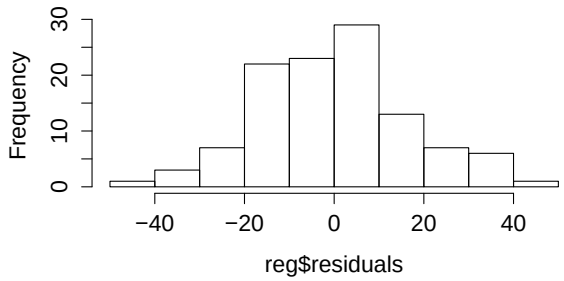
Rappel des hypothèses sur les erreurs e_i :

- les erreurs sont centrées ;
- les erreurs sont toutes de même variance ;
- les erreurs sont non corrélées entre elles.

Les erreurs e_i sont inconnues. On a montré comment les estimer par les résidus $\hat{e}_i$. On va donc utiliser ces résidus pour valider les hypothèses sur les erreurs.

Vérification des hypothèses sur les résidus :

☐ Les résidus sont centrés ?
Cette hypothèse est vérifiée par construction si on a une constante dans le modèle.



☐ Les résidus ont la même variance ? ("homoscédasticité")
Problème : on doit vérifier cette hypothèse sur les résidus qui ont des variances différentes par construction car $V(\hat{e}) = \sigma^2(I - X(X^TX)^{-1}X^T)$

$\Rightarrow$ On regarde plutôt les résidus standardisés :

$$t_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}},$$

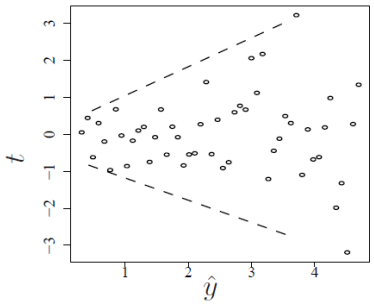
où $h_{ii} = [X(X^TX)^{-1}X^T]_{ii}$.

Les résidus standardisés sont donc tous de variance égale à 1.

Pour détecter une hétéroscédasticité, on trace le nuage de points des t_i en fonction des valeurs ajustées $\hat{y}_i$.

Si une structure apparaît (tendance, cône, vague), cela contredit l'hypothèse d'égalité des variances.

Exemple :



Les résidus ont une dispersion qui augmente quand $\hat{y}$ augmente $\Rightarrow$ variance non homogène.

Remarque : en pratique on utilise souvent les résidus "studentisés" au lieu des résidus "standardisés".

Les résidus studentisés s'écrivent :

$$t_i = \frac{\hat{e}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}},$$

où $\hat{\sigma}_{(i)}$ est l'estimateur dans le modèle linéaire privé de l'observation i.

Remarque : on peut tester l'homoscédasticité grâce au test de Breusch-Pagan (`bptest` sous R).

☐ Les résidus sont non corrélés entre eux ?
Cette question se pose surtout pour les données temporelles (observations ordonnées dans le temps).

Pour détecter une corrélation, on trace le nuage de points des $\hat{e}_i$ en fonction du temps.

⇒ Si une structure apparaît, possibilité d'une dépendance temporelle

On peut le tester avec :

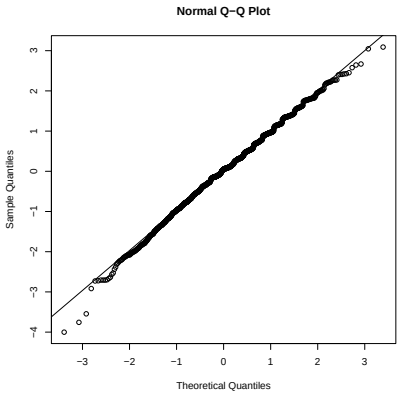
- le test de Durbin-Watson (`dwtest` sous R) ;
- le test de Breusch-Godfrey (`bgtest` sous R).

3.4. Validation de l'hypothèse gaussienne

Rappel : l'hypothèse gaussienne sur les erreurs e_i permet de construire des intervalles de confiance et de mettre en place des tests.

Remarque : Si le nombre d'individus est assez grand, il est inutile de la vérifier.

Sinon : vérification avec un qq-plot (droite de Henry) des résidus



Les points sont alignés $\Rightarrow$ les résidus suivent bien une loi gaussienne.

3.5. Etude des individus atypiques/influents

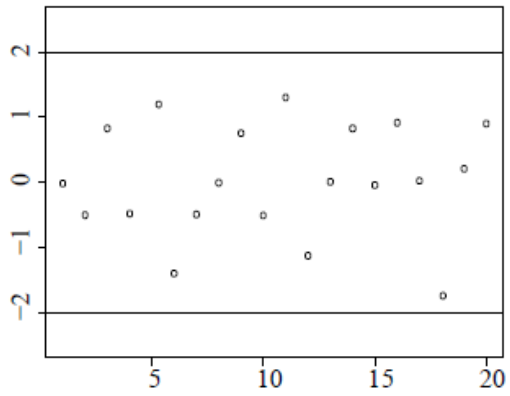
☐ Individu atypique :

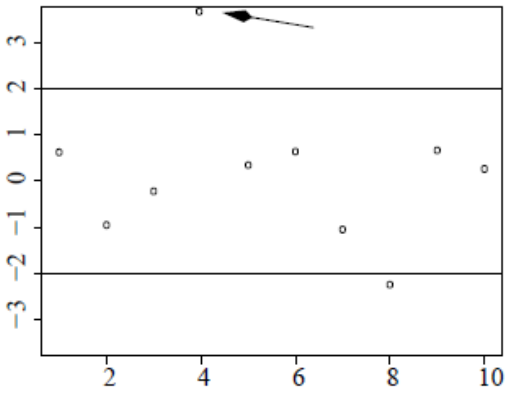
On se base sur l'analyse des résidus : un individu atypique est un individu ayant un résidu $\hat{e}_i$ anormalement élevé.

$\Rightarrow$ Critère : les résidus standardisés doivent être entre -2 et 2, sinon l'individu est considéré comme atypique.

Remarque : on prend la valeur 2 qui est proche du quantile à $97,5\%$ d'une loi gaussienne. Cela veut dire qu'en théorie 5% des valeurs sont considérées comme atypiques !

Exemple :





☐ Individu influent/levier :

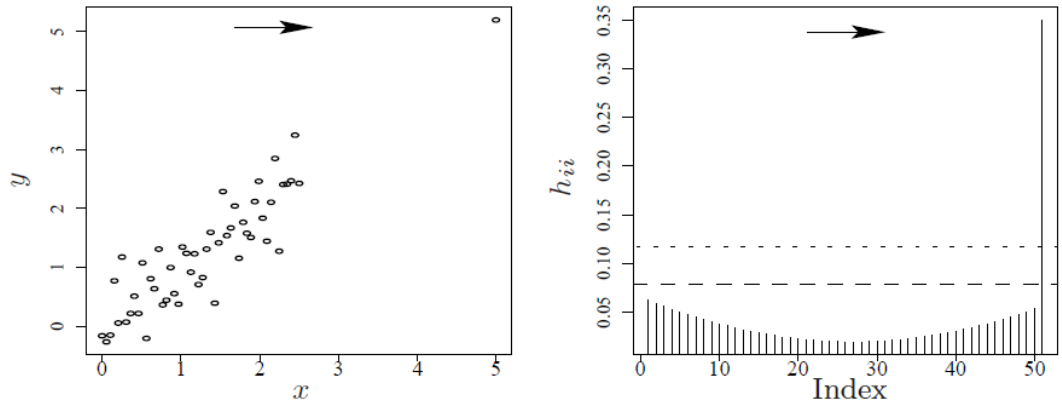
On se base sur la valeur de $h_{ii} = [X(X^TX)^{-1}X^T]_{ii}$. C'est le "poids" de l'observation i sur sa propre estimation.

Un individu est influent/levier si la valeur de h_{ii} dépasse :

- $\frac{2p}{n}$
- $\frac{3p}{n}$ pour $p > 6$ et $n - p > 12$

Un individu influent n'a pas forcément un résidu élevé mais influence fortement l'estimation des coefficients.

Exemple :



Le point isolé est levier mais pas atypique (il est loin du centre de gravité mais son résidu sera faible).

☐ Distance de Cook :

Critère pour mettre en évidence les points atypiques ou levier.

La distance de Cook C_i mesure l'influence de l'observation i sur l'estimation de β : on regarde la distance entre $\hat{\beta}$ et $\hat{\beta}_{(i)}$ calculé sans l'observation i

$$C_i = \frac{h_{ii}}{p(1-h_{ii})^2}\frac{\hat{e}_i^2}{\hat{\sigma}^2}$$

$\Rightarrow C_i$ est élevée si h_{ii} élevé ou/et $\frac{\hat{e}_i^2}{\hat{\sigma}^2}$ grand
$\Rightarrow C_i$ est élevée pour un individu atypique, influent, ou les deux

Seuils en pratique : $C_i > f_{p,n-p}(0.1)$ ("raisonnable") ou $C_i > f_{p,n-p}(0.5)$ ("préoccupant")

4. Choix de modèles
4.1. Critères de choix de modèles

Plusieurs critères de choix peuvent être utilisés :

☐ Le critère du R^2 (à maximiser) :

Rappel : le R^2 augmente avec le nombre de variables.

C'est donc un critère pour choisir entre plusieurs modèles ayant le même nombre de variables.

☐ Le critère du R^2 ajusté (à maximiser) :

Permet de choisir entre des modèles de tailles différentes.

☐ Autres critères à minimiser :

Le C_p de Mallows, l'AIC, le BIC.

☐ Choix entre deux modèles emboîtés :

On peut conclure avec le test de Fisher de modèles emboîtés.
Rappel des hypothèses :

$$\begin{cases} H_0 : \text{Modèle avec certains coefficients nuls} \\ H_1 : \text{Au moins un de ces coefficients est non nul} \end{cases}$$

4.2. Procédures de sélection de modèles

Objectif : éliminer les variables explicatives "inutiles" pour avoir le modèle le plus petit possible.

Intérêts :

- Diminuer le nombre de paramètres inconnus à estimer ;
- Avoir un modèle plus facile à interpréter.

Deux familles de méthodes :

- Méthodes exhaustives ;
- Méthodes pas à pas.

□ Méthode exhaustive :

Pour p variables explicatives il y a $2^p - 1$ modèles possibles.

Principe :

- On choisit un critère mesurant la qualité du modèle (R^2 ajusté, BIC) ;
- On calcule le critère pour les $2^p - 1$ modèles possibles ;
- On sélectionne le "meilleur" modèle.

$\Rightarrow$ Pas adapté si p est trop grand !

☐ Méthodes pas à pas :

Si p est trop grand, ces méthodes évitent la recherche exhaustive. Elles permettent de trouver un "bon" modèle, mais pas forcément le "meilleur".

Il existe différentes variantes :

- Elimination en arrière ("backward") ;
- Sélection en avant ("forward") ;
- Méthode mixte ("stepwise").

et chaque variante peut soit :

- utiliser des tests ;
- utiliser un critère de choix.

Elimination "backward" basée sur les tests :

- On part d'un gros modèle ;
- On vérifie si toutes les variables sont significatives (la "significativité" d'une variable est mesurée par la p-value du test de Student associé). Si c'est le cas on conserve le modèle complet ;
- Sinon, on enlève la variable la moins significative (ayant la p-value la plus grande) et on recommence avec le modèle réduit jusqu'à ce que tous les tests soient significatifs.

Elimination "backward" basée sur un critère de choix (exemple BIC) :

- On part d'un gros modèle ;
- Pour chaque variable : on la retire, on calcule le nouveau BIC du sous-modèle ;
- On garde le sous-modèle ayant le plus petit BIC ;
- Si aucun sous-modèle n'a un BIC plus faible, on garde le modèle courant.

Sélection "forward" basée sur les tests :

- On part du modèle réduit avec tous les coefficients nuls (sauf la constante) ;
- On ajoute une à une les variables les plus significatives (p-value la plus faible).

Sélection "forward" basée sur un critère de choix (exemple BIC) :

- On part du modèle réduit avec tous les coefficients nuls (sauf la constante) ;
- On ajoute la variable qui conduit au BIC le plus faible ;
- On s'arrête si le critère BIC ne diminue plus.

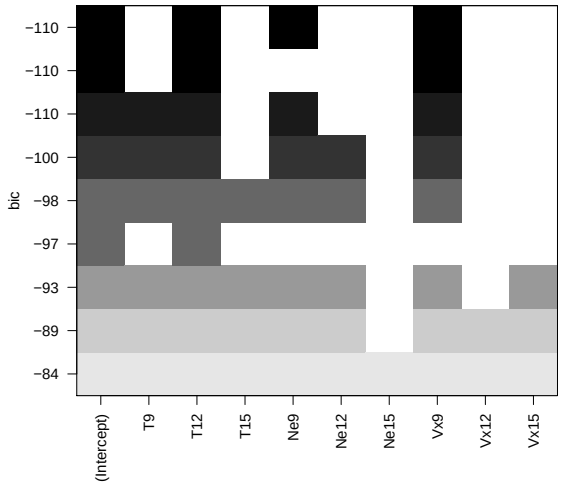
Sélection "stepwise" :

Même principe que pour les méthodes "forward", sauf que l'on peut éliminer des variables déjà introduites (il peut arriver que des variables introduites ne soient plus significatives après introduction de nouvelles variables).
$\Rightarrow$ A chaque étape de la sélection en avant, on effectue une étape d'élimination en arrière : on vérifie si toutes les variables sont significatives, on enlève celles qui ne le sont pas.

Exemple : Recherche exhaustive pour le jeu de données ozone :

```
library(leaps)
choix =regsubsets(maxO3~., nvmax=10, method="exhaustive",data=ozone)
plot(choix, scale="bic")
plot(choix,scale="Cp")
plot(choix,scale="adjr2")
```

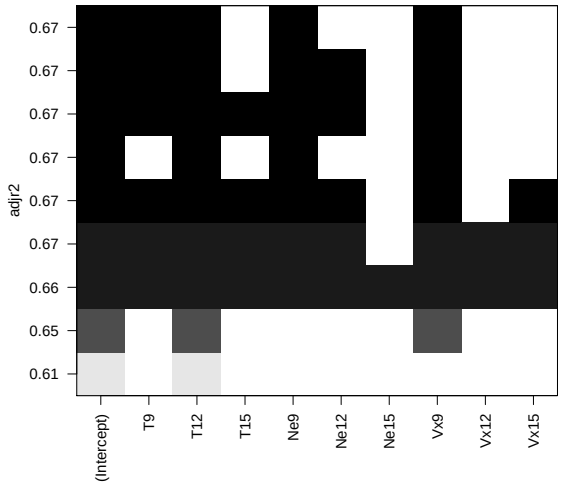
Résultat pour le critère BIC :



⇒ Modèle retenu :

$$\text{maxO3}_i = \beta_0 + \beta_1 \text{T12}_i + \beta_2 \text{Ne9}_i + \beta_3 \text{Vx9}_i + e_i \quad \forall i$$

Résultat pour le critère du R^2 ajusté :



⇒ Modèle retenu :

$$\text{maxO3}_i = \beta_0 + \beta_1 \text{T9}_i + \beta_2 \text{T12}_i + \beta_3 \text{Ne9}_i + \beta_4 \text{Vx9}_i + e_i \quad \forall i$$

Remarque : le critère du R^2 ajusté conduit souvent à sélectionner des modèles plus gros que pour les autres critères. Ici la sélection a conservé T9 et T12, alors que ces deux variables sont corrélées.

5. Analyse de la variance à 1 facteur

5.1. Introduction

En régression simple et multiple, la variable à expliquer et les variables explicatives étaient quantitatives. Mais il peut arriver que certaines variables explicatives soient qualitatives.

Exemple : On veut expliquer la concentration en ozone (O3) en fonction de la direction du vent : variable qualitative à 4 modalités (E,N,O,S)

Extrait du tableau de données :

Individu	O3	Vent
1	64	E
2	90	N
3	79	E
4	81	S
5	88	O
⋮	⋮	⋮

5.2. Notations

On note y la variable à expliquer (ici O3) et x la variable explicative (ici Vent).

x est une variable qualitative à I modalités (ici $I = 4$).

Les observations de la variable à expliquer y sont partagées en I groupes : le $i^{\text{ème}}$ groupe est constitué des n_i observations de la variable y admettant la modalité i de la variable explicative x.

Tableau des données regroupées par groupes :

Vent	E	N	O	S
O3	90 112.3 104	81 75	64 79	88 102.1 92.5

On a au total n observations avec $n = \sum_{i=1}^{I} n_i$.

$\Rightarrow$ Dans l'exemple : $I = 4$ et $n_1 = n_4 = 3$ et $n_2 = n_3 = 2$ donc $n = \sum_{i=1}^{4} n_i = 10$.

On note y_{ik} la valeur de la variable à expliquer pour l'individu k et la modalité i de la variable explicative.

$\Rightarrow$ Dans l'exemple : $y_{13} = 104$ et $y_{32} = 79$.

On note $\bar{y}_i$ la moyenne des valeurs de la variable à expliquer pour tous les individus ayant la modalité i :

$$\bar{y}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} y_{ik}$$

$\Rightarrow$ Dans l'exemple : $\bar{y}_4 = 94.2$.

On note $\bar{y}$ la moyenne générale des valeurs de la variable à expliquer pour tous les individus :

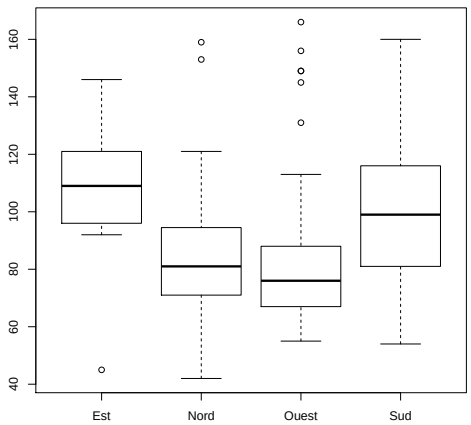
$$\bar{y} = \frac{1}{n} \sum_{i=1}^{I} \sum_{k=1}^{n_i} y_{ik} = \frac{1}{n} \sum_{i=1}^{I} n_i \bar{y}_i$$

$\Rightarrow$ Dans l'exemple : $\bar{y} = \frac{1}{10}(3 * 102.1 + 2 * 78 + 2 * 71.5 + 3 * 94.2) = 88.79$.

5.3. Etude graphique

On fait la représentation graphique de la variable à expliquer (quantitative) pour chacune des modalités de la variable explicative.

$\Rightarrow$ Ici on représente l'ozone en fonction des directions du vent :



Rappel de la question d'intérêt : la direction du vent a-t-elle un effet sur l'ozone ?

$\Rightarrow$ Oui d'après cette première étude graphique. On voit qu'en moyenne la concentration en pollution est :

- plus élevée quand le vent vient de l'est ;
- moins élevée quand le vent vient de l'ouest.

$\Rightarrow$ On va construire l'analyse de la variance pour répondre à cette question.

5.4. Modélisation et problématique

On note μ_i l'espérance de la variable à expliquer y pour la modalité i. On peut poser le modèle suivant :

$$y_{ik} = \mu_i + e_{ik} \quad \forall i = 1, \ldots, I \quad \text{et} \quad \forall k = 1, \ldots, n_i$$

Pour chaque groupe i, les valeurs de la variable à expliquer fluctuent donc autour d'une valeur moyenne μ_i. La fluctuation est décrite par e_{ik} pour chaque individu k.

On a :

- Les e_{ik} sont centrées, non corrélés entre elles et de même variance σ^2 ;
- Les paramètres $\mu_1, \ldots, \mu_I$ et σ^2 sont inconnus.

Pour répondre à la question "les facteurs de la variable explicative ont-ils un effet sur la variable à expliquer ?" on va chercher à tester l'égalité des μ_i pour toutes les modalités.

On veut donc tester :

$$\begin{cases} H_0 : \mu_1 = \ldots = \mu_I \\ H_1 : \exists (i,j) \quad \text{tel que} \quad \mu_i \neq \mu_j \end{cases}$$

Si on ne rejette pas H_0, c'est qu'il n'y a aucun effet.

5.5. Modélisation sous forme d'une régression

Le modèle précédent peut être réécrit sous la forme d'une régression multiple avec I variables explicatives :

$$y_k = \mu_1 x_{1k} + \mu_2 x_{2k} + \ldots + \mu_I x_{Ik} + e_k$$

où :

$$\begin{cases} x_{1k} = \mathbb{1}_{k \in \text{ groupe } 1} = \begin{cases} 0 & \text{si individu} \quad k \notin \text{groupe } 1 \\ 1 & \text{si individu} \quad k \in \text{groupe } 1 \end{cases} \\ \vdots \\ x_{Ik} = \mathbb{1}_{k \in \text{groupe } I} = \begin{cases} 0 & \text{si individu} \quad k \notin \text{groupe } I \\ 1 & \text{si individu} \quad k \in \text{groupe } I \end{cases} \end{cases}$$

Les variables explicatives $x_1, \ldots, x_I$ sont des variables binaires.

Ici on regarde chaque individu k du tableau de données : si l'individu k est associé à la modalité i on retrouve bien le modèle précédent :

$$y_{ik} = \mu_i + e_{ik}$$

car :

- $x_{ik} = 1$;
- $x_{i'k} = 0$ pour toutes les modalités i' autres que i.

Exemple :

Individu	O3	vent
1	64	E
2	90	N
3	79	E
4	81	S
5	88	O
⋮	⋮	⋮

Alors :

- $x_{11} = 1$, $x_{21} = 0$, $x_{31} = 0$, $x_{41} = 0$
- $x_{12} = 0$, $x_{22} = 1$, $x_{32} = 0$, $x_{42} = 0$
- $x_{13} = 1$, $x_{23} = 0$, $x_{33} = 0$, $x_{43} = 0$
- ...

$\Rightarrow$ On peut écrire le modèle sous la forme $Y = X\mu + e$ avec :

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ \vdots & & & \end{pmatrix}$$

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_I \end{pmatrix}$$

$$e = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

5.6. Contraintes sur le modèle

Ce modèle est un cas particulier d'un modèle de régression multiple avec constante :

$$y_k = m + \alpha_1 x_{1k} + \alpha_2 x_{2k} + \ldots + \alpha_I x_{Ik} + e_k$$

Remarque : avec la constante, la matrice des variables explicatives devient :

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ \vdots & & & & \end{pmatrix}$$

$\Rightarrow$ On a alors un problème de colinéarité car la première colonne est la somme des autres.

Pour résoudre ce problème, on impose une contrainte au modèle en retirant une variable.

On peut par exemple imposer :

- $m = 0$. Alors : $\alpha_i = \mu_i$;
- $\alpha_1 = 0$ (on retire la variable x_1). Alors :

$$\begin{cases} m = \mu_1 \\ \alpha_i = \mu_i - \mu_1 \end{cases}$$

- D'autres contraintes sont possibles...

Remarque : La deuxième contrainte (α_1=0, retrait de la première variable) est celle utilisée par le logiciel R en pratique.

Avantage de cette contrainte : tester l'égalité des μ_i ($\mu_1 = \ldots = \mu_I$) dans le modèle d'analyse de la variance revient à tester la nullité des α_i ($\alpha_1 = \ldots = \alpha_I = 0$) dans le modèle de régression multiple.

$\Rightarrow$ on pourra utiliser le test de Fisher global vu dans le cours de régression.

5.7. Estimation des paramètres

Sous la contrainte $\alpha_1 = 0$, les estimateurs des moindres carrés ordinaires des paramètres inconnus (m et $\alpha_1, \cdots, \alpha_I$) sont :

$$\begin{cases} \hat{m} = \bar{y}_1 \\ \hat{\alpha}_i = \bar{y}_i - \bar{y}_1 \end{cases}$$

Remarque : ce sont ces valeurs qui sont affichées dans la sortie R.

Donc, dans le modèle d'analyse de la variance :

$$\begin{cases} \hat{\mu}_1 = \bar{y}_1 \\ \hat{\mu}_i = \hat{m} + \hat{\alpha}_i = \bar{y}_i \end{cases}$$

Enfin, σ^2 (la variance des erreurs) est estimée par :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{I} \sum_{k=1}^{n_i} (y_{ik} - \bar{y}_i)^2}{n - I}$$

Exemple de l'ozone. Application avec le logiciel R.

- Estimation :

  ```
  > mod <- lm(O3~vent, data=ozone)
  ```

- Résultats :

  ```
  > summary(mod)
  ```

  On obtient :

```
Call:
lm(formula = maxO3 ~ vent, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-60.600 -16.807  -7.365  11.478  81.300

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
  (10 observations deleted due to missingness)
Multiple R-squared: 0.08602,    Adjusted R-squared: 0.06063
F-statistic: 3.388 on 3 and 108 DF,  p-value: 0.02074
```


R a choisi par défaut la contrainte α_1=0. Cela revient à prendre le groupe "Est" comme groupe de référence (première modalité dans l'ordre alphabétique).

Donc :

- "Intercept" $= \hat{m} = \bar{y}_1$ (moyenne du groupe de référence)
- "ventNord" $= \hat{\alpha}_2 = \bar{y}_2 - \bar{y}_1$
- "ventOuest" $= \hat{\alpha}_3 = \bar{y}_3 - \bar{y}_1$
- "ventSud" $= \hat{\alpha}_4 = \bar{y}_4 - \bar{y}_1$

5.8. Test d'influence de la variable explicative

Rappel : On a vu que le modèle d'analyse de la variance

$$y_{ik} = \mu_i + e_{ik} \quad \forall i = 1, \ldots, I \quad \text{et} \quad \forall k = 1, \ldots, n_i$$

peut s'écrire sous la forme d'un modèle de régression multiple

$$y_k = m + \alpha_1 x_{1k} + \alpha_2 x_{2k} + \ldots + \alpha_I x_{Ik} + e_k$$

avec (sous la contrainte $\alpha_1 = 0$) :

$$\begin{cases} m = \mu_1 \\ \alpha_i = \mu_i - \mu_1 \end{cases}$$

Pour conclure si la variable explicative à I modalités a un effet sur la variable à expliquer, on s'intéresse au test :

$$\begin{cases} H_0 : \mu_1 = \ldots = \mu_I \\ H_1 : \exists (i,j) \quad \text{tel que} \quad \mu_i \neq \mu_j \end{cases}$$

avec : si on ne rejette pas H_0, il n'y a aucun effet.

À partir de l'écriture du modèle sous forme d'une régression multiple, on voit que ce test est équivalent à :

$$\begin{cases} H_0 : \alpha_1 = \ldots = \alpha_I = 0 \\ H_1 : \exists i \quad \text{tel que} \quad \alpha_i \neq 0 \end{cases}$$

$\Rightarrow$ On est dans le cadre du test de Fisher global vu dans le cours de régression.

Rappels :

- i est l'indice pour la modalité de la variable explicative ;
- n_i est le nombre d'observations dans le groupe i ;
- y_{ik} est l'observation de y pour l'individu k du groupe i ;
- $\bar{y}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} y_{ik}$ est la moyenne pour le groupe i ;
- $\bar{y}$ est la moyenne générale.

La statistique de test du test d'influence de la variable explicative s'écrit alors :

$$F = \frac{n-I}{I-1} \frac{\sum_{i=1}^{I} n_i (\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{I} \sum_{k=1}^{n_i} (y_{ik} - \bar{y}_i)^2}$$

et on a sa loi sous l'hypothèse H_0 :

$$F \underset{H_0}{\sim} \mathcal{F}(I-1, n-I).$$

$\Rightarrow$ On rejette H_0 au niveau α si $F > f_{I-1,n-I}(1-\alpha)$
où $f_{I-1,n-I}(1-\alpha)$ est le quantile de niveau α de la loi de Fisher $(I-1, n-I)$.

ou bien :

on rejette H_0 au niveau α si la p-value (probabilité pour la statistique de test sous H_0 de dépasser la valeur estimée) est plus petite que α.

Dans ce cas on conclut que la variable explicative a un effet sur la variable à expliquer.

Exemple de l'ozone. Application avec le logiciel R.

- On a estimé le modèle :

```
> mod <- lm(O3~vent, data=ozone)
> summary(mod)
```

- On teste l'influence du vent sur la pollution :

```
> anova(mod)
```

On obtient :

```
Analysis of Variance Table

Response: maxO3
           Df Sum Sq Mean Sq F value  Pr(>F)
vent        3   7586 2528.69  3.3881 0.02074 *
Residuals 108  80606  746.35
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


On a dans cette sortie :

- "Df" : $I - 1 = 3$ et $n - I = 108$
- "Sum Sq" : $\sum_{i=1}^{I} n_i(\bar{y}_i - \bar{y})^2 = 7586$ et $\sum_{i=1}^{I}\sum_{k=1}^{n_i}(y_{ik} - \bar{y}_i)^2 = 80606$
- "Mean Sq" : $\frac{\sum_{i=1}^{I} n_i(\bar{y}_i - \bar{y})^2}{I-1} = 2528.69$ et $\frac{\sum_{i=1}^{I}\sum_{k=1}^{n_i}(y_{ik} - \bar{y}_i)^2}{n-I} = 746.35$
- "F value" : $F = 3.3881$
- "Pr(>F)" : p-value $= 0.02074$

$\Rightarrow$ On conclut ici au rejet de H_0 au niveau 5%.

$\Rightarrow$ La direction du vent a donc un effet sur la concentration en ozone.

Remarque : on lit également le résultat dans la dernière ligne de `summary(mod)` :

```
Call:
lm(formula = maxO3 ~ vent, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-60.600 -16.807  -7.365  11.478  81.300

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
  (10 observations deleted due to missingness)
Multiple R-squared: 0.08602,    Adjusted R-squared: 0.06063
F-statistic: 3.388 on 3 and 108 DF,  p-value: 0.02074
```


5.9. Décomposition de la variance

Remarque : La variance totale $\frac{1}{n}\sum_{i=1}^{I}\sum_{k=1}^{n_i}(y_{ik}-\bar{y})^2$ se décompose en somme de deux termes :

- la variance inter-groupes $\frac{1}{n}\sum_{i=1}^{I} n_i(\bar{y}_i-\bar{y})^2$
- la variance intra-groupes $\frac{1}{n}\sum_{i=1}^{I}\sum_{k=1}^{n_i}(y_{ik}-\bar{y}_i)^2$

La statistique de test pour le test d'influence de la variable explicative s'écrit donc :

$$F = \frac{n-I}{I-1}\,\frac{\text{Variance inter-groupes}}{\text{Variance intra-groupes}}.$$

$\Rightarrow$ Ceci exlique l'intitulé du cours "analyse de la variance".

Intuition du test : Si les moyennes des groupes sont proches ($\Leftrightarrow$ aucun effet de la variable explicative), la variance inter-groupes est faible, on a tendance à ne pas rejeter H_0.

Preuve de la décomposition :

$$\frac{1}{n}\sum_{i=1}^{I}\sum_{k=1}^{n_i}(y_{ik}-\bar{y})^2 = \frac{1}{n}\sum_{i=1}^{I}\sum_{k=1}^{n_i}(y_{ik}-\bar{y}_i+\bar{y}_i-\bar{y})^2$$

$$= \frac{1}{n}\sum_{i=1}^{I}\sum_{k=1}^{n_i}(y_{ik}-\bar{y}_i)^2 + \frac{2}{n}\sum_{i=1}^{I}\sum_{k=1}^{n_i}(y_{ik}-\bar{y}_i)(\bar{y}_i-\bar{y}) + \frac{1}{n}\sum_{i=1}^{I}\sum_{k=1}^{n_i}(\bar{y}_i-\bar{y})^2$$

Or on a :

$$\sum_{i=1}^{I}\sum_{k=1}^{n_i}(y_{ik}-\bar{y}_i)(\bar{y}_i-\bar{y})$$

$$= \sum_{i=1}^{I} n_i\bar{y}_i^2 - \sum_{i=1}^{I} n_i\bar{y}\bar{y}_i - \sum_{i=1}^{I} n_i\bar{y}_i^2 + \sum_{i=1}^{I} n_i\bar{y}\bar{y}_i \quad \text{car} \quad \sum_{k=1}^{n_i} y_{ik} = n_i\bar{y}_i$$

$$= 0$$

5.10. Comparaisons multiples de moyennes

Rappel : lorsqu'on teste l'effet de la variable explicative, on se base sur les hypothèses suivantes :

$$\begin{cases} H_0 : \mu_1 = \ldots = \mu_I \\ H_1 : \exists(i,j) \quad \text{tel que} \quad \mu_i \neq \mu_j \end{cases}$$

$\Rightarrow$ le rejet de H_0 amène à conlure qu'au moins deux moyennes sont différentes. Mais on ne sait pas quelles sont les paires de moyennes qui diffèrent.

Pour répondre à cela on fait un test de comparaisons multiples de moyennes : le test de Tukey.

Il y a I sous-groupes. Il y a donc $\frac{I(I-1)}{2}$ paires de moyennes à comparer. Pour chaque paire (μ_i, μ_j), on veut tester :

$$\begin{cases} H_0 : \mu_i = \mu_j \\ H_1 : \mu_i \neq \mu_j \end{cases}$$

On utilise la statistique de test :

$$T = \frac{\max\limits_{i=1,\dots,I} \bar{y}_i - \min\limits_{i=1,\dots,I} \bar{y}_i}{\hat{\sigma}}$$

On connait la loi de cette statistique sous H_0 : loi appelée "Etendue studentisée".

Pour chaque test, on compare alors $\frac{\bar{y}_i - \bar{y}_j}{\hat{\sigma}}$ au quantile $q_{1-\alpha}$ de niveau $1-\alpha$ de cette loi.

$\Rightarrow$ En pratique on regarde la p-value pour chaque paire.

Avantage du test : on a un risque de première espèce global égal à α, contrairement à une succession de comparaison de moyennes qui augmente le risque.

Exemple de l'ozone (données hebdomadaires) :

- $I = 4$
- $\frac{I(I-1)}{2} = 6$ paires de moyennes à comparer

$\Rightarrow$ On obtient pour un niveau $\alpha = 5\%$:

- μ_N et μ_E sont significativement différentes
- μ_O et μ_E sont significativement différentes


```
Analysis of Variance Table

Response: maxO3
          Df  Sum Sq Mean Sq F value    Pr(>F)
vent       3  9859.8  3286.6  8.3383 0.0001556 ***
Residuals 46 18131.4   394.2



Tukey multiple comparisons of means
  95% family-wise confidence level

Fit: aov(formula = ozone$maxO3 ~ ozone$vent)

$ozone$vent
          diff       lwr         upr     p adj
N-E -25.561111 -47.61084  -3.5113791 0.0172421
S-E  -9.507143 -33.48826  14.4739772 0.7171820
O-E -32.272222 -50.45490 -14.0895474 0.0001236
S-N  16.053968 -10.61488  42.7228174 0.3861919
O-N  -6.711111 -28.31535  14.8931259 0.8408998
O-S -22.765079 -46.33723   0.8070757 0.0618631
```

5.11. Validation du modèle

☐ Validation de l'hypothèse gaussienne :

Le test d'influence de la variable explicative repose sur l'hypothèse de normalité des erreurs.

On doit donc vérifier :

$$e_{ik} \sim \mathcal{N}(0, \sigma^2)$$

⇒ On se base sur l'observation graphique de la distribution des résidus : Histogramme et droite de Henry (qq-plot).

Exemple : comparaison de 3 traitements contre l'asthme.

Traitement A	Traitement B	Traitement C
26 ; 27 ; 35 ; 36	29 ; 42 ; 44 ; 44	26 ; 26 ; 30 ; 30
38 ; 38 ; 41 ; 42	45 ; 48 ; 48 ; 52	33 ; 36 ; 38 ; 38
45 ; 50 ; 65	56 ; 56 ; 58 ; 58	39 ; 46 ; 47 ; 51
	60 ; 61 ; 63 ; 63	51 ; 56 ; 75
	69	

On pose le modèle $Duree_{ik} = \mu_i + e_{ik}$ où μ_i est la moyenne des durées pour chaque traitement (A, B ou C).

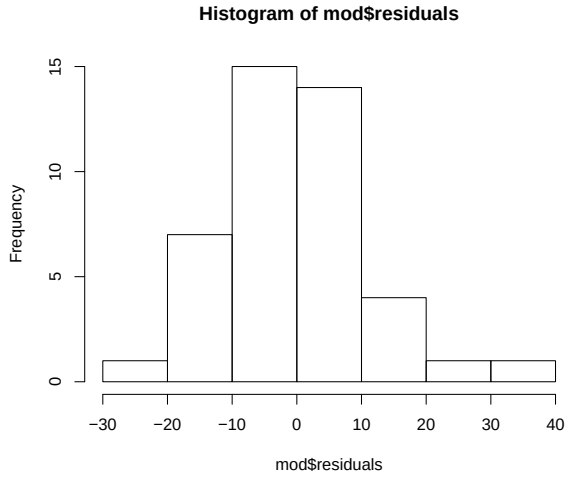
Les résidus sont définis par :

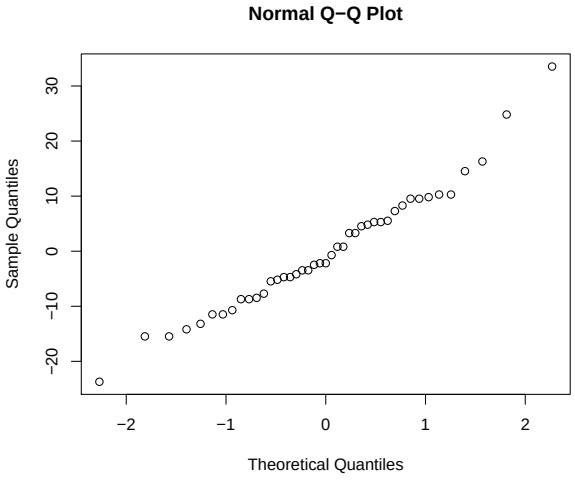
$$\hat{e}_{ik} = Duree_{ik} - \hat{\mu}_i = Duree_{ik} - \overline{Duree_i}$$

où $\overline{Duree_i}$ est la moyenne empirique dans chaque groupe.

$\Rightarrow$ le résidu d'un individu est l'écart à la moyenne de son groupe.


```
> mod=lm(Duree~Traitement, data=asthme)
> hist(mod$residuals)
> qqnorm(mod$residuals)
```





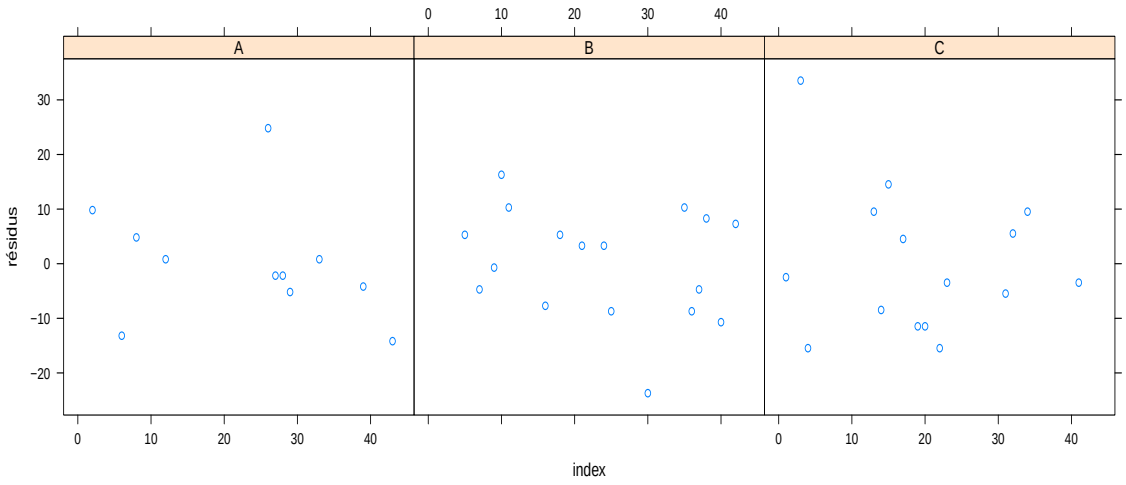
☐ Validation de l'hypothèse d'homoscédasticité :

On doit vérifier que la variance des erreurs est la même dans chacun des I sous-groupes.

⇒ Pour cela on commence par une validation graphique, puis on mettra en place des tests statistiques.

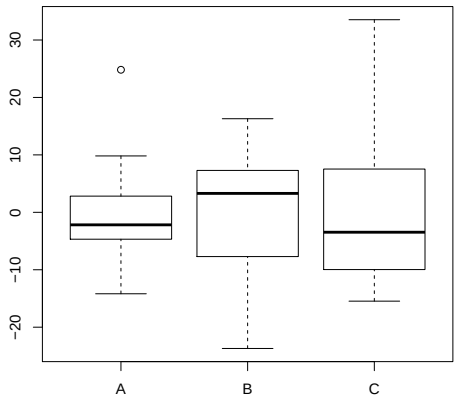
- Validation graphique :

  Exemple des traitements contre l'asthme : on observe la dispersion des résidus dans chaque sous-groupe (A,B et C)



  ⇒ L'hypothèse d' homoscédasticité semble vérifiée.

On peut aussi utiliser le graphique suivant pour comparer les dispersions :



- Validation par des tests :

  On suppose que l'hypothèse gaussienne est vérifiée. On peut alors utiliser différents tests :

  - Test de Fisher d'égalité de variances (si le facteur n'a que deux modalités) ;
  - Tests de Bartlett, Levene, etc... (pour un facteur à plus de deux modalités).

Premier cas : le facteur a deux modalités.

$\Rightarrow$ Test de Fisher d'égalité de variances dans les deux sous-groupes.

Les hypothèses de test sont :

$$\begin{cases} H_0 : \sigma_1 = \sigma_2 \\ H_1 : \sigma_1 \neq \sigma_2 \end{cases}$$

On se base sur la statistique de test :

$$F = \frac{\frac{n_1}{n_1-1} s_1^2}{\frac{n_2}{n_2-1} s_2^2}$$

où :

- n_i est le nombre d'individus du sous-groupe i $(i = 1, 2)$;
- s_i^2 est la variance empirique du sous-groupe i $(i = 1, 2)$.

On a : $F \underset{H_0}{\sim} \mathcal{F}(n_1 - 1, n_2 - 1)$.

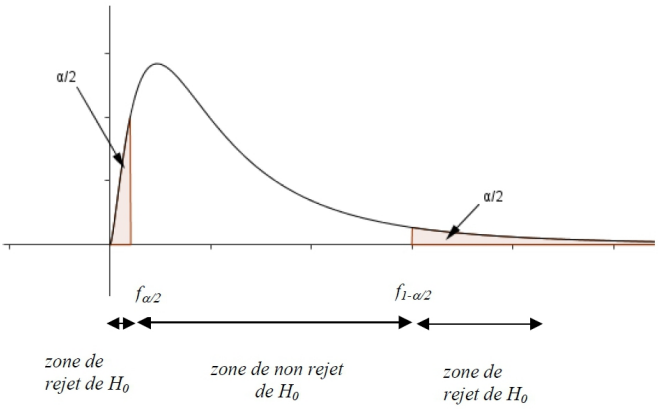
On rejette H_0 dans deux situations :

- s_1^2 est trop grande par rapport à $s_2^2 \Rightarrow$ grandes valeurs de F ;
- s_1^2 est trop petite par rapport à $s_2^2 \Rightarrow$ petites valeurs de F.

Finalement, rejet de H_0 au niveau α dans deux cas :

- $F > f_{n_1-1,n_2-1}(1-\frac{\alpha}{2})$;
- $F < f_{n_1-1,n_2-1}(\frac{\alpha}{2})$.

$\Rightarrow$ Dans ce cas l'hypothèse d'homoscédasticité n'est pas vérifiée.



Deuxième cas : le facteur a I modalités.

$\Rightarrow$ Test de Bartlett d'égalité des variances dans les I sous-groupes.

Les hypothèses de test sont :

$$\begin{cases} H_0 : \sigma_1 = \sigma_2 = \ldots = \sigma_I \\ H_1 : \exists (i,j) \quad \text{tel que} \quad \sigma_i \neq \sigma_j \end{cases}$$

On se base sur la statistique de test :

$$X = \frac{(N-I)\ln(s^2) - \sum_{i=1}^{I}(n_i-1)\ln(s_i^2)}{1 + \frac{1}{3(I-1)}\left(\sum_{i=1}^{I}\left(\frac{1}{n_i-1}\right) - \frac{1}{N-I}\right)}$$

où :

- $N = \sum_{i=1}^{I} n_i$
- $s^2 = \frac{1}{n-I} \sum_{i=1}^{I} (n_i - 1) s_i^2$

On sait que X suit approximativement une loi $\chi^2(I-1)$.

On rejette H_0 si $X > \chi^2_{1-\alpha}(I-1)$.

$\Rightarrow$ Dans ce cas l'hypothèse d'homoscédasticité n'est pas vérifiée.

Alternative : le test de Levene.

Les hypothèses de test sont toujours :

$$\begin{cases} H_0 : \sigma_1 = \sigma_2 = \ldots = \sigma_I \\ H_1 : \exists (i,j) \quad \text{tel que} \quad \sigma_i \neq \sigma_j \end{cases}$$

On se base sur la statistique de test :

$$W = \frac{(N-I) \sum_{i=1}^{I} n_i (z_{i\cdot} - z_{\cdot\cdot})^2}{(I-1) \sum_{i=1}^{I} \sum_{j=1}^{n_i} (z_{ij} - z_{i\cdot})^2}$$

où :

$$z_{ij} = |y_{ij} - \bar{y}_i|$$

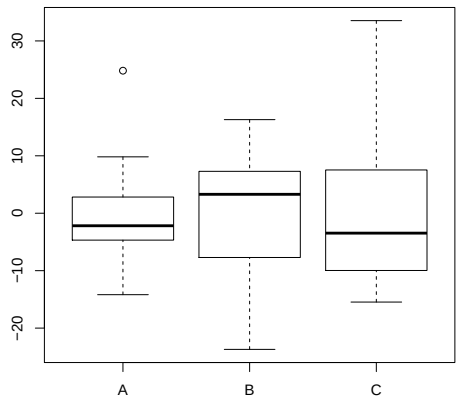
On a : $W \underset{H_0}{\sim} \mathcal{F}(I-1, n-I)$.

On rejette H_0 si $W > f_{1-\alpha}(I-1, n-I)$.

$\Rightarrow$ Dans ce cas l'hypothèse d'homoscédasticité n'est pas vérifiée.

Exemple 1 : comparaison de traitements contre l'asthme.

Rappel de la vérification graphique de l'homogénéité des variances :



☐ Test de Bartlett :

```
> bartlett.test(residus~Traitement, data=asthme)
```

Résultat :

```
Bartlett test of homogeneity of variances

data:  residus by Traitement
Bartlett's K-squared = 1.1687, df = 2, p-value = 0.5575
```

⇒ On ne rejette pas H_0, on peut donc valider l'hypothèse d'homoscédasticité.

☐ Test de Levene :

```
> library(car)
> leveneTest(residus~Traitement,data=asthme, center=mean)
```

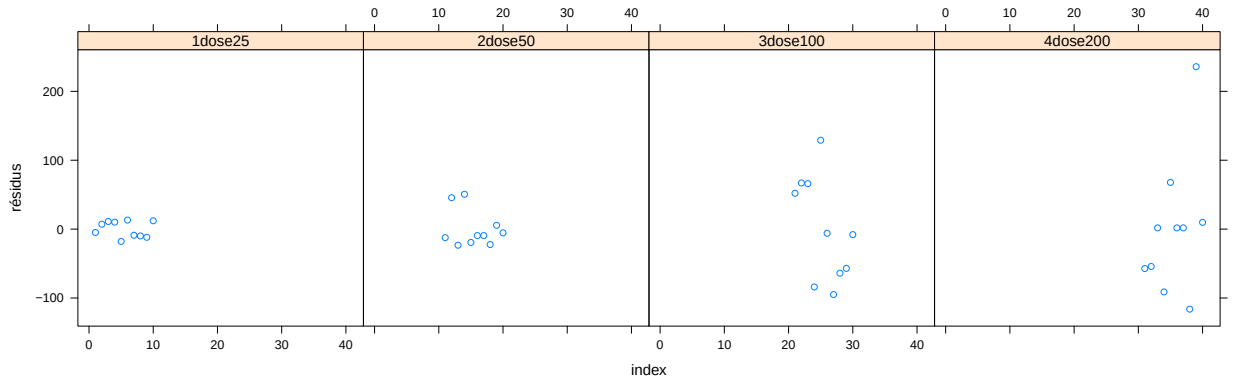
Résultat :

```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  2    0.63 0.5378
      40
```

⇒ Même conclusion.

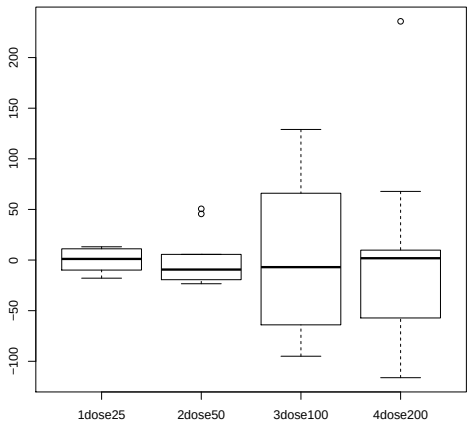
Exemple 2 : concentration dans le sang (en ng/ml) d'un certain produit chimique chez 40 patients ayant absorbé un médicament dosé à 25, 50, 100 ou 200 mg de substance active.

Observation graphique des résidus pour chaque sous-groupe :



⇒ On ne valide pas l'hypothèse d'égalité des variances dans les sous-groupes.

Autre graphique possible :



⇒ Même conclusion.

Confirmation par les tests statistiques :

```
Bartlett test of homogeneity of variances

data:  residus by dose
Bartlett's K-squared = 33.6901, df = 3, p-value = 2.303e-07
```

```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value   Pr(>F)
group  3  4.4767 0.009028 **
      36
```

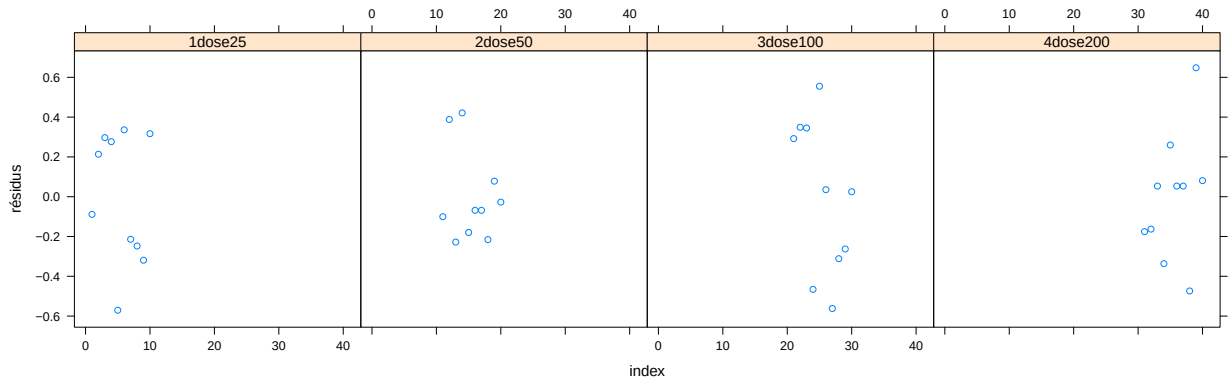
$\Rightarrow$ Les deux tests confirment l'observation des graphiques : l'homoscédasticité n'est pas vérifiée.

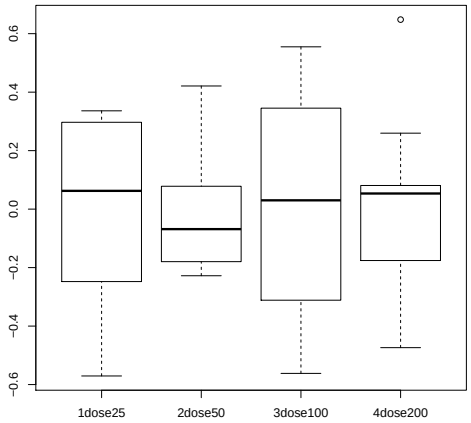
Solutions possibles quand l'hypothèse d'homoscédasticité n'est pas valide :

- Essayer de transformer la variable à expliquer pour se ramener à l'homoscédasticité ;
- ANOVA de type Welch ;
- MCG (moindres carrés généralisés).

Retour à l'exemple 2 : on applique une transformation de type log à la variable à expliquer (la concentration).

Nouveaux graphiques des résidus :





$\Rightarrow$ Il est maintenant raisonnable de valider l'hypothèse d'égalité des variances dans les sous-groupes.

Confirmation avec les tests :

```
Bartlett test of homogeneity of variances

data:  residus by dose
Bartlett's K-squared = 2.1285, df = 3, p-value = 0.5462
```

```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  3  1.4608 0.2415
      36
```

$\Rightarrow$ Les deux tests confirment l'observation des graphiques : l'hypothèse d'homoscédasticité est vérifiée.

6. Analyse de la covariance

On cherche à expliquer une variable quantitative par une variable quantitative et une variable qualitative.

Exemple : On veut expliquer la concentration en ozone par la température à 12h et le vent ("Est", "Nord", "Ouest", "Sud").

Table des données :

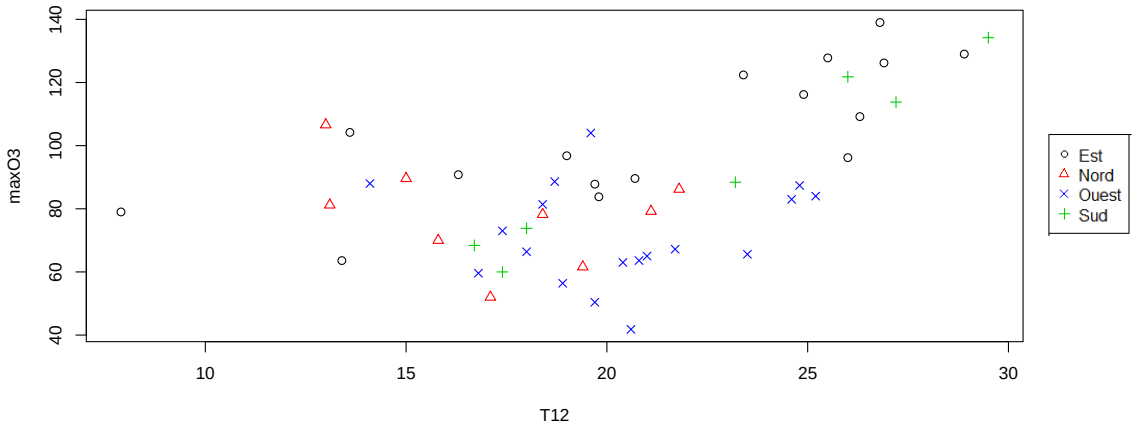
Individu	O3	T12	vent
1	63.6	18.3	E
2	81.2	18.4	N
3	139	25	E
4	68.4	17	S
5	87.4	19.2	O
⋮	⋮	⋮	⋮

6.1. Représentation graphique

On commence par faire une représentation graphique avec :

- la variable à expliquer (quantitative) en ordonnée ;
- la variable explicative (quantitative) en abscisse ;
- le facteur explicatif avec des couleurs (ou des symboles).

Exemple :



6.2. Modélisation

Première approche : on fait une régression linéaire simple pour chaque modalité i du facteur ($i = 1, \ldots, I$).

On note y la variable à expliquer et y_{ik} la valeur prise par l'individu k ayant la modalité i pour le facteur.

On note x la variable explicative quantitative et x_{ik} la valeur prise par l'individu k ayant la modalité i pour le facteur.

On pose donc le modèle :

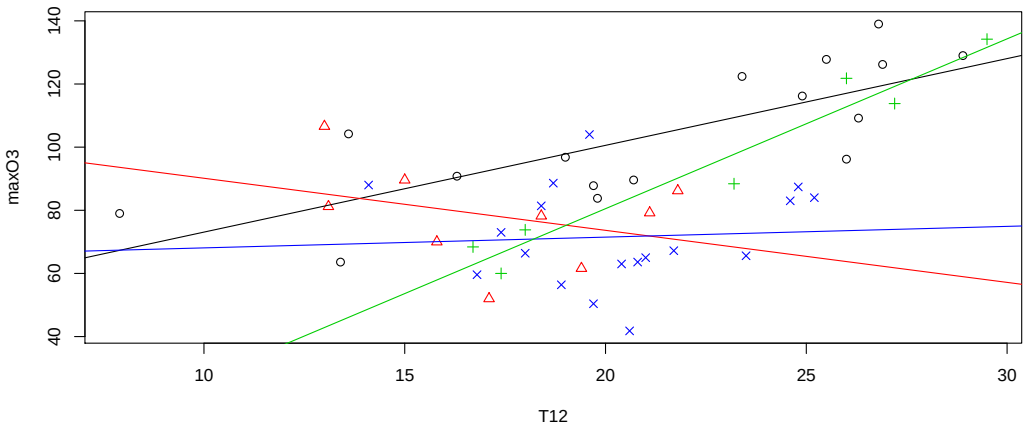
$$y_{ik} = \alpha_i + \gamma_i x_{ik} + e_{ik} \quad \forall k = 1, \ldots, n_i \quad i = 1, \ldots, I$$

où :

- α_i est l'ordonnée à l'origine du modèle i ;
- γ_i est la pente du modèle i ;
- les e_{ik} sont les erreurs.

Exemple de l'ozone : on pose un modèle de régression simple (entre l'ozone et la température) pour chacune des directions du vent.

On obtient après estimation :



Ce modèle d'analyse de la covariance comporte $2I + 1$ paramètres :

- I ordonnées à l'origine différentes ;
- I pentes différentes ;
- la variance des erreurs σ^2.

Mais en statistique, on recherche toujours le modèle le plus parcimonieux possible.

$\Rightarrow$ Est-il possible de simplifier ce modèle (réduire le nombre de paramètres à estimer) ?

6.3. Simplifications du modèle

☐ Première simplification possible :

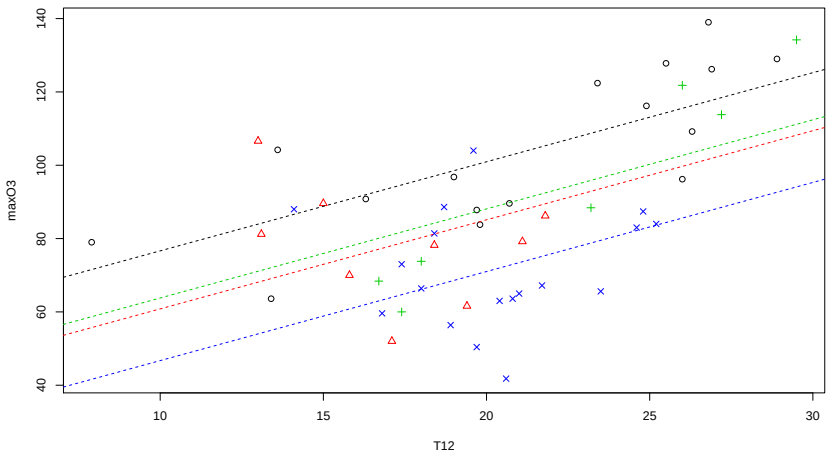
Si on peut considérer que les pentes des I modèles de régression sont égales, on pose donc le modèle :

$$y_{ik} = \alpha_i + \gamma x_{ik} + e_{ik} \quad \forall k = 1, \ldots, n_i \quad i = 1, \ldots, I$$

On obtient alors I droites de régression parallèles.

Ce modèle a $I + 2$ paramètres.

Exemple de l'ozone :



☐ Autre simplification possible :

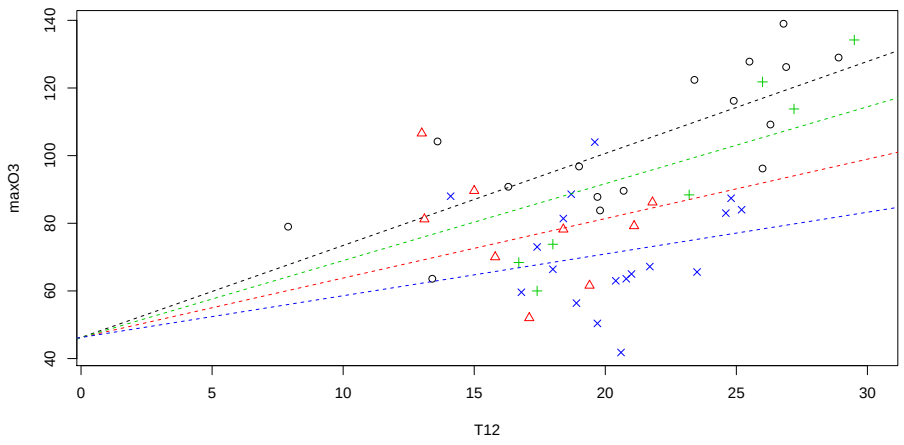
Si on peut considérer que les ordonnées à l'origine des I modèles de régression sont égales, on pose donc le modèle :

$$y_{ik} = \alpha + \gamma_i x_{ik} + e_{ik} \quad \forall k = 1, \ldots, n_i \quad i = 1, \ldots, I$$

On obtient alors I droites de régression de pentes différentes mais d'ordonnées à l'origine identiques.

Ce modèle a $I + 2$ paramètres.

Exemple de l'ozone :



6.4. Sélection du meilleur modèle

On va mettre en place des tests pour choisir le meilleur modèle. Pour cela on suppose que l'hypothèse gaussienne sur les erreurs est vérifiée ($e_{ik} \sim \mathcal{N}(0, \sigma^2)$).

On commence par comparer :

- le modèle complet ;
- le modèle avec pente commune.

Ces deux modèles sont emboîtés $\Rightarrow$ on utilise le test de Fisher pour modèles emboîtés.

Les hypothèses de test sont :

$$\begin{cases} H_0 : \gamma_1 = \ldots = \gamma_I = \gamma \quad \text{(modèle avec pente commune)} \\ H_1 : \exists (i,j) \quad \text{tel que} \quad \gamma_i \neq \gamma_j \quad \text{(modèle complet)} \end{cases}$$

Statistique de test pour le test de Fisher emboîté :

$$F = \frac{n-2I}{I-1} \frac{SCR_0 - SCR}{SCR}$$

et $F \underset{H_0}{\sim} \mathcal{F}_{(I-1,n-2I)}$

$\Rightarrow$ On rejette H_0 au niveau α si $F > f_{I-1,n-2I}(1-\alpha)$
où $f_{I-1,n-2I}(1-\alpha)$ est le quantile de niveau α de la loi de Fisher $(I-1, n-2I)$.

ou bien : on rejette H_0 si la p-value est plus petite que α

$\Rightarrow$ Dans ce cas on conserve le modèle complet.

On compare ensuite :

- le modèle complet ;
- le modèle avec ordonnée à l'origine commune.

Ces deux modèles sont emboîtés $\Rightarrow$ on utilise le test de Fisher pour modèles emboîtés.

Les hypothèses de test sont :

$$\begin{cases} H_0 : \alpha_1 = \ldots = \alpha_I = \alpha \quad \text{(modèle avec ordonnée à l'origine commune)} \\ H_1 : \exists (i,j) \quad \text{tel que} \quad \alpha_i \neq \alpha_j \quad \text{(modèle complet)} \end{cases}$$

Comme pour le test précédent, la statistique de test est :

$$F = \frac{n-2I}{I-1} \frac{SCR_0 - SCR}{SCR}$$

$\Rightarrow$ On rejette H_0 au niveau α si $F > f_{I-1,n-2I}(1-\alpha)$.

ou bien : on rejette H_0 si la p-value est plus petite que α

$\Rightarrow$ Dans ce cas on conserve le modèle complet.

A l'issue de ces deux tests, il y a plusieurs situations possibles :

1. Rejet de H_0 pour les deux tests ;
2. Rejet de H_0 pour un des deux tests ;
3. Non rejet de H_0 pour les deux tests.

☐ Cas 1 : On conserve le modèle complet à $2I + 1$ paramètres.

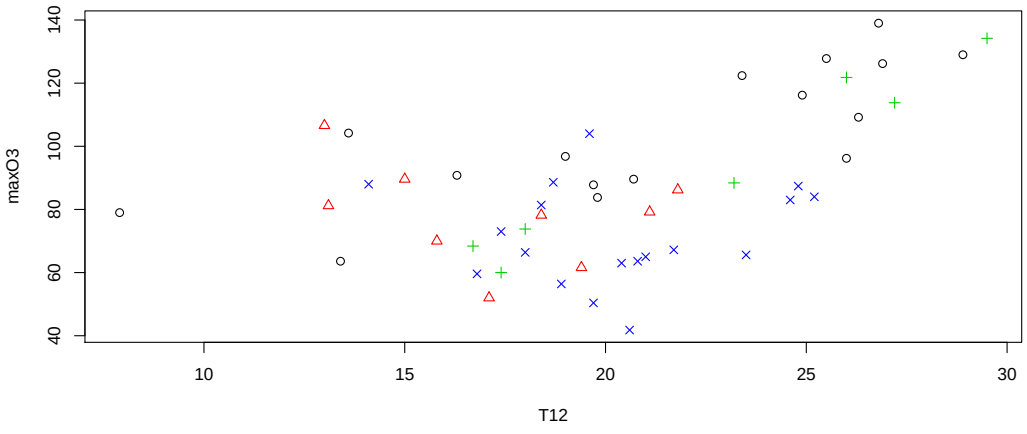
☐ Cas 2 : On conserve le modèle simplifié à $I + 2$ paramètres.

☐ Cas 3 : On doit choisir entre deux modèles simplifiés ayant chacun $I + 2$ paramètres.

$\Rightarrow$ On peut les comparer avec le critère du R^2.

6.5. Exemple

Exemple de l'ozone. Application avec le logiciel R.



Modèle d'analyse de la covariance complet :

```
mod1=lm(maxO3 ~ -1+vent+vent:T12,data=ozone)
summary(mod1)
```

Résultat :

```
Residuals:
    Min      1Q  Median      3Q     Max
-29.903  -9.163   1.153  10.319  32.638

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
ventEst       45.6090    13.9343   3.273 0.002133 **
ventNord     106.6345    28.0341   3.804 0.000456 ***
ventSud      -27.0602    26.5389  -1.020 0.313737
ventOuest     64.6840    24.6208   2.627 0.011967 *
ventEst:T12    2.7480     0.6342   4.333 8.96e-05 ***
ventNord:T12  -1.6491     1.6058  -1.027 0.310327
ventSud:T12    5.3786     1.1497   4.678 3.00e-05 ***
ventOuest:T12  0.3407     1.2047   0.283 0.778709
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.71 on 42 degrees of freedom
Multiple R-squared: 0.9773,     Adjusted R-squared: 0.973
F-statistic: 226.1 on 8 and 42 DF,  p-value: < 2.2e-16
```

$\Rightarrow \hat{\alpha}_1 = 45.6$, $\hat{\alpha}_2 = 106.6$, ... et $\hat{\gamma}_1 = 2.7$, $\hat{\gamma}_2 = -1.6$, ...

Premier modèle simplifié (pente commune) :

```
mod2=lm(maxO3 ~ -1+vent+T12,data=ozone)
summary(mod2)
```

Résultat :

```
Residuals:
    Min      1Q  Median      3Q     Max
-30.669 -11.115  -2.376  11.756  38.490

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
ventEst    52.3484    12.3370   4.243 0.000109 ***
ventNord   36.5193    10.9423   3.337 0.001704 **
ventSud    39.4935    13.8853   2.844 0.006675 **
ventOuest  22.4101    11.7611   1.905 0.063125 .
T12         2.4300     0.5476   4.438 5.82e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.74 on 45 degrees of freedom
Multiple R-squared: 0.9685,     Adjusted R-squared: 0.965
F-statistic: 276.7 on 5 and 45 DF,  p-value: < 2.2e-16
```

$\Rightarrow \hat{\alpha}_1 = 52.3$, $\hat{\alpha}_2 = 36.5$, ... et $\hat{\gamma} = 2.4$.

Deuxième modèle simplifié (ordonnée à l'origine commune) :

```
mod3=lm(maxO3 ~ vent:T12,data=ozone)
summary(mod3)
```

Résultat :

```
Residuals:
    Min      1Q  Median      3Q     Max
-29.865 -11.670  -0.747  11.776  37.520

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   46.2347    11.3300   4.081 0.000181 ***
ventEst:T12    2.7206     0.5306   5.128 6.01e-06 ***
ventNord:T12   1.7573     0.7102   2.474 0.017185 *
ventSud:T12    2.2739     0.5487   4.144 0.000148 ***
ventOuest:T12  1.2345     0.5799   2.129 0.038794 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.24 on 45 degrees of freedom
Multiple R-squared: 0.5762,     Adjusted R-squared: 0.5385
F-statistic: 15.29 on 4 and 45 DF,  p-value: 5.718e-08
```

$\Rightarrow \hat{\alpha} = 46.2$, $\hat{\gamma}_1 = 2.7$, $\hat{\gamma}_2 = 1.8$, ...

Comparaison des modèles 1 (complet) et 2 (pente commune) :

```
anova(mod2,mod1)
```

Résultat :

```
Analysis of Variance Table

Model 1: maxO3 ~ -1 + vent + T12
Model 2: maxO3 ~ -1 + vent + vent:T12
  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1     45 12612.0
2     42  9087.4  3    3524.5 5.4298 0.003011 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rejet de H_0 donc on préfère le modèle complet.

Comparaison des modèles 1 (complet) et 3 (ordonnée à l'origine commune) :

`anova(mod3,mod1)`

Résultat :

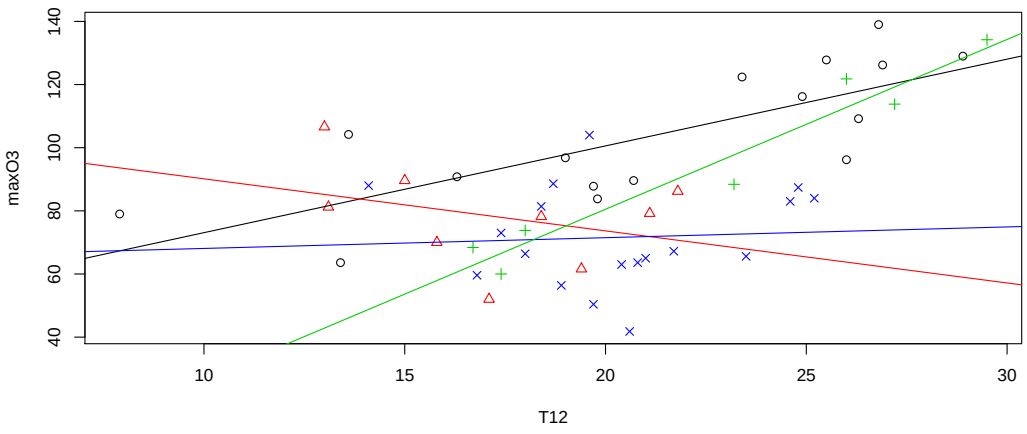
```
Analysis of Variance Table

Model 1: maxO3 ~ vent:T12
Model 2: maxO3 ~ -1 + vent + vent:T12
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     45 11864.1
2     42  9087.4  3    2776.6 4.2776 0.01008 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rejet de H_0 donc on préfère à nouveau le modèle complet.

Bilan : pour expliquer l'ozone par la température à 12h et le vent, on conserve donc le modèle complet avec une régression pour chaque modalité du vent, ayant I ordonnées à l'origine et pentes différentes.

$\Rightarrow$ Modèle estimé :



$\Rightarrow$ Le facteur vent apporte donc de l'information dans l'explication de la concentration en ozone.