Introduction à la Régression Linéaire

Olivier Martin

INRA Avignon, BioSP

Plan du cours

1. Cadre, rappels et objectifs
2. La régression linéaire simple
3. La régression linéaire multipe
4. Validation du modèle, analyse des résidus
5. Difficultés en régression multiple

Cadre et objectifs

On dispose de 2 caractères X et Y. On distingue deux objectifs :

1. On cherche à savoir s'il existe un lien entre X et Y
2. On cherche à savoir si X a une influence sur Y et éventuellement prédire Y à partir de X.

1. Liaison entre X et Y. On définit un indice de liaison : coeff. de corrélation, statistique du Khi-2,...
   Estimation : mesure de l'intensité de la liaison
   Test : Existence du lien
2. Infulence de x sur Y. On modélise l'influence de x sur Y : régression logistique, analyse de la variance, régression linéaire,...
   Estimation : description de l'influence et prédiction
   Test : validation d'hypothèse particulière : absence d'influence, inlfuence linéaire, quadratique,...

Cadre et objectifs

Définition : Pour deux variables X et Y, le coeff. de corrélation linéaire $r = \rho(X, Y)$ vaut :

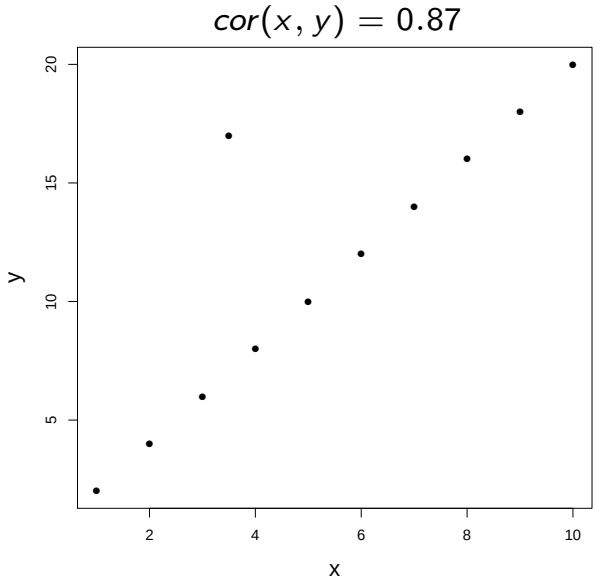
$$r = \rho(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \in [-1; 1]$$

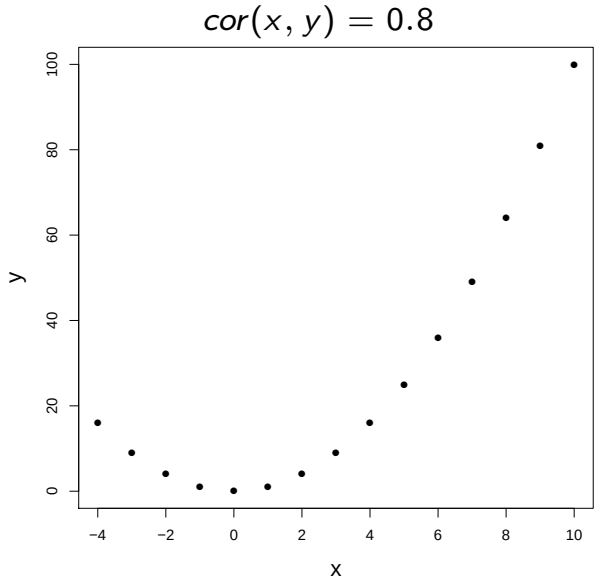
ρ est une mesure symétrique qui mesure le lien linéaire entre X et Y :
$\rho = -1$: X et Y sont proportionnels et varient en sens opposé
$\rho = 1$: X et Y sont proportionnels et varient dans le même sens
$\rho = 0$: X et Y ne sont pas corrélés
La corrélation n'indique aucune causalité.

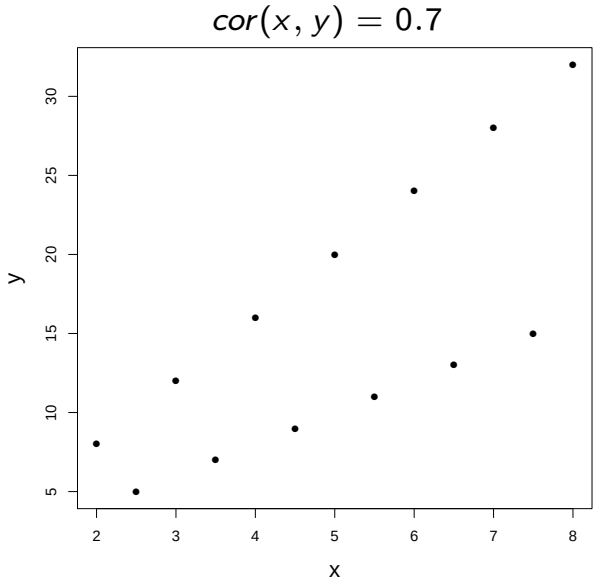
Propriétés :

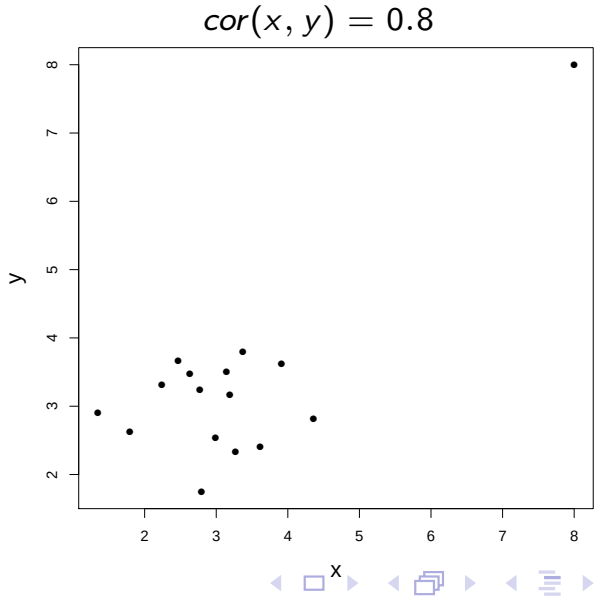
1. Si X et Y sont indépendants, alors $\rho(X, Y) = 0$.
2. Si X et Y sont gaussiens, il y a équivalence entre indépendance et corrélation nulle.

Cadre et objectifs









Cadre et objectifs

Rappels variance et covariance empirique :

$cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

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

$$r = \rho(X, Y) = \frac{cov(X, Y)}{S_X S_Y} \in [-1; 1]$$

Cadre et objectifs

Test sur le coef. de corrélation :

Dans le cas où X et Y sont supposés gaussiens, on peut réaliser un test sur le coeff de corrélation $\mathcal{H}_0 : r = 0$ contre $\mathcal{H}_1 : r \neq 0$. On utlise la statistique

$$\frac{R}{\sqrt{1-R^2}}\sqrt{n-2} \underset{\mathcal{H}_0}{\sim} \mathcal{T}_{n-2}$$

(R^2 est le coeff. de détermination, cf. plus loin dans le cours)

Corrélation partielle et multiple :

- On définit aussi la corrélation multiple entre une variables Y et p variables $X_1, ..., X_p$.
- Dans le cas de plusieurs variables, on définit également le coeff. de corrélation partiel pour s'assurer que la corrélation entre 2 variables n'est pas due en réalité aux variations d'une troisième variable.

La régression simple

Objectif : On souhaite expliquer les variations de la variable Y à partir des valeurs observées pour la variable x.
Le problème n'est pas symétrique : les 2 variables n'ont pas le même statut

Définition :
$Y =$ variable à expliquer ou réponse, supposée alétaoire
$x =$ variable explicative ou covariable ou régresseur, supposée fixe

Modèle :

$$Y = f(x) + E$$

où E est un terme résiduel alétoire ou erreur.

La régression simple

Données : On observe n individus ($i = 1, ..., n$).

Régression linéaire : On suppose que pour tout i :

$$Y_i = a + bx_i + E_i \text{ avec } \{E_i\} \text{ i.i.d et } \sim \mathcal{N}(0, \sigma^2).$$

Formulation équivalente : Les variables réponses $\{Y_i\}$ sont indépendantes de lois respectives

$$Y_i \sim \mathcal{N}(a + bx_i, \sigma^2)$$

Hypothèses du modèle statistique :

- L'espérance de Y_i dépend linéairement de x_i : $\mathbb{E}(Y_i) = a + bx_i$.
- La variance des Y_i est cste : $\mathbb{V}(Y_i) = \mathbb{V}(E_i) = \sigma^2$.
- Les réponses et termes résiduels sont gaussiens **et** indépendants

La régression simple

Paramètres à estimer : a, b, σ^2

Deux approches : maximum de vraisemblance ou moindres carrés

L'estimation par max. de vraisemblance ou moindres carrés sont deux méthodes classiques pour l'estimation. Ici, les estimations sont (quasi) identiques.

Dans les 2 cas, on définit un critère qui mesure l'adéquation entre les paramètres du modèle et les données observées. On cherche alors les paramètres qui maximisent ou minimisent ce critère.

La régression simple

Le critère du maximum de vraisemblance

La densité pour la réponse Y_i est :

$$f(y_i; a, b, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(y_i - a - bx_i)^2}{2\sigma^2}\right]$$

Les données étant supposées indépendantes, la log-vraisemblance vaut :

$$\mathcal{L}(a, b, \sigma^2) = -\frac{n}{2} log(2\pi) - \frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2} \sum_i (Y_i - a - bx_i)^2$$

Les estimateurs du max. de vraisemblance de a, b, σ^2 sont les valeurs qui maximisent $\mathcal{L}(a, b, \sigma^2)$. Les estimateurs sont obtenus à partir des réalisations y_i : ce sont des **variables alétoires qui possèdent une loi**.

La régression simple

Le critère du maximum des moindres carrés (SCR)

On cherche les valeurs de a et b qui minimisent la somme des carrés des résidus, i.e. les écarts entre les observations (Y_i) et les prédictions $(a + bx_i)$ du modèle.

$$SCR(a, b) = \sum_i (Y_i - (a + bx_i))^2$$

On remarque que ce critère apparaît dans la log-vraisemblance...

Propriété :

Les critères du max. de vraisemblance et des moindres carrés donnent les même estimateurs pour a et b. Le critère des moindres carrés n'utilise pas l'hypothèse de distribution gaussienne des erreurs.

La régression simple

Les estimateurs A et B de a et b

$$A = \bar{Y} - B\bar{x} \text{ et } B = \frac{\sum_i (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Les estimations $\hat{a}$ et $\hat{b}$ de a et b

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \text{ et } \hat{b} = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{cov(x,y)}{\sigma_x^2}$$

L'estimateur de la variance σ^2 est donné par :

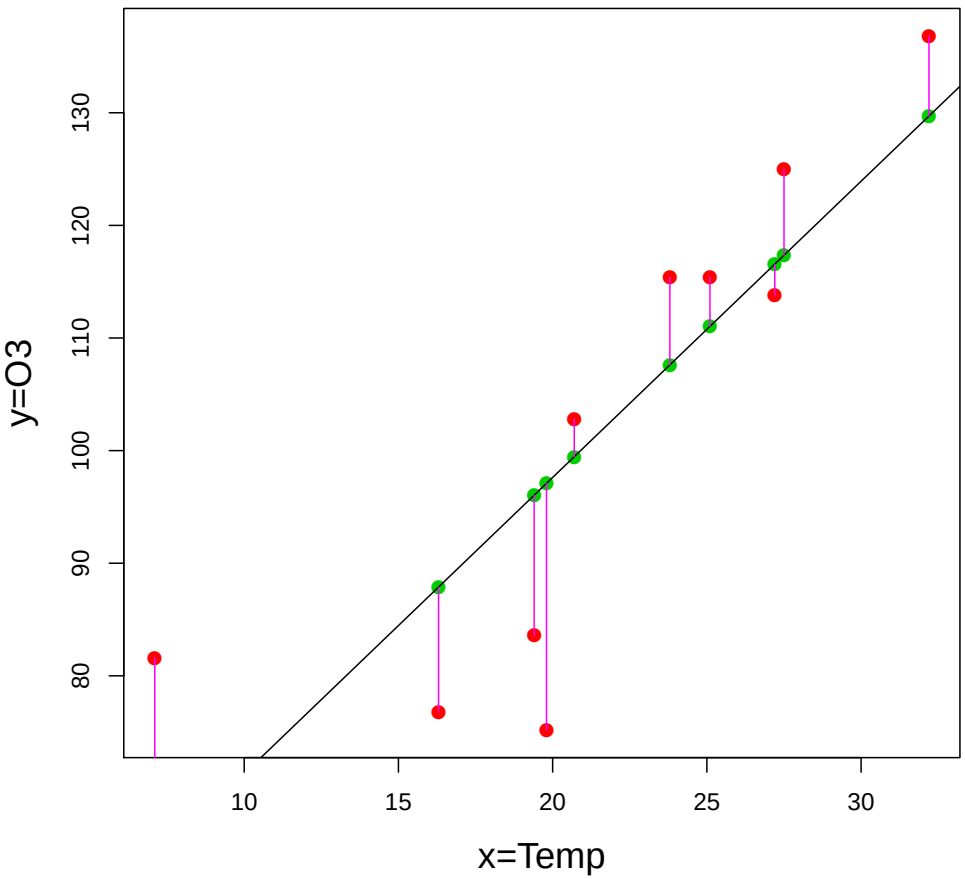
$$S_{n-2}^2 = \frac{1}{n-2} \sum_i (Y_i - A - Bx_i)^2$$

L'estimation $\hat{\sigma}^2$ de la variance σ^2 est :

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_i (y_i - (\hat{a} + \hat{b}x_i))^2 = \frac{1}{n-2} \sum_i \hat{\epsilon}_i^2$$

La régression simple

Temp.	23.8	16.3	27.2	7.1	25.1	27.5	19.4	19.8	32.2	20.7
O3	115.4	76.8	113.8	81.6	115.4	125.0	83.6	75.2	136.8	102.8



cor(x,y)=0.839

les données (x_i, y_i)
$\hat{a}$ et $\hat{b}$ les estimations

Prédiction : $\hat{a} + \hat{b}x_i$
Droite de régression : $\hat{a} + \hat{b}x$
Erreurs : $\hat{e}_i = y_i - \hat{a} - \hat{b}x_i$

La régression simple

Les estimateurs A, B et S^2_{n-2} sont des variables aléatoires.

En utlisant l'hypothèse de loi gaussienne sur les erreurs E_i, on obtient les lois de ces estimateurs.

On peut alors réaliser des tests sur les paramètres, définir des intervalles de confiance, des intervalles de prédiction, comparer des modèles de régression,...

La régression simple

Moments des estimateurs :

A, B et S^2_{n-2} sont des estimateurs sans biais : $\mathbb{E}(A) = a$, $\mathbb{E}(B) = b$ et de variance :

$$\mathbb{V}(A) = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}] \text{ et } \mathbb{V}(B) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}.$$

Comme σ^2 est inconnu, on obtient des estimations de ces variances en remplaçant σ^2 par $\hat{\sigma}^2$.

Loi des estimateurs :

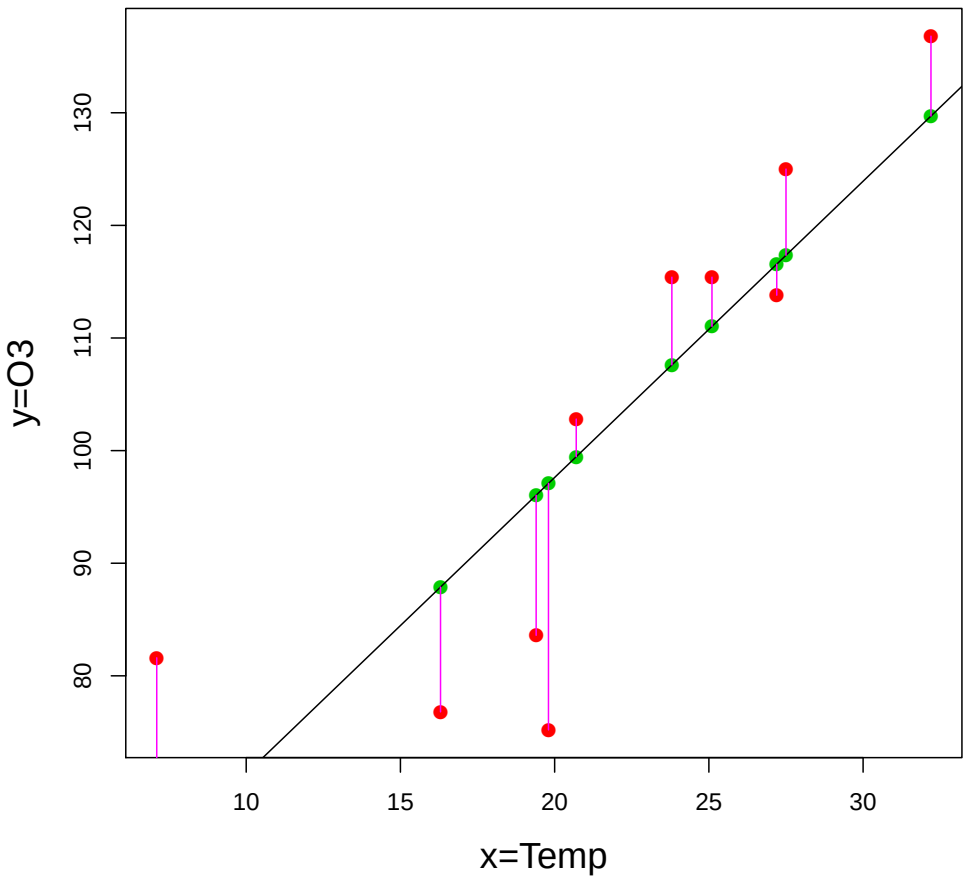
$A \sim \mathcal{N}(a, \mathbb{V}(A))$

$B \sim \mathcal{N}(b, \mathbb{V}(B))$

$\frac{(n-2)S^2_{n-2}}{\sigma^2} \sim \chi^2_{n-2}$

La régression simple

Temp.	23.8	16.3	27.2	7.1	25.1	27.5	19.4	19.8	32.2	20.7
O3	115.4	76.8	113.8	81.6	115.4	125.0	83.6	75.2	136.8	102.8



$$\hat{a} = 45 \text{ et } \sqrt{\hat{\mathbb{V}}(A)} = 13.805$$

$$\hat{b} = 2.63 \text{ et } \sqrt{\hat{\mathbb{V}}(B)} = 0.602$$

$$\hat{\sigma^2} = 160.64 \text{ et } \hat{\sigma} = 12.67$$

La régression simple

Résultat de la régression avec *lm()* sous R

```
>summary(lm(O3~Tp))
Call:
lm(formula=O3~Tp)

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 *
Tp          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
```

La régression simple

Tests sur les paramètres

On veut tester $\mathcal{H}_0 : b = 0$ contre $\mathcal{H}_1 : b \neq 0$

Loi de B

$$B \sim \mathcal{N}(b, \mathbb{V}(B)) \Rightarrow \frac{B-b}{\sqrt{\mathbb{V}(B)}} \sim \mathcal{N}(0,1) \Rightarrow \frac{B-b}{\sqrt{\hat{\mathbb{V}}(B)}} \sim \mathcal{T}_{n-2}$$

Statistique de test sous $\mathcal{H}_0$

$$T = \frac{B}{\sqrt{\hat{\mathbb{V}}(B)}} \underset{\mathcal{H}_0}{\sim} \mathcal{T}_{n-2}$$

Calcul de la p-valeur

$$p - value = 2 * P(\mathcal{T}_{n-2} > |\frac{\hat{b}}{\sqrt{\hat{\mathbb{V}}(B)}}|) = 2 * P(\mathcal{T}_{n-2} < -|\frac{\hat{b}}{\sqrt{\hat{\mathbb{V}}(B)}}|)$$

La régression simple

Test $\mathcal{H}_0 : b = 0$ contre $\mathcal{H}_1 : b \neq 0$

$\hat{b} = 2.63$, $\sqrt{\hat{\mathbb{V}}(B)} = 0.603$ donc $t = \frac{2.63}{0.603} = 4.36$

$$p - value = 2 * P(\mathcal{T}_{n-2} > |4.36|) = 0.0024$$

Test $\mathcal{H}_0 : a = 0$ contre $\mathcal{H}_1 : a \neq 0$

$\hat{a} = 45.00$, $\sqrt{\hat{\mathbb{V}}(A)} = 13.805$ donc $t = \frac{45}{13.805} = 3.260$

$$p - value = 2 * P(\mathcal{T}_{n-2} > |0.011|) = 0.0011$$

La régression simple

Résultat de la régression avec *lm()* sous R

```
>summary(lm(O3~Tp))
Call:
lm(formula=O3~Tp)

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 *
Tp          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
```

La régression simple

1. L'ajustement du modèle calculé pour une covariable x_0 est

$$\hat{Y}_0 = A + Bx_0.$$

$\hat{Y}_0$ est une variable gaussienne telle que :

$$\mathbb{E}(\hat{Y}_0) = a + bx_0 \text{ et } \mathbb{V}(\hat{Y}_0) = \sigma^2[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}]$$

2. Prédiction

On peut prédire la réponse Y_0 pour une valeur x_0 de la covariable :

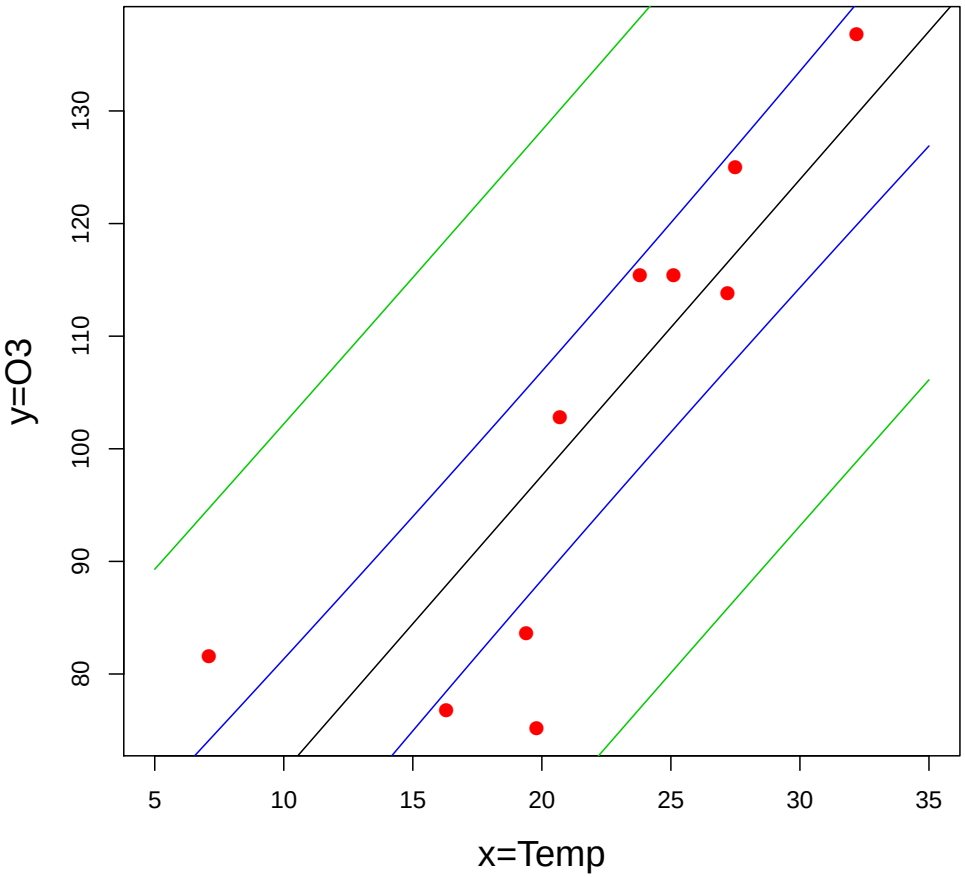
$$T_0 = A + Bx_0 + E_0$$

T_0 est une variable gaussienne telle que :

$\mathbb{E}(T_0) = a + bx_0$ et

$\mathbb{V}(T_0) = \mathbb{V}(\hat{Y}_0) + \sigma^2 = \sigma^2[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_i (x_i-\bar{x})^2} + 1]$

La régression simple



Ajustement : $\hat{y}_0 = \hat{a} + \hat{b}x_0$

Intervalle de confiance :
$[\hat{y}_0 \pm t_{n-2,1-\alpha/2}\sqrt{\mathbb{V}(\hat{Y}_0)}]$

Intervalle de prédiction :
$[t_0 \pm t_{n-2,1-\alpha/2}\sqrt{\mathbb{V}(T_0)}]$

La régression simple

Le coefficient d'ajustement ou de détermination R^2

Somme des carrés totale	$SCT = \sum_i (Y_i - \bar{Y})^2$	variabilité totale à expliquer
Somme des carrés due au modèle	$SCM = \sum_i (\hat{Y}_i - \bar{\hat{Y}})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2$	variabilité expliquée par le modèle
Somme des carrés résiduelle	$SCR = \sum_i (Y_i - \hat{Y}_i)^2$	variabilité non expliquée par le modèle

Formule d'analyse de variance : $SCT = SCM + SCR$

Coefficient d'ajustement R^2

Le R^2 mesure la part de variabilité expliquée par le modèle :

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

Remarque

Un bon ajustement linéaire implique un R^2 proche de 1 (attention, la réciproque est fausse). On montre la relation $R^2 = \rho^2$.

La régression simple

$$SCT = \sum_i (Y_i - \bar{Y})^2 = 4342.944$$
$$SCM = \sum_i (\hat{Y}_i - \bar{Y})^2 = 3057.806 \quad R^2 = 3057.806/4342.944 \approx 0.704$$
$$SCR = \sum_i (Y_i - \hat{Y}_i)^2 = 1285.138$$

On peut réaliser un test $\mathcal{H}_0 : b = 0$ contre $\mathcal{H}_0 : b \neq 0$ en utilisant la statistique

$$\frac{SCM}{SCR}(n-2) \underset{\mathcal{H}_0}{\sim} \mathcal{F}(1, n-2)$$

En fait, $\frac{SCM}{SCR} = \frac{R^2}{1-R^2}$. On retrouve donc le test sur le coeff. de corrélation définie au début et le fait que $(\mathcal{T}_{n-2})^2 = \mathcal{F}(1, n-2)$. Tester $\rho = 0$ ou $b = 0$ est en effet équivalent : pas de lien de linéarité.

La régression simple

Pour les données :

$\frac{R^2}{1-R^2}(n-2) = \frac{0.704}{1-0.704}(10-8) = 19.027$
et
$P(\mathcal{F}(1,8) > 19.027) = 0.0024$

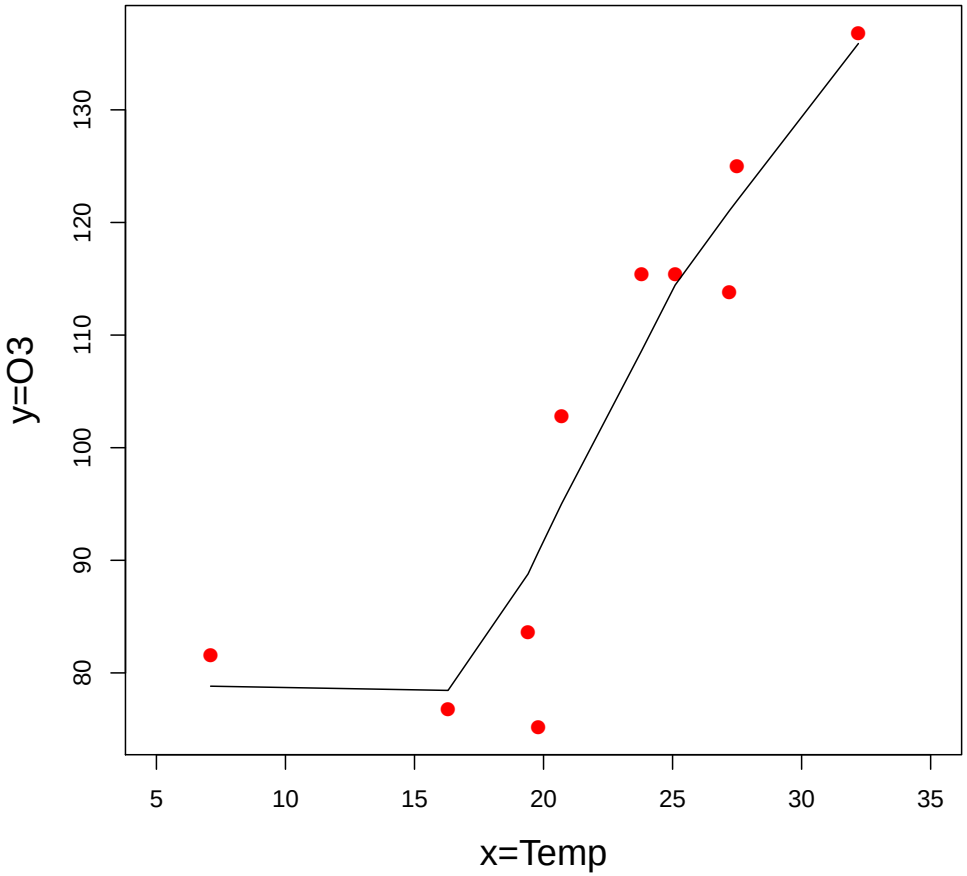
On a également $\rho^2 = 0.839^2 = 0.704 = R^2$.

La régression simple

Résultat de la régression avec *lm()* sous R

```
>summary(lm(O3~Tp))
Call:
lm(formula=O3~Tp)

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 *
Tp          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
```

La régression multiple



La régression multiple

Régression quadratique

Le lien entre Y et la covariable est peut-être une fonction plus complexe.

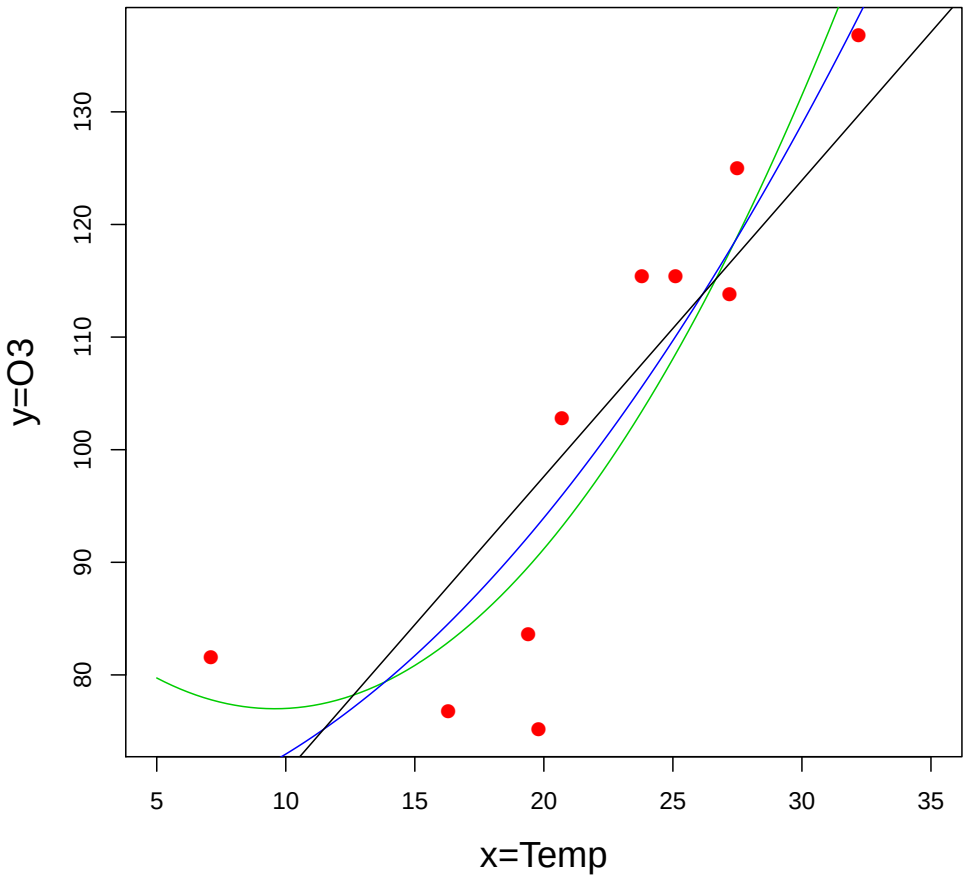
Exemples :

$\mathcal{M}_1 : Y_i = a + bx_i + cx_i^2 + E_i$ avec $\{E_i\}$ i.i.d. $\mathcal{N}(0, \sigma^2)$
$\mathcal{M}_2 : Y_i = a + bx_i^2 + E_i$ avec $\{E_i\}$ i.i.d. $\mathcal{N}(0, \sigma^2)$
$\mathcal{M}_3 : Y_i = a + bx_i + E_i$ avec $\{E_i\}$ i.i.d. $\mathcal{N}(0, \sigma^2)$

Remarque :

Ces modèles sont tous des modèles linéaires. Le modèle $\mathcal{M}_1$ est un modèle de régression multiple (plus de une covariable dans le modèle).
Les modèles $\mathcal{M}_2$ et $\mathcal{M}_3$ sont deux modèles emboités (cas particuliers) de $\mathcal{M}_1$.

La régression multiple



$$\mathcal{M}_1 : Y_i = a + bx_i + cx_i^2 + E_i$$

$$\mathcal{M}_2 : Y_i = a + bx_i^2 + E_i$$

$$\mathcal{M}_3 : Y_i = a + bx_i + E_i$$

La régression multiple

Cas de plusieurs covariables $x_1, ..., x_p$ avec $p < n$

Dans le cas de plusieurs variables, la première analyse consiste à faire des analyses descriptives des covariables. On peut utiliser par exemple les commandes *pairs()* et *boxplot()* sous R.

Modèle de régression $x_1, ..., x_p$

On pose le modèle : $Y_i = a_0 + a_1 x_{1,i} + ... + a_p x_{p,i} + E_i$ avec $E_i \sim \mathcal{N}(0, \sigma^2)$

De même que dans le cas du modéle linéaire simple :

- On peut estimer les paramètres $a_0, ..., a_p$ et σ^2
- Faire des tests sur les paramètres
- Calculer le R^2
- Faire un test sur le coef. de corrélation
- ...

La régression multiple

Analyse de variance de la régression multiple

On souhaite tester l'hypothèse de non-régression, i.e.
$\mathcal{H}_0 : a_1 = ... = a_p = 0$ contre $\mathcal{H}_1$: au moins un $a_i \neq 0$

On a alors la propriété suivante :

$$\frac{R^2}{1-R^2}\frac{n-p-1}{p} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(p, n-p-1)$$

Pour le cas $p = 1$, on retrouve bien le cas de la régression simple.

La régression multiple

Comparaison de modèles de régression

Attention, plus le nombre de variables sera grand et plus le R^2 sera grand. On définit le R^2 ajusté qui prend en compte le nombre de covariables.

Il existe également des critères numériques tel que AIC (An information criteria ou critère de Akaike) pour sélectionner des modèles. Ce critère est adapté pour un nombre pas trop important de covariables (< 20)

Le R^2 et le R^2 ajusté ne sont **surtout pas** les seuls critères à regarder pour comparer des modèles. L'analyse des résidus, des points extrêmes ou aberrants est tout aussi importante.

La régression multiple

Test pour la comparaison de modèles de régression

On souhaite comparer 2 modèles **emboités** : $\mathcal{M}_q$ avec q covariables et $\mathcal{M}_p$ avec p covariables (et la cste fait partie des 2 modèles).

Pour effectuer cette comparaion, on pose le test :

$$\mathcal{H}_0 : \text{le "bon" modèle est } \mathcal{M}_q$$
$$\mathcal{H}_1 : \text{le "bon" modèle est } \mathcal{M}_p$$

avec la condition : $q < p$

La régression multiple

Modèle 1 : $03 \sim Tp$
Modèle 2 : $03 \sim Tp + Tp^2$

```
> anova(lm(O3~Tp),lm(O3~Tp+I(Tp^2)))
Analysis of Variance Table
Model 1: O3 ~ Tp
Model 2: O3 ~ Tp + I(Tp^2)
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1      8 1285.14
2      7  711.18  1    573.96 5.6494 0.04911 *
```

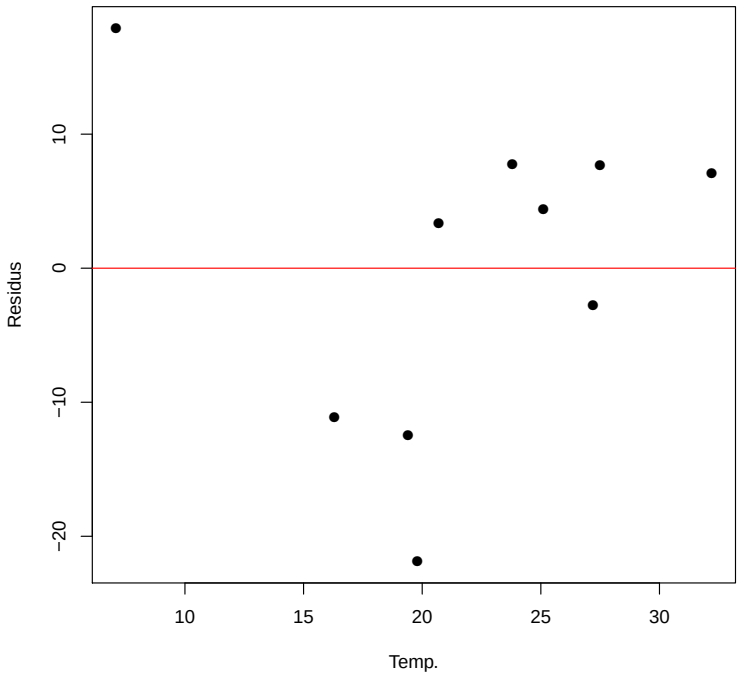
```
> anova(lm(O3[-4]~Tp[-4]),lm(O3[-4]~Tp[-4]+I(Tp[-4]^2)))
Analysis of Variance Table

Model 1: O3[-4] ~ Tp[-4]
Model 2: O3[-4] ~ Tp[-4] + I(Tp[-4]^2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      7 489.68
2      6 455.19  1    34.491 0.4546 0.5253
```

Validation du modèle

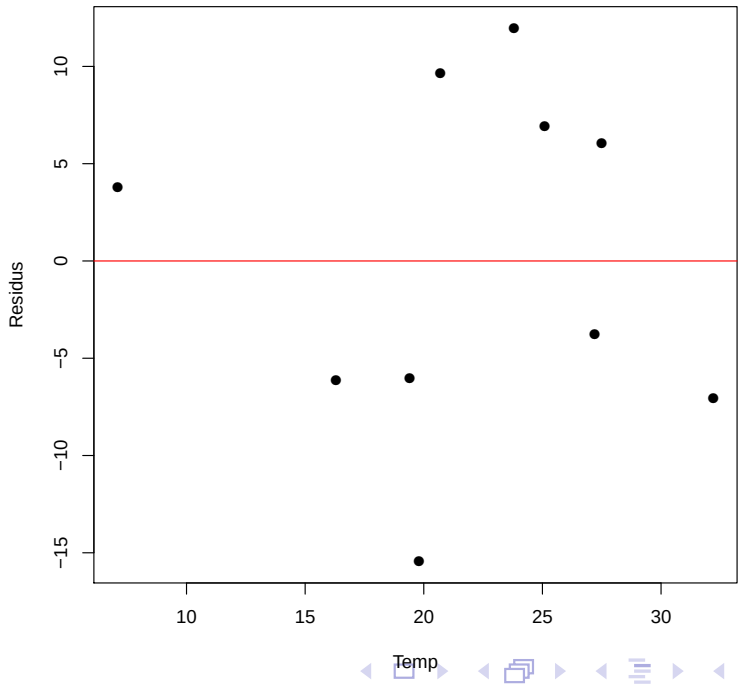
L'analyse des résidus

On estime l'erreur de l'ajustement par le résidu $Y_i - \hat{Y}_i$. On représente en abscisse x_i et en ordonnée $E_i = Y_i - \hat{Y}_i$. On peut également placer y_i en abscisse et $E_i = Y_i - \hat{Y}_i$ en ordonnée. Le graphique ne doit montrer **aucune structure particulière**.

Modele 1

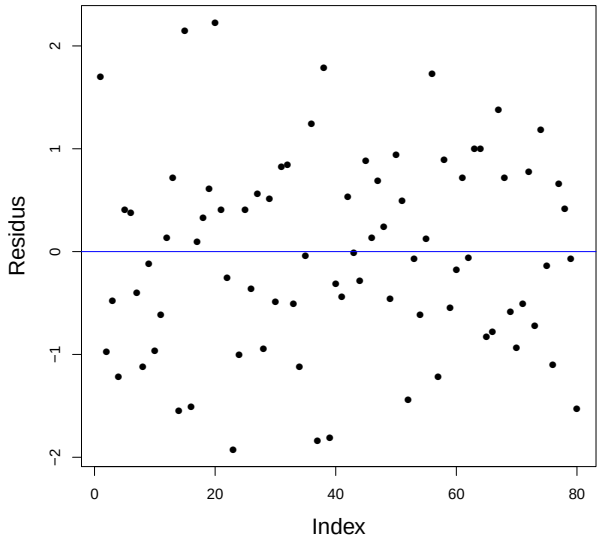


Modele 2

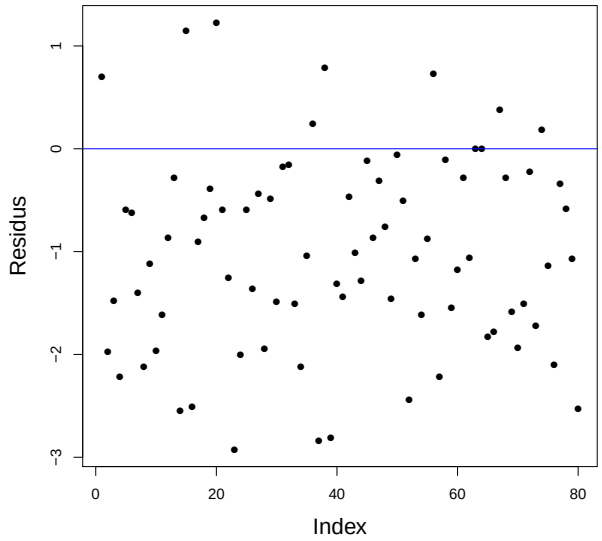


Validation du modèle

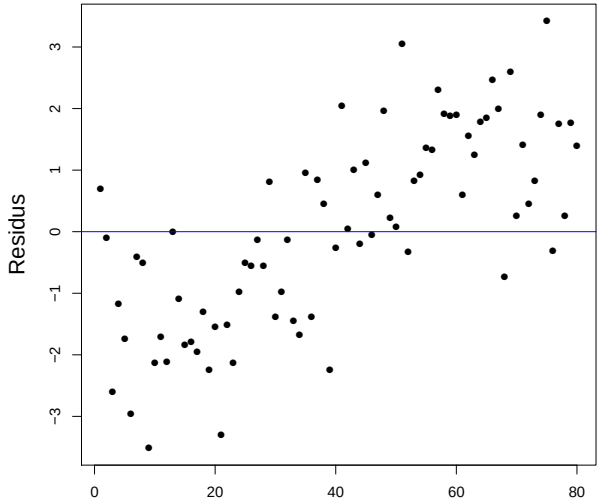
Hypothèses vérifiées



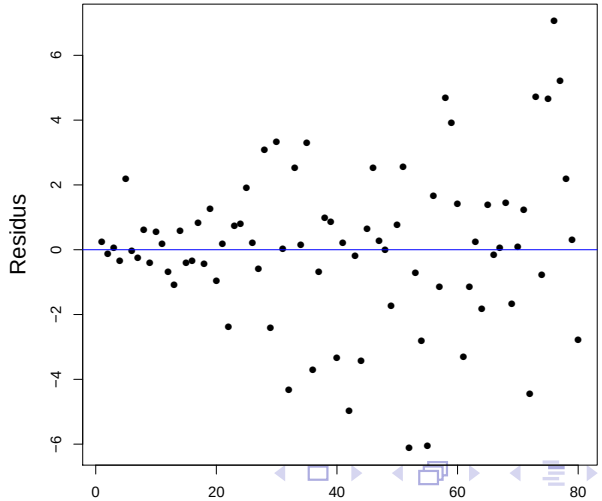
Hyp. gaussienne fausse



Linéarité fausse



Variance non-homogène



Validation du modèle

Hypothèse de variance homogène des résidus

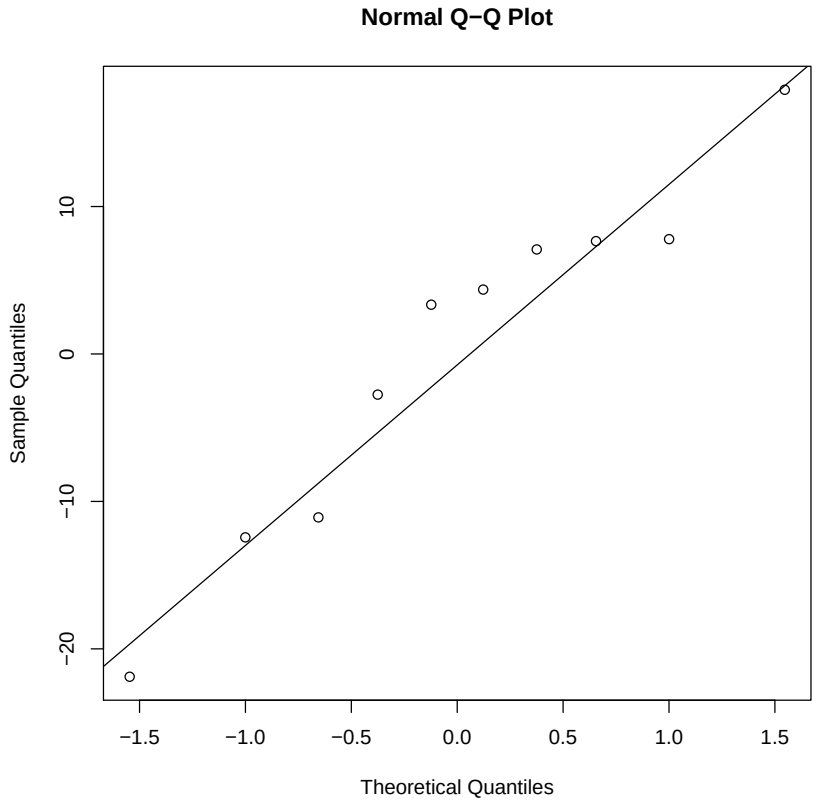
Lorsque une analyse des résidus permet d'identifier une variance non-homogène, on peut réaliser une transformation des variables Y ou x_i.

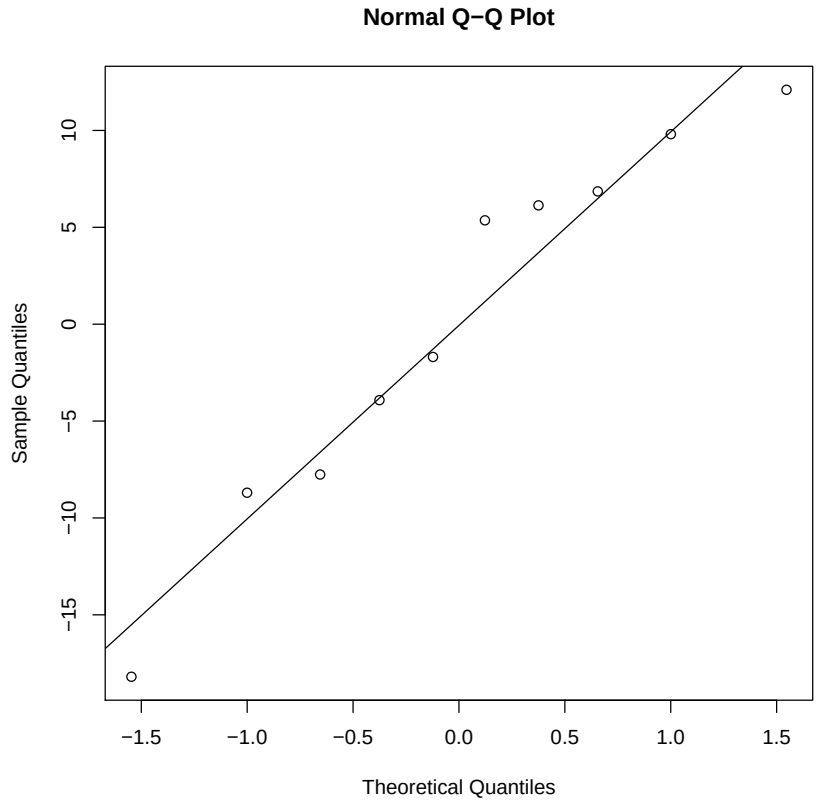
Les transformations classiques sont la transformation $\sqrt{.}$ ou la transformation log(.).

Validation du modèle

Normalité des résidus

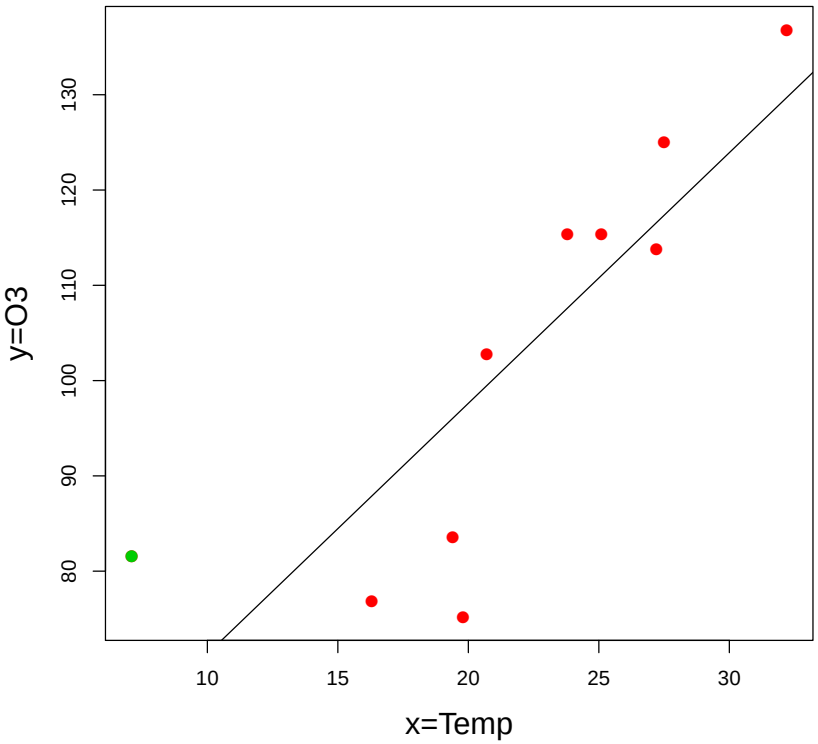
```
res=lm(03~Tp) ; qqnorm(res$residuals) ; qqline(res$residuals)
res2=lm(03~Tp^2) ; qqnorm(res2$residuals) ; qqline(res2$residuals)
```

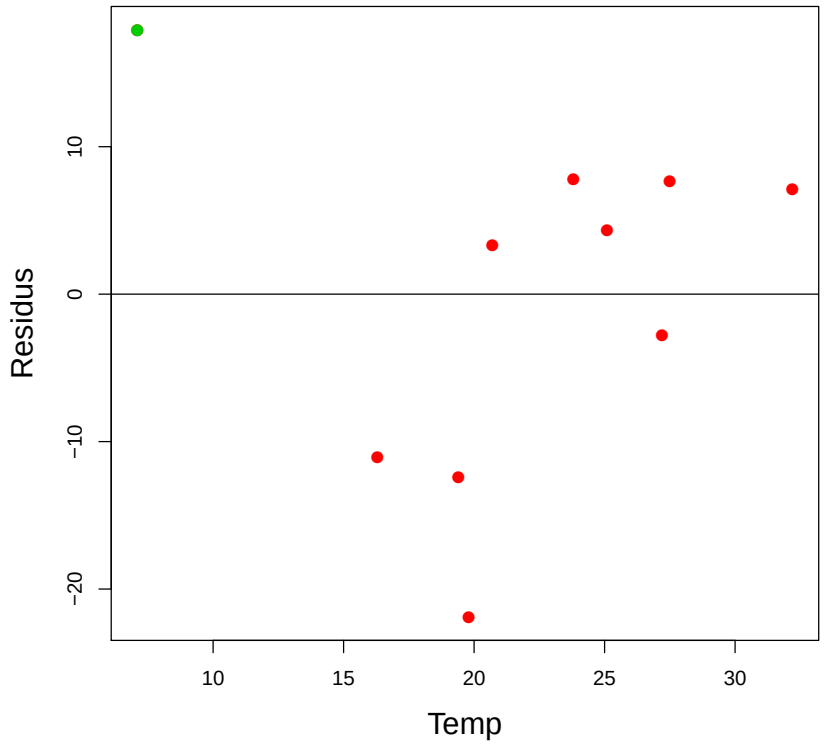




Validation du modèle

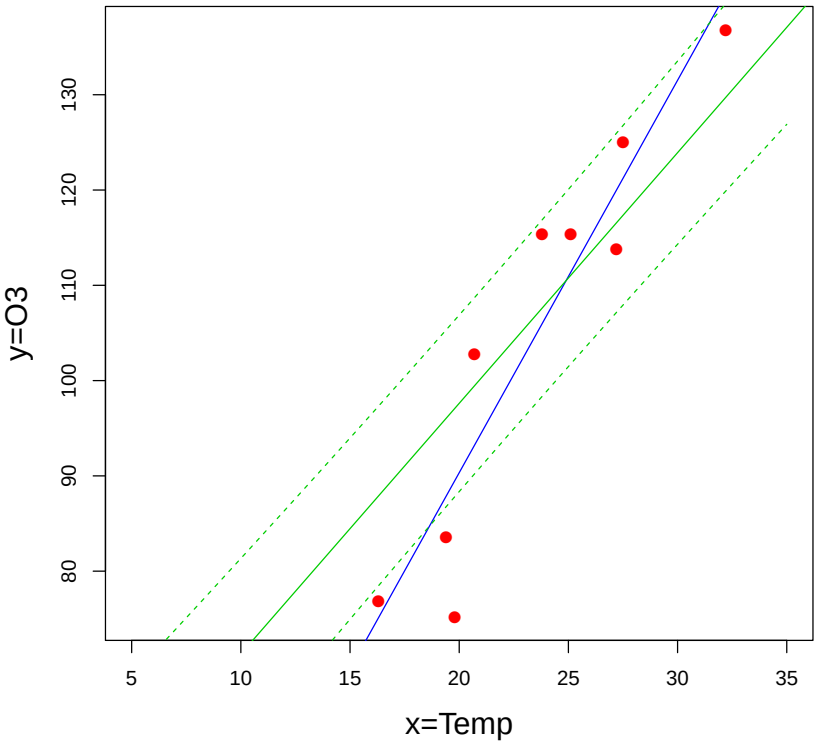
Influence de certains points

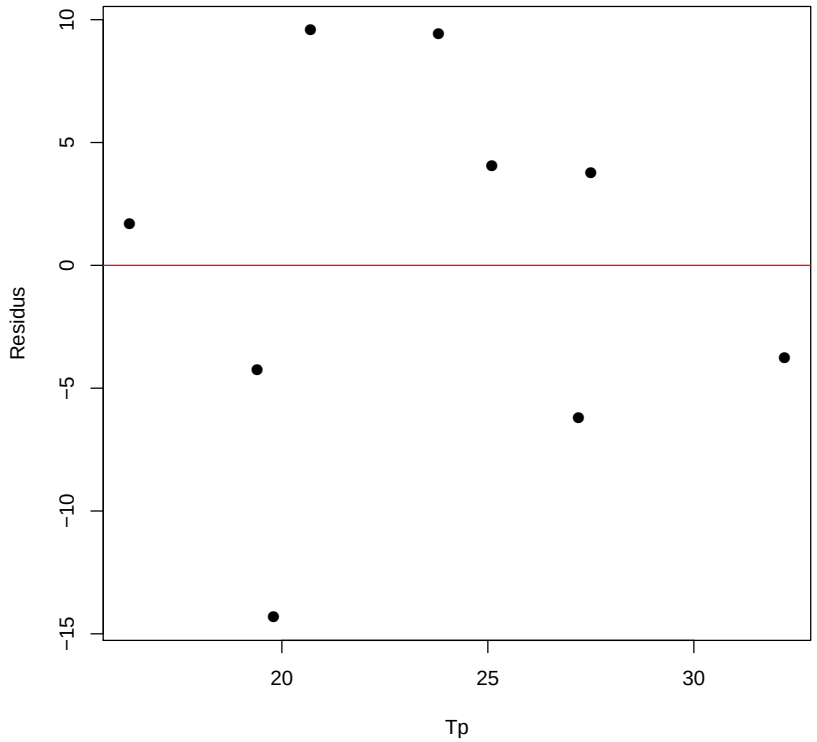




Validation du modèle

Influence de certains points





Validation du modèle

Influence de certains points

```
> summary(lm(O3~Tp))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.0044    13.8050   3.260   0.0115 *
Tp            2.6306     0.6029   4.363   0.0024 **
---
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


> summary(lm(O3[-4]~Tp[-4]))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.9669    14.2699   0.558 0.594039
Tp[-4]        4.1184     0.5941   6.932 0.000225 ***
---
Residual standard error: 8.364 on 7 degrees of freedom
Multiple R-squared: 0.8728,Adjusted R-squared: 0.8547
F-statistic: 48.05 on 1 and 7 DF,  p-value: 0.0002248
```

Difficulté en régression multiple

Deux points doivent être abordés avec attention :

1. Les échelles des covariales (vraie aussi en régression simple)
   Il est souvent judicieux de ramener toutes les variables à une moyenne nulle (centrage) et les variances empiriques de chacune des variables à 1 (centrer et réduire) : utiliser la commande *boxplot()* pour analyser ces différences d'échelle.
2. La corrélation entre les variables explicatives
   Les corrélations entre variables peuvent induire de mauvaises interprétations. Pour cela, on utilise parfois les axes d'une analyse en composantes principales (ACP) comme variables explicatives.
   L'inconvénient de cette approche est qu'il est alors nécessaire de donner le lien entre les co-variables de départ et les axes de l'ACP.
   On peut aussi utiliser la commande *pairs()* pour une première analyse et calculer les corrélations entre les covariables.

Difficulté en régression multiple

```
> summary(lm(O3~Tp+I(Tp^2)))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.96445   21.50810   4.136  0.00437 **
Tp          -2.50001    2.21118  -1.131  0.29546
I(Tp^2)      0.13057    0.05493   2.377  0.04911 *
---
Residual standard error: 10.08 on 7 degrees of freedom
Multiple R-squared: 0.8362, Adjusted R-squared: 0.7895
F-statistic: 17.87 on 2 and 7 DF,  p-value: 0.001777


> summary(lm(O3[-4]~Tp[-4]+I(Tp[-4]^2)))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -39.94750   72.59881  -0.550    0.602
Tp[-4]        8.24747    6.15501   1.340    0.229
I(Tp[-4]^2)  -0.08554    0.12687  -0.674    0.525
---
Residual standard error: 8.71 on 6 degrees of freedom
Multiple R-squared: 0.8818, Adjusted R-squared: 0.8424
F-statistic: 22.38 on 2 and 6 DF,  p-value: 0.001651
```

Difficulté en régression multiple

La corrélation partielle

Le coefficient de corrélation partielle mesure la liaison entre 2 variables lorsque l'influence d'une troisième (ou de plusieurs autres) est gardée constante sur les 2 variables comparées. Il a le même sens que le coefficient de corrélation classique.

$$\rho_{y,x_1|x_2} = \frac{\rho_{y,x_1} - \rho_{y,x_1}\rho_{y,x_2}}{\sqrt{1-\rho^2_{y,x_1}}\sqrt{1-\rho^2_{y,x_2}}}$$

Lien entre corrélation partielle et corrélation multiple

$$\rho_{y,x_{p+1}|x_1,\ldots,x_p} = \frac{R^2_{p+1} - R^2_p}{1 - R^2_p}$$

Le carré de la corrélation partielle, donne donc l'augmentation de R^2 relative á la portion de la variation de y inexpliquée par les variables $x_1, \ldots, x_p$ déjà dans l'équation.

Difficulté en régression multiple

La corrélation partielle

Pour le jeu de données, on obtient :

$cor(O3, Tp|Tp^2) = 0.09$ et $cor(O3, Tp^2|Tp) = 0.12$

La régression multiple

Les tests sur les paramètres dans le cadre de la régression multiple doivent être utilisés avec précaution en raison des possibles coorélations entre les variables.
On pourrait éliminer des variables du modèle sur la base de ces tests (les variables aux coefficients significativement nuls). Mais cette procédure est incorrecte. Il ne faut pas oublier que le test d'un coefficient est effectué alors que les autres variables sont fixées. Donc si deux variables sont trés corrélées, le test d'un des deux coefficients peut être non significatif puisque l'information apportée par la variable testée existe dans l'autre. On ne peut donc rien conclure sur l'estimation de ces coefficients et de leurs significativité.
La question de la sélèction des variables doit faire l'objet d'une analyse approfondie basée par exemple sur l'analyse des coeff. de corrélation partielle.

La régression multiple

La sélèction de variables ou choix de modèle peut égaement se réaliser grâce aux critères numériques tels que le AIC ou BIC.

Avec ces critères, on choisit le modèle dont le critère est le plus petit. Ces critères ne sont pas basés sur des tets, mais sur la pénalisation de la vraisemblance en fonction du nombre de paramètres dans le modèle.

Le package MuMIn (MultiModelInference) est intéressant car il calcule les AIC pour tous les modèles possibles, puis les range en fonction de la valeur du AIC.

Pour conclure ...

La régression soulève encore d'autres questions, comme la sélection des variables, les transformations des co-variables, le choix de modèle, la régression généralisée ou les effets aléatoires.

Il existe d'autres méthodes pour modéliser des observations comme la régression sur variables d'ACP, la régression PLS ou les Random Forest.

Pour plus d'informations :

- The R book, Michael J. Crawley
- Applied regression analysis, Drapper & Smith
- Probabilités, analyses des données et statistiques, Saporta.
- The elements of statistical learning, Hastie T., Tibshirani R., Friedman J.