La régression multiple

François Husson

Unité pédagogique de mathématiques appliquées
Agrocampus Ouest

Données, problématique

L'association de surveillance de la qualité de l'air Air Breizh mesure la concentration de polluants comme l'ozone (O_3) ainsi que les conditions météorologiques comme la température, la nébulosité, le vent, etc. Leur objectif est de prévoir la concentration en ozone pour le lendemain afin d'avertir la population en cas de pic de pollution.

Nous souhaitons analyser ici la relation entre le maximum journalier de la concentration en ozone (en $\mu g/m^3$) et les données météorologiques. Nous disposons de 112 données relevées durant l'été 2001 à Rennes.

	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v
2001-06-01	87	15.6	18.5	18.4	4	4	8	0.69	-1.71	-0.69	84
2001-06-02	82	17.0	18.4	17.7	5	5	7	-4.33	-4.00	-3.00	87
2001-06-03	92	15.3	17.6	19.5	2	5	4	2.95	1.88	0.52	82
2001-06-04	114	16.2	19.7	22.5	1	1	0	0.98	0.35	-0.17	92
2001-06-05	94	17.4	20.5	20.4	8	8	7	-0.50	-2.95	-4.33	114
2001-06-06	80	17.7	19.8	18.3	6	6	7	-5.64	-5.00	-6.00	94
2001-06-07	79	16.8	15.6	14.9	7	8	8	-4.33	-1.88	-3.76	80
...	...	...									

Peut-on prévoir le taux d'ozone du lendemain ?

Problématique

Exemples :

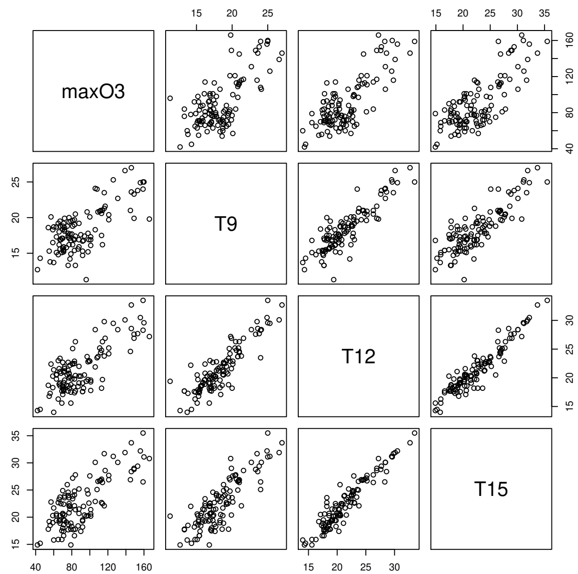
- Prévision de l'appréciation d'un produit en fonction de sa composition
- Optimisation d'une réaction chimique en fonction du temps de réaction et de la température
- ...

Objectifs :

- Expliquer une variable quantitative Y en fonction de p variables quantitatives $x_1, \ldots, x_p$
- Prédire de nouvelles valeurs pour Y

Analyse exploratoire : outils graphiques

```
pairs(ozone[,1:4])
```



Rappel régression simple

$$\begin{cases} \forall i = 1, \dots, n \quad Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \\ \forall i = 1, \dots, n \quad \varepsilon_i \text{ i.i.d. }, \quad \mathbb{E}(\varepsilon_i) = 0, \quad \mathbb{V}(\varepsilon_i) = \sigma^2 \\ \forall i \neq k \quad cov(\varepsilon_i, \varepsilon_k) = 0 \end{cases}$$

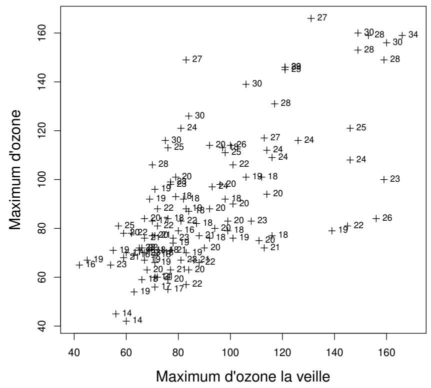
ε_i correspond à :

- erreur de mesure
- erreur d'échantillonnage
- facteur mal contrôlé
- oubli de facteurs

$\Rightarrow$ Introduction de variables supplémentaires pour réduire cette variabilité résiduelle

Rappel régression simple

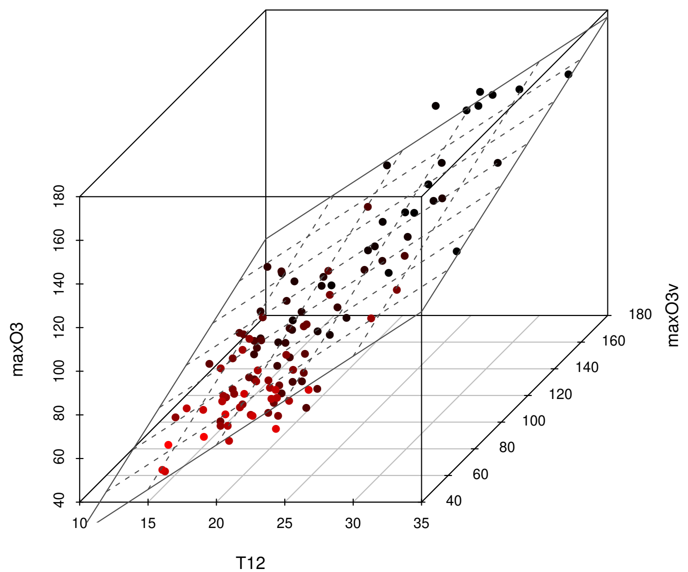
Régression du maximum d'ozone en fonction du maximu d'ozone de la veille




```
summary(lm(maxO3~maxO3v, data=ozone))
Coefficients:
          Estimate Std. Err t value Pr(>|t|)
(Inter) 28.50249  6.57153   4.337 3.21e-05 ***
maxO3v   0.68235  0.06929   9.848  < 2e-16 ***

Residual standard error: 20.64 on 110 DF
Multiple R-squared: 0.4686, Adj-R2: 0.4637
Fstat: 96.99 on 1 and 110 DF, p-value<2.2e-16
```


Rappel régression simple



$\Rightarrow$ Prendre en compte simultanément l'effet des deux variables

Définition du modèle de régression multiple

Sous forme indicée :

$$\begin{cases} \forall i = 1, \dots, n \quad Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i \\ \forall i = 1, \dots, n \quad \varepsilon_i \text{ i.i.d. }, \quad \mathbb{E}(\varepsilon_i) = 0, \quad \mathbb{V}(\varepsilon_i) = \sigma^2 \\ \forall i \neq k \quad cov(\varepsilon_i, \varepsilon_k) = 0 \end{cases}$$

Modèle traduit l'influence de chaque variable sur Y

- linéarité du modèle : linéarité par rapport aux paramètres
- additivité : les effets des variables s'additionnent
- modèle polynomial possible : $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$

Définition du modèle de régression multiple

$$
\begin{aligned}
Y_1 &= \beta_0 + \beta_1 x_{11} + \ldots + \beta_j x_{1j} + \ldots + \beta_p x_{1p} + \varepsilon_1 \\
\ldots & \quad \ldots \\
Y_i &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_j x_{ij} + \ldots + \beta_p x_{ip} + \varepsilon_i \\
\ldots & \quad \ldots \\
Y_n &= \beta_0 + \beta_1 x_{n1} + \ldots + \beta_j x_{nj} + \ldots + \beta_p x_{np} + \varepsilon_n
\end{aligned}
$$

Matriciellement :

$$Y = X\beta + E \quad \text{avec} \quad \mathbb{E}(E) = 0, \quad \mathbb{V}(E) = \sigma^2 Id$$

$$
\begin{bmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix}
=
\begin{bmatrix}
1 & x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\
\vdots & \vdots & & \vdots & & \vdots \\
1 & x_{i1} & & x_{ij} & & x_{ip} \\
\vdots & \vdots & & \vdots & & \vdots \\
1 & x_{n1} & \cdots & x_{nj} & \cdots & x_{np}
\end{bmatrix}
\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{bmatrix}
+
\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}
$$

Estimation des paramètres du modèle

Critère des moindres carrés

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}))^2 = \arg\min_{\beta} \|Y - X\beta\|^2$$

Dérivée matricielle par rapport à β (règles de dérivation : $\frac{\partial (A'Z)}{\partial A} = \frac{\partial (Z'A)}{\partial A} = Z$)

$$
\begin{aligned}
0 &= \frac{\partial \|Y - X\beta\|^2}{\partial \beta} = \frac{\partial (Y - X\beta)'(Y - X\beta)}{\partial \beta} \\
&= \frac{\partial (Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta)}{\partial \beta} = -X'Y - X'Y + X'X\beta + X'X\beta \\
\Longrightarrow & \quad X'X\hat{\beta} = X'Y
\end{aligned}
$$

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad \text{si } X'X \text{ est inversible}$$

Propriétés :

$\mathbb{E}(\hat{\beta}) = \beta\,;\ \ \mathbb{V}(\hat{\beta}) = (X'X)^{-1}\sigma^2\,;\ \ \mathbb{V}(\hat{\beta}_j) = \left[(X'X)^{-1}\right]_{jj}\sigma^2$

Estimation des paramètres du modèle

Prédiction et résidus

Valeurs prédites :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_j x_{ij} + \ldots + \hat{\beta}_p x_{ip}$$

Résidus :

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

Estimateur de la variabilité résiduelle σ^2 :

$$\hat{\sigma}^2 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{\ldots} = \frac{\sum_i \varepsilon_i^2}{\ldots} \qquad \mathbb{E}(\hat{\sigma}^2) = \sigma^2$$

Décomposition de la variabilité

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

Variabilité	totale	modèle	résiduelle
ddl	...	...	...

Source Variation	Somme des carrés	ddl	Carré Moyen
Modèle	$\sum_i (\hat{y}_i - \bar{y})^2$	...	$\frac{SCM}{p}$
Résidu	$\sum_i (y_i - \hat{y}_i)^2$	...	$\frac{SCR}{n-p-1}$
Total	$\sum_i (y_i - \bar{y})^2$	...	

Coefficient de détermination

$$R^2 = \frac{SC_{modele}}{SC_{total}} = 1 - \frac{SC_{residuelle}}{SC_{total}}$$

R^2 : pourcentage de variabilité de Y expliqué par le modèle

Propriétés :

- $0 \leq R^2 \leq 1$
- $R^2 = 0 \iff SC_{modele} = 0$
- $R^2 = 1 \iff SC_{modele} = SC_{total}$

Rq : $R^2 = r^2(y_i, \hat{y}_i)$

Inférence : test global

Objectifs : Le R^2 est-il significatif ? Le modèle est-il intéressant ?

Hypothèses :
H_0 :"$\forall j = 1, .., p \quad \beta_j = 0$" contre H_1 : "$\exists j = 1, .., p \ / \ \beta_j \neq 0$"

Si H_0 est vraie :

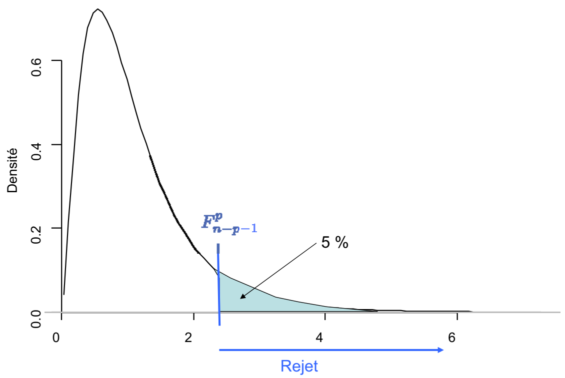
$$\mathbb{E}\left(\frac{SC_M}{p}\right) = \mathbb{E}\left(CM_M\right) = \sigma^2$$

$$\mathbb{E}\left(\frac{SC_R}{n-p-1}\right) = \mathbb{E}\left(CM_R\right) = \sigma^2$$

Principe du test : ...

Inférence : test global

Statistique de test : $F_{obs} = \dfrac{SC_M/p}{SC_R/(n-p-1)} = \dfrac{CM_M}{CM_R}$
Loi de la statistique de test : Sous H_0, $\mathcal{L}(F_{obs}) = \mathcal{F}^p_{n-p-1}$
Décision : $F_{obs} > \mathcal{F}^p_{n-p-1}(1-\alpha) \implies$ rejet de H_0 au seuil α



Inférence : test d'un coefficient de régression

$$\mathcal{L}(\hat{\beta}_j) = \mathcal{N}\left(\beta_j, \sigma^2_{\hat{\beta}_j}\right) \qquad \text{avec} \quad \sigma^2_{\hat{\beta}_j} = (X'X)^{-1}_{jj}\sigma^2$$

$$\mathcal{L}\left(\frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}}\right) = ...$$

$$\mathcal{L}\left(\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}}\right) = ...$$

Construction de tests ou d'intervalles de confiance sur les paramètres

Inférence : test d'un coefficient de régression

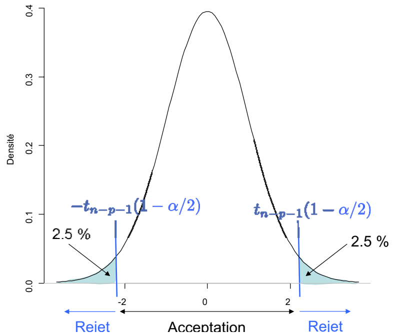
Hypothèses : H_0 : "$\beta_j = 0$" contre H_1 : "$\beta_j \neq 0$"

H_0 : la variable j n'apporte pas d'information supplémentaire intéressante sachant que les autres variables sont déjà dans le modèle

Statistique de test : $T_{obs} = \dfrac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$

Loi de la statistique de test sous H_0 **:** $\mathcal{L}(T_{obs}) = \mathcal{T}_{\nu=n-p-1}$

Décision : $|T_{obs}| > t_{n-p-1}(1-\alpha/2) \implies$ rejet de H_0 au seuil α



Exemple sur l'ozone

```
> summary(lm(max03~ ., data=ozone))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.24442   13.47190   0.909   0.3656
T9          -0.01901    1.12515  -0.017   0.9866
T12          2.22115    1.43294   1.550   0.1243
T15          0.55853    1.14464   0.488   0.6266
Ne9         -2.18909    0.93824  -2.333   0.0216 *
Ne12        -0.42102    1.36766  -0.308   0.7588
Ne15         0.18373    1.00279   0.183   0.8550
Vx9          0.94791    0.91228   1.039   0.3013
Vx12         0.03120    1.05523   0.030   0.9765
Vx15         0.41859    0.91568   0.457   0.6486
max03v       0.35198    0.06289   5.597 1.88e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.36 on 101 degrees of freedom
Multiple R-squared: 0.7638,     Adjusted R-squared: 0.7405
F-statistic: 32.67 on 10 and 101 DF,  p-value: < 2.2e-16
```


Sélection de variables

Comment construire un modèle ne contenant que des variables qui apportent de l'information ?

Plusieurs stratégies :

- Méthode descendante (backward) : on construit le modèle complet ; on reconstruit un modèle sans la variable explicative la moins intéressante ; on itère jusqu'à ce que toutes les variables explicatives soient intéressantes
- Méthode ascendante (forward) : on part du modèle avec la variable la plus intéressante ; on ajoute la variable qui, connaissant les autres variables du modèle, apporte le plus d'information complémentaire ; on itère jusqu'à ce qu'aucune variable n'apporte d'information intéressante
- Méthode stepwise : compromis entre les 2 méthodes ci-dessus
- Méthode du R^2 : on construit tous les sous-modèles possibles et on retient celui pour lequel la probabilité critique du test du R^2 est la plus petite (on rejette le plus fortement l'hypothèse : le modèle n'est pas intéressant)

Exemple sur l'ozone : sélection de variables

```
> library(FactoMineR)
> RegBest(y=ozone[,1],x=ozone[,-1],nbest=1)

$all[[1]]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -27.4196     9.0335  -3.035    0.003 **
T12           5.4687     0.4125  13.258   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.57 on 110 degrees of freedom
Multiple R-squared: 0.6151,     Adjusted R-squared: 0.6116
F-statistic: 175.8 on 1 and 110 DF,  p-value: < 2.2e-16

$all[[2]]
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -29.43810    8.00289  -3.678 0.000366 ***
T12           4.07197    0.44195   9.214 2.66e-15 ***
max03v        0.35425    0.06318   5.607 1.57e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.55 on 109 degrees of freedom
Multiple R-squared: 0.7012,     Adjusted R-squared: 0.6958
F-statistic: 127.9 on 2 and 109 DF,  p-value: < 2.2e-16

...

$all[[10]]
```

Exemple sur l'ozone : sélection de variables

```
> library(FactoMineR)
> RegBest(y=ozone[,1],x=ozone[,-1],nbest=1)

$summary
                            R2       Pvalue
Model with 1 variable   0.6150674 1.512025e-24
Model with 2 variables  0.7012408 2.541031e-29
Model with 3 variables  0.7519764 1.457692e-32
Model with 4 variables  0.7622198 1.763434e-32
Model with 5 variables  0.7630603 1.449905e-31
Model with 6 variables  0.7635768 1.130263e-30
Model with 7 variables  0.7637610 8.556709e-30
Model with 8 variables  0.7638390 6.076804e-29
Model with 9 variables  0.7638407 4.066941e-28
Model with 10 variables 0.7638413 2.545665e-27

$best
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.76225   11.10038   0.879    0.381
T12          2.85308    0.48052   5.937 3.57e-08 ***
Ne9         -3.02423    0.64342  -4.700 7.71e-06 ***
max03v       0.37571    0.05801   6.477 2.85e-09 ***

Residual standard error: 14.23 on 108 degrees of freedom
Multiple R-squared: 0.752,      Adjusted R-squared: 0.7451
F-statistic: 109.1 on 3 and 108 DF,  p-value: < 2.2e-16
```

$\Rightarrow$ Le meilleur modèle en prévision contient 3 variables
$\Rightarrow$ Ajouter d'autres variables améliore l'ajustement mais pas la prévision

Inférence : intervalle de confiance d'un coefficient

$$\mathcal{L}\left(\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}}\right) = \mathcal{T}_{n-p-1}$$

$$-t_{n-p-1}(1-\alpha/2) \leq \left(\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}}\right) \leq t_{n-p-1}(1-\alpha/2)$$

Intervalle de confiance :

$$\beta_j \in \left[\hat{\beta}_j - t_{n-p-1}(1-\alpha/2) \times \hat{\sigma}_{\hat{\beta}_j} \; ; \; \hat{\beta}_j + t_{n-p-1}(1-\alpha/2) \times \hat{\sigma}_{\hat{\beta}_j}\right]$$

```
> model = lm(max03~T12+Ne9+max03v,data=ozone)
> confint(model)

                  2.5 %     97.5 %
(Intercept) -12.2406259 31.7651193
T12           1.9005988  3.8055572
Ne9          -4.2995961 -1.7488656
max03v        0.2607251  0.4907008
```


Prévisions

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots \hat{\beta}_p x_{ip}$$

```
> xnew <- matrix(c(19,8,70,23,10,95),nrow=2,byrow=TRUE)
> xnew
     [,1] [,2] [,3]
[1,]   19    8   70
[2,]   23   10   95
> colnames(xnew) <- c("T12","Ne9","max03v")
> xnew <- as.data.frame(xnew)
> xnew
  T12 Ne9 max03v
1  19   8     70
2  23  10     95
> predict(model,xnew,interval="pred")
       fit      lwr       upr
1 66.07679 37.52847  94.62512
2 80.83347 51.58514 110.08179
```


Analyse graphique des résidus du modèle

```
> model = lm(max03~T12+Ne9+max03v,data=ozone)
> hist(residuals(model),main="Histogramme des résidus",xlab="Résidus")
```