Régression linéaire multiple

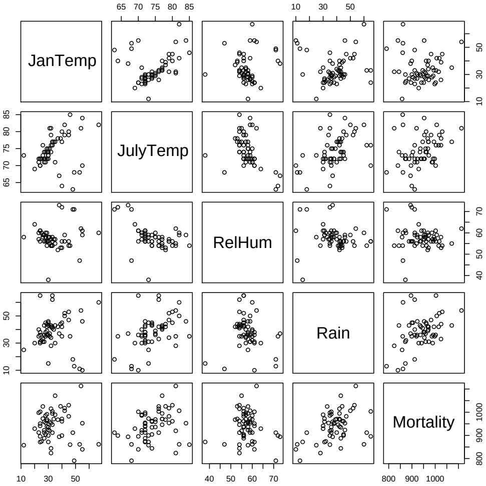
Thierry Denœux

Printemps 2010

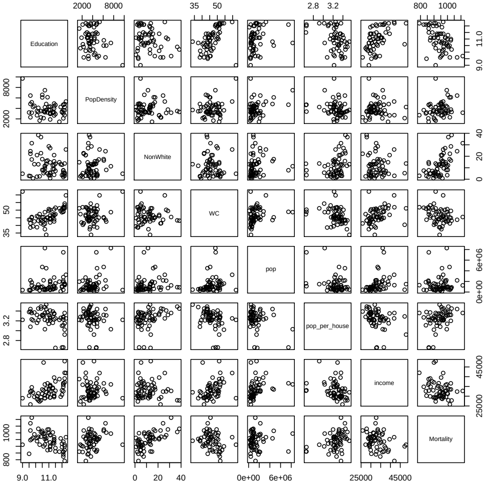
Données SMSA

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

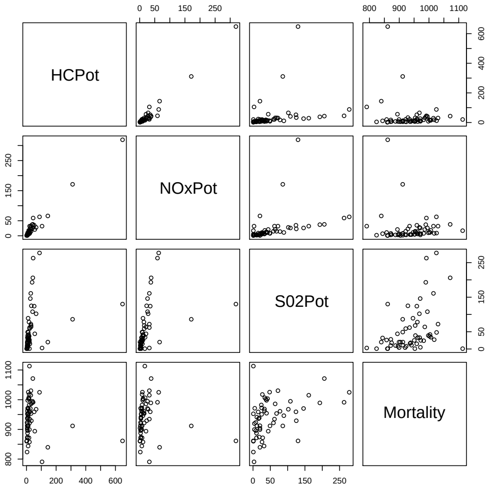
Variables climatiques et mortalité



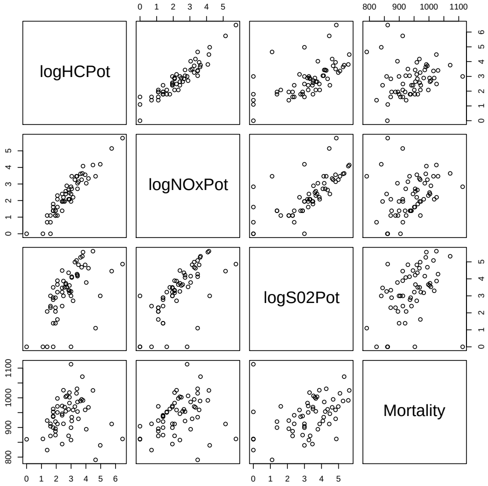
Variables sociologiques et mortalité



Variables de pollution et mortalité



Variables de pollution transformées et mortalité



Le problème

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

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

Le modèle

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

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

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

- Matriciellement, on peut écrire

$$Y = Xb + \varepsilon$$

avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix} \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} b = \begin{pmatrix} b_0 \\ \vdots \\ b_p \end{pmatrix}$$

Plan

Critère des moindres carrés

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

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

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

Solution

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

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

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

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

  est un estimateur sans biais de σ^2.

Application en R

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

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

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

Plan

Coefficient de détermination

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

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

soit

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

- Cette équation montre que la quantité

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

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

R^2 et R^2 ajusté

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

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

Coefficients de détermination en R

```
> summary(reg.smsa)

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

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

⋮

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

Plan

Hypothèse de normalité des résidus

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

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

Significativité du R^2

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

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

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

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

Significativité des coefficients de régression

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

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

- Sous H_0,

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

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

- On en déduit le degré de signification :

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

Tests de significativité en R

```
> summary(reg.smsa)

:

Coefficients :
              Estimate     Std. Error   t value   Pr(>|t|)
 (Intercept)  1.333e+03    2.917e+02    4.569     3.94e-05    ***
 JanTemp      -2.305e+00   8.795e-01    -2.621    0.0120      *
 JulyTemp     -1.657e+00   2.051e+00    -0.808    0.4236
 RelHum       4.067e-01    1.070e+00    0.380     0.7058
 Rain         1.444e+00    5.847e-01    2.469     0.0175      *
 Education    -9.458e+00   9.080e+00    -1.042    0.3033
 PopDensity   4.509e-03    4.311e-03    1.046     0.3014
 NonWhite     5.194e+00    1.005e+00    5.167     5.55e-06    ***
:
Residual standard error : 34.58 on 44 degrees of freedom
Multiple R-squared : 0.7672, Adjusted R-squared : 0.6931
F-statistic : 10.36 on 14 and 44 DF, p-value : 9.864e-10
```

Plan

Principe

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

Analyse des résidus

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

Résidus standardisés

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

Résidus studentisés

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

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

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

Distance de Cook

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

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

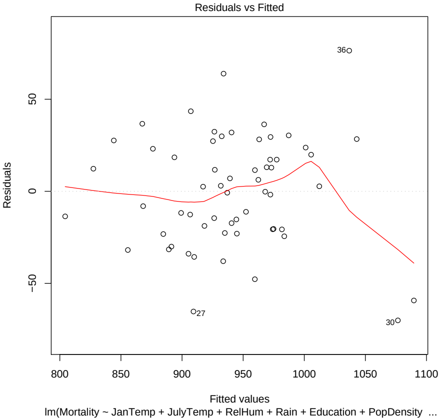
- Une distance de Cook supérieure à 1 indique en général une influence anormale.

Diagnostic de la régression en R

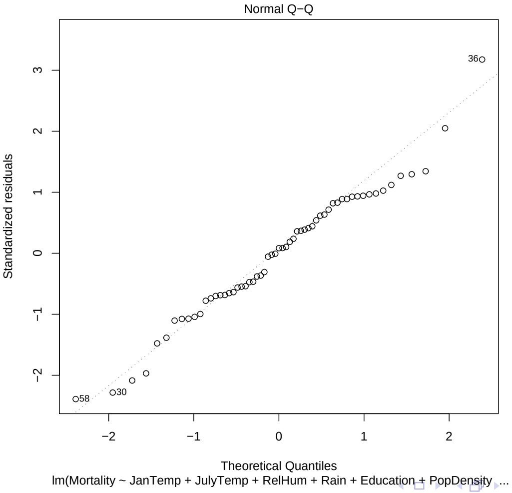
- Prédictions $\widehat{y}_i$: `fitted(reg.smsa)`
- Résidus bruts $\widehat{\varepsilon}_i$: `resid(reg.smsa)`
- Résidus standardisés $\widehat{\varepsilon}'_i$: `rstandard(reg.smsa)`
- Résidus studentisés : $\widehat{\varepsilon}^*_i$: `rstudent(reg.smsa)`
- Distances de Cook D_i : `cooks.distance(reg.smsa)`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

Application aux données SMSA (1)

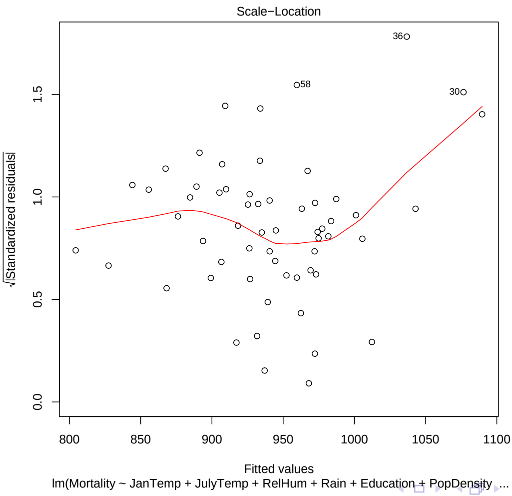
```
> plot(reg.smsa)
```



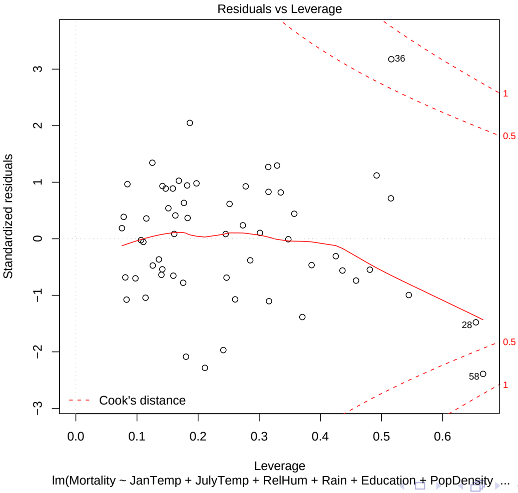
Application aux données SMSA (2)



Application aux données SMSA (3)



Application aux données SMSA (4)



Principe

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

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

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

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

et

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

Intervalles de prévision et de confiance

On en déduit :

- l'intervalle de prévision sur Y_0 :

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

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

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

Exemple en R

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

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

Plan

Objectif de la sélection

Après avoir vérifié la validité des hypothèses du modèle linéaire et effectué les tests de significativité, plusieurs situations peuvent se rencontrer :

1. le R^2 et les coefficients sont tous significatifs ;
2. le R^2 et les coefficients sont tous non significatifs ;
3. autres cas (ex : certains coefficients significatifs et d'autres non).

Dans le cas 3, il convient sûrement de supprimer des variables (raisons de stabilité numérique et critères économiques), mais lesquelles ? (suppression des variables non significatives incorrect).

Approche générale

- Il faut donc envisager des techniques systématiques de sélection de variables, permettant de trouver un modèle satisfaisant utilisant $m < p$ variables parmi les p variables initiales.
- Remarques :
  - Il n'existe pas nécessairement un seul meilleur ensemble : le choix final peut prendre en compte des critères extra-statistiques.
  - L'existence de techniques automatiques de sélection ne dispense pas d'une réflexion sérieuse sur la nature des variables en question !
- Il faut définir :
  - un critère de choix de modèle
  - un algorithme de recherche parmi tous les modèles possibles.

Critère de choix d'un modèles

- Le R^2 est peu adapté en général (sauf si m fixé), car il augmente de façon monotone avec le nombre de variables.
- Le critère

$$\widehat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\widehat{\epsilon}_i^2 = \frac{n}{n-1}(1-\overline{R}^2)S_Y^2$$

  est intéressant car non monotone (revient à maximiser $\overline{R}^2$).
- Il existe de nombreux autres critères (AIC, AICc, BIC,...). Par exemple, le critère AIC est défini par par -2 fois la log-vraisemblance + 2 fois le nombre de paramètres. En régression, cela revient à :

$$AIC = n\log\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\epsilon}_i^2\right) + 2(p+1)$$

Plan

Recherche exhaustive

- Il s'agit de trouver le meilleur sous-ensemble de variables parmi les p variables initiales.
- Il y a en tout $2^p - 1$ sous-ensembles non vides parmi l'ensemble des p variables :
  - 31 pour $p = 5$,
  - 1023 pour $p = 10$,
  - 1048575 pour $p = 20$!
- En pratique, il faut faire une recherche heuristique qui permet de trouver un « bon » modèle, pas nécessairement le meilleur.

Sélection pas à pas

- Principe : élimination successive ou ajout successif de variables.
- Méthode ascendante : on ajoute incrémentalement des variables en maximisant à chaque fois un critère ($\overline{R}^2$ ou AIC par exemple).
- Méthode descendante : on commence avec les p variables, puis on retire à chaque pas la variable dont la suppression fait décroître le moins le critère.
- Méthode "stepwise" : en partant d'un modèle quelconque, on considère trois types modifications possibles :
  - ajout d'une variable,
  - suppression d'une variable,
  - échange d'une variable dans le modèle avec une variable non encore dans le modèle.

  L'algorithme s'arrête lorsqu'un optimum local a été trouvé.

Sélection de modèle en R

```
> reg.smsa1 <- step(reg.smsa,trace=F)
> summary(reg.smsa1)

Call :
lm(formula = Mortality ~ JanTemp + Rain + Education + NonWhite +
WC + logNOxPot)

Coefficients :
              Estimate   Std. Error   t value    Pr(>|t|)
 (Intercept)  1031.9491    80.2930    12.852    < 2e-16    ***
 JanTemp        -2.0235     0.5145    -3.933    0.000250   ***
 Rain            1.8117     0.5305     3.415    0.001245    **
 Education     -10.7463     7.0797    -1.518    0.135098
 NonWhite        4.0401     0.6216     6.500    3.10e-08   ***
 WC             -1.4514     1.0451    -1.389    0.170817
 logNOxPot      19.2481     4.5220     4.257    8.70e-05   ***
Residual standard error : 33.72 on 52 degrees of freedom
Multiple R-squared : 0.7383, Adjusted R-squared : 0.7081
F-statistic : 24.45 on 6 and 52 DF, p-value : 1.543e-13
```