MAT7381 Chapitre 5
Régression multiple

5.1 Le modèle

La régression multiple est une généralisation de la régression simple qui permet d'expliquer une variable endogène y en fonction de p variables exogènes $x_1, x_2, \ldots, x_p$. Nous disposons de n observations $(y_i, x_{i1},\ldots, x_{ip})$, $i = 1,\ldots, n$. Le modèle s'écrit

$$y_j = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i,\ i = 1, \ldots, n$$

Dans le langage matriciel, ces équations s'écrivent

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

où

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},\ \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix},\ \boldsymbol{\beta} = \begin{bmatrix} \beta_o \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Donc $\boldsymbol{y}$ est un vecteur $n \times 1$, $\mathbf{X}$ est une matrice $n \times q$, où $q = p + 1$ et $\boldsymbol{\beta}$ un vecteur $q \times 1$. Nous avons la table d'analyse de la variance suivante, où $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$:

Source	*Somme des carrés*	*d.l.*	*Moyenne des carrés*	*F*
Régression	SCE = $\boldsymbol{y}'\left(\mathbf{H} - \frac{1}{n}\boldsymbol{e}\boldsymbol{e}'\right)\boldsymbol{y}$	p	MCE = SCE/p	MCE/MCR
Résiduelle	SCR = $\boldsymbol{y}'(\mathbf{I}-\mathbf{H})\boldsymbol{y}$	n-p-1	MCR = SCR/(n-p-1)	
Total	SCT = $\boldsymbol{y}'\left(\mathbf{I} - \frac{\boldsymbol{e}\boldsymbol{e}'}{n}\right)\boldsymbol{y}$	n-1	MCT = SCT/(n-1)	

Remarque *La somme des carrés expliquée peut également s'écrire*

$$\text{SCE} = \boldsymbol{y}'\mathbf{H}\boldsymbol{y} - \frac{\boldsymbol{y}'\boldsymbol{e}\boldsymbol{e}'\boldsymbol{y}}{n} = \boldsymbol{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y} - \frac{\boldsymbol{y}'\boldsymbol{e}\boldsymbol{e}'\boldsymbol{y}}{n} = \boldsymbol{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y} - \frac{\boldsymbol{y}'\boldsymbol{e}\boldsymbol{e}'\boldsymbol{y}}{n}$$

$$= (\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{X}\hat{\boldsymbol{\beta}}) - (\bar{y}\boldsymbol{e})'(\bar{y}\boldsymbol{e}) = \|\hat{\boldsymbol{\mu}}\|^2 - \|\hat{\boldsymbol{\mu}}_o\|^2$$

SCE *est donc une différence de longueurs au carré entre deux estimateurs de l'espérance de* $\boldsymbol{y}$: *l'estimateur sous le modèle,* $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, *et l'estimateur sous l'hypothèse que* $\beta_1 = \ldots = \beta_p = 0$, $\hat{\boldsymbol{\mu}}_o = \bar{y}\boldsymbol{e}$.

La somme des carrés résiduelle peut s'écrire $\text{SCR} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \|\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$, *le carré de la longueur du vecteur des résidus* $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ ■

Soit $[x_{01}; x_{02}; \ldots; x_{0p}]$ une valeur du vecteur des variables exogènes $x_1, x_2,\ldots, x_p$, et supposons qu'on veuille estimer la valeur moyenne $\mu_o = \beta_o + \beta_1 x_{01} + \beta_1 x_{02} + \ldots + \beta_p x_{0p}$, ou faire une prévision de la valeur de $y_o = \mu_o + \varepsilon_o$ correspondant à $[x_{01}; x_{02}; \ldots; x_{0p}]$, soit $\mu_o + \varepsilon_o$. Les intervalles de confiance à 100(1-α)% sont donnés par

$$\boldsymbol{x}_o'\hat{\boldsymbol{\beta}} - t_{n-q;\alpha/2}\,\hat{\sigma}_{\boldsymbol{x}_o'\hat{\boldsymbol{\beta}}} \le \boldsymbol{x}_o'\boldsymbol{\beta} \le \boldsymbol{x}_o'\hat{\boldsymbol{\beta}} + t_{n-q;\alpha/2}\,\hat{\sigma}_{\boldsymbol{x}_o'\hat{\boldsymbol{\beta}}}$$

où

$$\hat{\sigma}_{\boldsymbol{x}_o'\hat{\boldsymbol{\beta}}} = \hat{\sigma}\sqrt{\boldsymbol{x}_o'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{x}_o}$$

et $\boldsymbol{x}_o = [1, x_{01}, x_{02},..., x_{0p}]'$ est un vecteur $(p + 1) \times 1$.

Les limites de prédiction à 100(1-α)% sont données par

$$\boldsymbol{x}_o'\hat{\boldsymbol{\beta}} - t_{n-q;\alpha/2}\,\hat{\sigma}_{y_o-\hat{y}_o} \le y_o \le \boldsymbol{x}_o'\hat{\boldsymbol{\beta}} + t_{n-q;\alpha/2}\,\hat{\sigma}_{y_o-\hat{y}_o}$$

où

$$\hat{\sigma}_{y_o-\hat{y}_o} = \hat{\sigma}\sqrt{1+\boldsymbol{x}_o'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{x}_o}\,.$$

Le premier est un intervalle de confiance pour $\mu_o = \beta_o+\beta_1x_{01}+\beta_1x_{02}+...+\beta_px_{0p} = E(y_o|\boldsymbol{x}_o)$. Le deuxième est une prévision concernant une valeur future $y_o = \beta_o+\beta_1x_{01}+\beta_1x_{02}+...+\beta_px_{0p}+\varepsilon$.

5.2 Coefficients de corrélation

Deux sortes de coefficients de corrélation sont utiles dans la régression multiple, le *coefficient de corrélation multiple*, et *le coefficient de corrélation partielle*.

Coefficient de corrélation multiple Le coefficient de corrélation multiple est la généralisation immédiate du coefficient de corrélation ordinaire, et en fait son carré R^2, appelé *coefficient de détermination* est défini exactement comme r^2:

$$R^2 = \frac{\text{SCE}}{\text{SCT}} = \frac{\boldsymbol{y}'(\mathbf{H}-\boldsymbol{e}(\boldsymbol{e}'\boldsymbol{e})^{-1}\boldsymbol{e}')\boldsymbol{y}}{\boldsymbol{y}'\mathbf{C}\boldsymbol{y}} = \frac{\boldsymbol{y}'\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'-\boldsymbol{e}(\boldsymbol{e}'\boldsymbol{e})^{-1}\boldsymbol{e}'\right)\boldsymbol{y}}{\boldsymbol{y}'\left(\mathbf{I}-\boldsymbol{e}(\boldsymbol{e}'\boldsymbol{e})^{-1}\boldsymbol{e}'\right)\boldsymbol{y}} -$$

ce qui peut s'écrire

$$R^2 = \frac{\boldsymbol{y}'\left((\mathbf{I}-\boldsymbol{e}(\boldsymbol{e}'\boldsymbol{e})^{-1}\boldsymbol{e}')-(\mathbf{I}-\mathbf{P})\right)\boldsymbol{y}}{\boldsymbol{y}'\mathbf{C}\boldsymbol{y}} = \frac{\text{SCR}_o-\text{SCR}}{\text{SCR}_o},$$

où $\text{SCR}_o = \text{SCT} = \boldsymbol{y}'\mathbf{C}\boldsymbol{y}$ est la somme des carrés dans le modèle $\boldsymbol{y} = \beta\boldsymbol{e} + \boldsymbol{\varepsilon}$. En d'autres termes, R^2 est la réduction relative de la somme des carrés résiduelle due à l'ajout de l'ensemble des variables exogènes.

Il y a d'autres interprétations possibles. On peut montrer que le coefficient de corrélation multiple R est la corrélation *maximale* entre y et une combinaison linéaire de $x_1, x_2, ..., x_p$. C'est-à-dire, si $r(\beta_1, ..., \beta_p)$ désigne le coefficient de corrélation entre y et une combinaison linéaire $\beta_1x_1 + ... , \beta_px_p$, alors le coefficient de corrélation multiple est la valeur maximale de $r(\beta_1, ..., \beta_p)$. Les valeurs de $\beta_1, ..., \beta_p$ qui maximisent cette corrélation sont justement les estimateurs $\hat{\beta}_1, ..., \hat{\beta}_p$ des paramètres de régression.

Autrement dit, le coefficient de corrélation multiple est le coefficient de corrélation entre les y_i et les $\hat{y}_i$:

$$R = \frac{\boldsymbol{y}'\mathbf{C}\hat{\boldsymbol{y}}}{\sqrt{\boldsymbol{y}'\mathbf{C}\boldsymbol{y}}\sqrt{\hat{\boldsymbol{y}}'\mathbf{C}\hat{\boldsymbol{y}}}} = \frac{\boldsymbol{y}'\mathbf{C}\mathbf{X}\hat{\boldsymbol{\beta}}}{\sqrt{\boldsymbol{y}'\mathbf{C}\boldsymbol{y}}\sqrt{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{C}\mathbf{X}\hat{\boldsymbol{\beta}}}}$$

Coefficient de corrélation partielle Dans le contexte de la régression multiple, le coefficient de corrélation partielle mesure la dépendance linéaire entre y et l'une des variables exogènes, disons x_1, *conditionnellement aux autres variables*. Il est défini par

$$r^2_{y1.2} = \frac{\boldsymbol{y}'(\mathbf{H}-\mathbf{H}_{o2})\boldsymbol{y}}{\boldsymbol{y}'(\mathbf{I}-\mathbf{H}_{o2})\boldsymbol{y}} = \frac{\boldsymbol{y}'\left[(\mathbf{I}-\mathbf{H}_{o2})-(\mathbf{I}-\mathbf{H})\right]\boldsymbol{y}}{\boldsymbol{y}'(\mathbf{I}-\mathbf{H}_{o2})\boldsymbol{y}}$$

où $\mathbf{H}$ est le projecteur orthogonal sur $\mathcal{C}(\mathbf{X})$, $\mathbf{X} = [\boldsymbol{x}_o, \boldsymbol{x}_1, \mathbf{X}_2]$, $\mathbf{H}_{o2}$ est le projecteur orthogonal sur $\mathcal{C}[\boldsymbol{x}_o \mid \mathbf{X}_2]$.

C'est-à-dire, $r^2_{y1.2}$ est la réduction relative de la somme des carrés résiduelle due à l'introduction de la variable x_1.

On peut aussi montrer que $r_{y1.2}$ est le coefficient de corrélation entre $\boldsymbol{y} - \hat{\boldsymbol{y}} = (\mathbf{I} - \mathbf{H}_{o2})\boldsymbol{y}$ et $\boldsymbol{x}_1 - \hat{\boldsymbol{x}}_1 = (\mathbf{I} - \mathbf{H}_{o2})\boldsymbol{x}_1$, c'est-à-dire, les résidus des régression de y sur $\mathbf{X}_2$ et de x_1 sur $\mathbf{X}_2$:

$$r_{y1.2} = \frac{\boldsymbol{y}'(\mathbf{I} - \mathbf{H}_{o2})\mathbf{C}(\mathbf{I} - \mathbf{H}_{o2})\boldsymbol{x}_1}{\sqrt{\boldsymbol{y}'(\mathbf{I} - \mathbf{H}_{o2})\mathbf{C}(\mathbf{I} - \mathbf{H}_{o2})\boldsymbol{y}}\sqrt{\boldsymbol{x}_1'(\mathbf{I} - \mathbf{H}_{o2})\mathbf{C}(\mathbf{I} - \mathbf{H}_{o2})\boldsymbol{x}_1}} = \frac{\boldsymbol{y}'(\mathbf{I} - \mathbf{H}_{o2})\boldsymbol{x}_1}{\sqrt{\boldsymbol{y}'(\mathbf{I} - \mathbf{H}_{o2})\boldsymbol{y}}\sqrt{\boldsymbol{x}_1'(\mathbf{I} - \mathbf{H}_{o2})\boldsymbol{x}_1}}$$

car $\mathbf{C}(\mathbf{I} - \mathbf{H}_{o2}) = (\mathbf{I} - \mathbf{H}_{o2})$.

5.3 Un exemple

Exemple. *On analyse des réponses données par* 146 *classes à des questionnaires d'évaluation de professeurs. Les variables exogènes sont les cotes moyennes sur chacun des points suivants* :

x_1: `Interet`: L'intérêt porté au cours par l'étudiant(e).
x_2: `Maniere`: La façon de traiter les étudiants
x_3: `Cours`: La qualité et pertinence du cours lui-même
x_4: `Repond`: La manière dont le professeur répond aux questions
x_5: `Taille`: La taille du groupe

La variable endogène est

y: `Prof`: L'évaluation globale donnée au professeur

Voici les résultats fournis par le logiciel MINITAB:

```
The regression equation is
Prof = - 0,012 - 0,0117 Interet + 0,528 Maniere + 0,518 Cours + 0,00706 Repond + 0,000337
Taille
```

Une anomalie est immédiatement remarquée : le coefficient de `Interet` est négatif : plus la classe trouve le cours intéressant, moins elle aime le professeur. On attribuera cette anomalie à un accident du hasard si on peut conclure que le coefficient n'est pas significativement différent de 0. Le tableau suivant teste les hypothèses $\beta_o = 0$, $\beta_1 = 0$, ..., $\beta_6 = 0$. En effet, la variable `Interet` n'est pas significative, puisque $p = 0{,}835$.

Predictor	Coef	StDev	T	P
Constant	-0,0116	0,1377	-0,08	0,933
Interet	-0,01171	0,05605	-0,21	0,835
Maniere	0,52802	0,04358	12,12	0,000
Cours	0,51794	0,08888	5,83	0,000
Repond	0,007060	0,007065	1,00	0,320
Taille	0,0003374	0,0009206	0,37	0,715

S = 0,2050 R-Sq = 89,9% R-Sq(adj) = 89,5%

Le *coefficient de détermination* de 89,9% signifie que les variables exogènes expliquent ensemble une bonne part de la dispersion des y. L'analyse de variance dans le tableau suivant indique que la dépendance observée est globalement significative (on rejette l'hypothèse que *tous* les coefficients à part β_o sont nuls).

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	5	46.2371	9.2474	219.97	0.000
Error	124	5.2128	0.0420		
Total	129	51.4499			

Nous répétons l'analyse en éliminant la variable qui semble être la moins significative, `Interet` :

```
The regression equation is
Prof = - 0,009 + 0,530 Maniere + 0,504 Cours + 0,00704 Repond +0,000332 Taille

Predictor        Coef       StDev          T        P
Constant      -0,0086      0,1364      -0,06    0,950
Maniere       0,52956     0,04278      12,38    0,000
Cours         0,50409     0,05894       8,55    0,000
Repond       0,007039    0,007037       1,00    0,319
Taille      0,0003324   0,0009168       0,36    0,717
S = 0,2042      R-Sq = 89,9%     R-Sq(adj) = 89,5%
```

```
Analysis of Variance

Source            DF          SS          MS         F        P
Regression         4      46,235      11,559    277,07    0,000
Residual Error   125       5,215       0,042
Total            129      51,450
```

Le coefficient de détermination n'a presque pas changé (si plus de décimales avaient été montrées, on aurait constaté une légère baisse due à l'élimination de `Interet`). Ceci montre à quel point la variable `Interet` ne contribuait pas à la prédiction de *y en présence des autres variables exogènes*.

Tous les coefficients ont le bon signe. Mais on voit dans le tableau, que la variable `Taille` demeure non significative. On l'élimine:

```
The regression equation is
Prof = 0,060 + 0,534 Maniere + 0,486 Cours + 0,00545 Repond

Predictor        Coef       StDev          T        P
Constant       0,0598      0,1283       0,47    0,642
Maniere       0,53449     0,04319      12,38    0,000
Cours         0,48633     0,05909       8,23    0,000
Repond       0,005447    0,006854       0,79    0,428

S = 0,2139      R-Sq = 88,6%     R-Sq(adj) = 88,3%
```

Finalement, nous découvrons que la variable `Répond` est elle aussi non significative. Nous fixons donc comme prédicateurs les seules variables `Maniere` et `cours`. Ceci signifie que les facteurs qui déterminent l'évaluation globale faite par les étudiants sont la qualité du cours, et la gentillesse du professeur. Voici les résultats. On constate entre autres que le coefficient de détermination est presque aussi élevé qu'il l'était avec toutes les variables introduites au début.

```
The regression equation is
Prof = 0,126 + 0,535 Maniere + 0,485 Cours

Predictor        Coef       StDev          T        P
Constant      0,12643     0,09694       1,30    0,194
Maniere       0,53503     0,04313      12,41    0,000
Cours         0,48511     0,05900       8,22    0,000

S = 0,2136      R-Sq = 88,5%     R-Sq(adj) = 88,4%
```

```
Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      50,364      25,182    551,87    0,000
Residual Error   143       6,525       0,046
Total            145      56,889
```

Le tableau suivant donne un intervalle de confiance et des limites de prédiction pour `Maniere` = 2,5

et Cours = 2,5.

```
     Fit  Stdev.Fit          95% C.I.          95% P.I.
  2.6768     0.0193   ( 2.6386, 2.7150)   ( 2.2527, 3.1009)
```

Il serait intéressant d'obtenir un tableau de corrélations pour expliquer les résultats auxquels nous avons abouti:

```
           Prof  Maniere    Cours  Interet   Taille
Maniere   0,912
Cours     0,873    0,806
Interet   0,703    0,623    0,847
Taille    0,027    0,045   -0,019   -0,012
Repond    0,009   -0,008   -0,022   -0,019   -0,061
```

Chose curieuse: la variable Interet, qui a été éliminée en premier, est en fait assez fortement corrélée avec *y*. À elle seule, elle a une valeur de prédiction certaine. Pourquoi n'a-t-elle pas été conservée? Le fait est que Interet est fortement corrélée avec Maniere et Cours, les deux variables qui ont été retenues. Donc en présence de ces deux variables, elle n'apporte pas beaucoup d'information supplémentaire, c'est-à-dire, de l'information qui ne soit déjà contenue dans Maniere et Cours. Lorsque Maniere et Cours sont fixes, il n'y a pas une forte corrélation entre Interet et *y*.

Remarquez, d'ailleurs, qu'on aurait pu substituer INTEREST à COURSE sans grande perte :

```
The regression equation is
INSTRUCR = 0.410 + 0.697 MANNER + 0.223 INTEREST

Predictor       Coef       Stdev    t-ratio        p
Constant     0.41001     0.09139       4.49    0.000
MANNER       0.69705     0.03590      19.41    0.000
INTEREST     0.22260     0.04017       5.54    0.000

s = 0.2352      R-sq = 86.1%     R-sq(adj) = 85.9%
```

5.4 La régression Polynomiale

La régression polynomiale est une application particulière de la régression multiple : une seule variable *x* est considérée, mais l'équation de prédiction est un polynôme de degré supérieur à un. Le modèle est

$$y_i = \beta_o + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \varepsilon_i \quad , i = 1, \ldots, n$$

Pour voir si c'est utile de prendre un polynôme de degré p plutôt qu'un polynôme de degré p-1, on garde le polynôme de degré p si

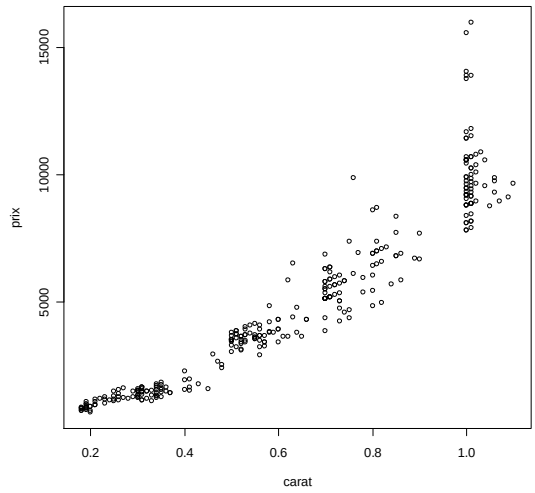
$$\frac{\mathrm{SCR}(p-1) \; - \; \mathrm{SCR}(p)}{\mathrm{SCR}(p)/(n-p-1)} > F_{1,n\text{-}p\text{-}1;\alpha}$$

où SCR(q) désigne la somme des carrés résiduelle dans le modèle polynomial de degré q.

La régression polynomiale constitue une façon de tester l'hypothèse de linéarité dans un modèle de régression simple : on suppose au départ un modèle polynomial de degré p (≥2) et on teste successivement les hypothèses $\beta_p = 0$, $\beta_{p-1} = 0$, etc.

Exemple Le graphique suivant montre la relation entre la grosseur (en nombre de carats) d'un diamant et son prix. Il révèle ce qu'on sait déjà : le prix ne croît pas linéairement avec la grosseur. Les variables sont

carat La grosseur du diamant, en nombre de carats
prix Le prix du diamant en dollars US



On pourrait donc considérer un modèle polynomial. Commençons par un polynôme de degré 3 (`cara2` = `carat`^{2} et `carat3`=`carat`^{3}) :

```
> summary(lm(prix~carat+cart2+carat3))
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     786.3      765.4   1.027   0.3051
carat         -2564.2     4636.9  -0.553   0.5807
carat2        16638.9     8185.3   2.033   0.0429 *
carat3        -5162.5     4341.9  -1.189   0.2354
Residual standard error: 1017 on 304 degrees of freedom
Multiple R-Squared: 0.9116,     Adjusted R-squared: 0.9107
F-statistic:  1045 on 3 and 304 DF,  p-value: < 2.2e-16
```

Les tests montrent que le terme cubique est non significatif, le terme linéaire pas du tout, et le terme quadratique tout juste significatif. Pourtant le test global montre que la relation est quand même très significative. La contradiction est due à la forte dépendance entre les variables `carat`, `carat2` et `carat3`. C'est pour cela qu'il vaut mieux utiliser la procédure `anova`, qui introduit les termes un à la fois.

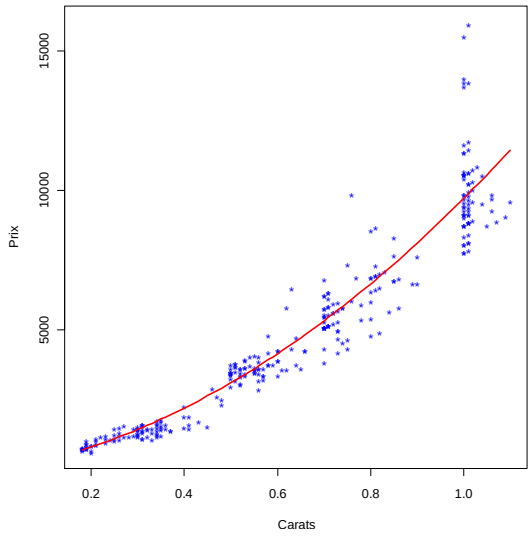
```
> anova(lm(prix~carat+cart2+carat3))
Analysis of Variance Table

Response: prix
           Df     Sum Sq    Mean Sq   F value    Pr(>F)
carat       1 3173248722 3173248722 3069.6825 < 2.2e-16 ***
carat2      1   66460799   66460799   64.2917 2.335e-14 ***
carat3      1    1461352    1461352    1.4137    0.2354
Residuals 304  314256474    1033738
```

Il est clair qu'on peut se passer du terme cubique. Voici donc la régression avec seuls les termes linéaire et quadratique.

```
> summary(lm(prix~carat+cart2))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -42.51     316.37  -0.134   0.8932
carat        2786.10    1119.61   2.488   0.0134 *
carat2       6961.71     868.83   8.013  2.4e-14 ***

Residual standard error: 1017 on 305 degrees of freedom
Multiple R-Squared: 0.9112,     Adjusted R-squared: 0.9106
F-statistic:  1565 on 2 and 305 DF,  p-value: < 2.2e-16
```

Remarquez que le coefficient de détermination n'est que très légèrement inférieur à celui du modèle cubique. ■

Voici la courbe polynomiale, superposée au nuage de points :



On constate que l'ajustement est bon, sauf pour les très gros diamants, dont les prix peuvent aller très loin. Il serait probablement futile de tenter d'inclure dans un même modèle des diamants exceptionnellement gros. On constate par ailleurs, des signes certains d'hétéroscédasticité.

Polynômes orthogonaux

La procédure `anova` du logiciel R est équivalente à une orthogonalisation des colonnes de la matrice **X** qui, bien qu'elle soit le plus souvent appliquée à la régression polynomiale, peut en fait s'appliquer à toute régression multiple. C'est une procédure qui vise à éliminer les dépendances entre les colonnes en choisissant une base orthonormale de l'espace $\mathcal{C}(\mathbf{X})$. La matrice **X** normalement comprend une première colonne $\boldsymbol{e}$ dont tous les éléments sont égaux à 1. Posons donc

$$\mathbf{X} = [\boldsymbol{e} \mid x_1 \mid \ldots \mid \boldsymbol{x}_p]$$

On remplace cette matrice par une matrice $\mathbf{Z} = [\boldsymbol{e} \mid z_1 \mid \ldots \mid \boldsymbol{z}_p]$ construite de la façon suivante :

- La première colonne reste telle quelle : $\boldsymbol{e}$
- La deuxième, $\boldsymbol{z}_1$, est la projection de $\boldsymbol{x}_1$ sur le complément orthogonale de $\boldsymbol{e}$: $\boldsymbol{z}_1 = (\mathbf{I}-\boldsymbol{e}\boldsymbol{e}'/n)\boldsymbol{x}_1$.
- La troisième, $\boldsymbol{z}_2$, est la projection de $\boldsymbol{x}_2$ sur le complément orthogonale de $\mathcal{C}([\boldsymbol{e} \mid \boldsymbol{z}_1])$: $\boldsymbol{z}_2 = [\mathbf{I} - \mathbf{Z}_{01}(\mathbf{Z}_{01}'\mathbf{Z}_{01})^{-1}\mathbf{Z}_{01}']\boldsymbol{x}_2$,où $\mathbf{Z}_{01} = [\boldsymbol{e} \mid \boldsymbol{z}_1]$
- La quatrième, $\boldsymbol{z}_3$, est la projection de $\boldsymbol{x}_3$ sur le complément orthogonale de $\mathcal{C}([\boldsymbol{e} \mid \boldsymbol{z}_1 \mid \boldsymbol{z}_2])$: $\boldsymbol{z}_3 = [\mathbf{I} - \mathbf{Z}_{012}(\mathbf{Z}_{012}'\mathbf{Z}_{012})^{-1}\mathbf{Z}_{012}']\boldsymbol{x}_2$,où $\mathbf{Z}_{012} = [\boldsymbol{e} \mid \boldsymbol{z}_1 \mid \boldsymbol{z}_2]$
- …
- La p^e, $\boldsymbol{z}_p$, est la projection de $\boldsymbol{x}_p$ sur le complément orthogonale de $([\boldsymbol{e} \mid \boldsymbol{z}_1 \mid \ldots \mid \boldsymbol{z}_{p-1}])$.

Les colonnes $\mathbf{z}_1, \ldots, \mathbf{z}_p$ sont ensuite normalisées (divisées par leur longueur, $\sqrt{\mathbf{z}_i'\mathbf{z}_i}$).
Les colonnes de $\mathbf{Z}$ sont orthonormales (sauf pour la première, orthogonale aux autres mais non normale).

5.5 Transformation d'une variable exogène

L'idée essentielle d'un test de linéarité c'est de considérer un modèle plus général, un modèle dont la régression linéaire simple est un sous-modèle. Une possibilité, c'est de commencer avec un modèle polynomial $E(y) = \beta_o+\beta_1 x+\ldots+\beta_p x^p$ et de tester successivement les hypothèses $\beta_p = 0$, $\beta_{p-1} = 0$, etc. Un autre modèle général qu'on pourrait considérer est

$$E(y) = \beta_o + \beta_1 x + \beta_2 [x \ln(x)].$$

L'hypothèse de linéarité est alors simplement l'hypothèse que $\beta_2 = 0$. Cette transformation est motivée par l'argument suivant. On généralise la régression linéaire simple en proposant le modèle

$$E(y) = \beta_o + \gamma x^{\delta}.$$

La fonction x^{δ}, n'étant pas linéaire (en δ), on la remplace par les deux premiers termes de son développement en série par rapport à $\delta = 1$:

$$x^{\delta} = x^1 + \left(\left. \frac{dx^{\delta}}{d\delta} \right|_{\delta=1} \right)(\delta - 1) = x^1 + (x \ln x)(\delta - 1)$$

d'où

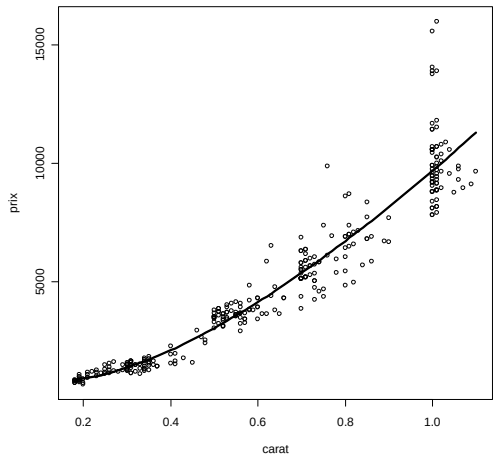
$$E(y) = \beta_o + \gamma\left[x^1 + (x \ln x)(\delta - 1)\right] = \beta_o + \gamma x^1 + \gamma(\delta - 1)(x \ln x) = \beta_o + \beta_1 x + \beta_2 x \ln(x)$$

où $\beta_1 = \gamma$ et $\beta_2 = \gamma(\delta\text{-}1)$. Si $\hat{\beta}_o, \hat{\beta}_1,$ et $\hat{\beta}_2$ sont les estimateurs habituels, on peut alors estimer γ par $\hat{\gamma} = \hat{\beta}_1$ et δ par $\hat{\delta} = \frac{\hat{\beta}_2}{\hat{\beta}_1} + 1$ et la relation estimée est $\hat{E}(y) = \hat{\beta}_o + \hat{\gamma} x^{\hat{\delta}}$. Évidemment, si on accepte— et on décide d'intégrer— l'hypothèse que $\beta_2 = 0$, on retombe sur le modèle de régression linéaire simple.

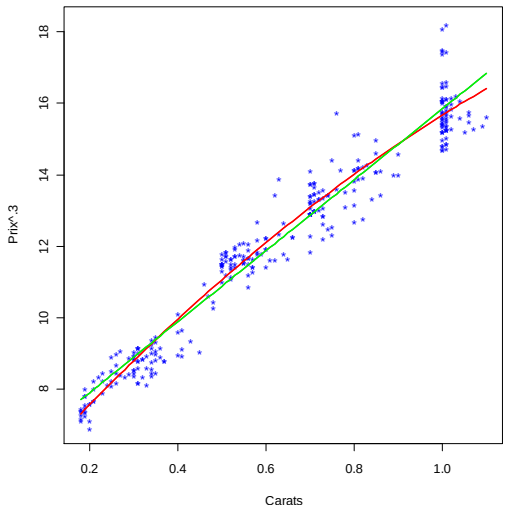
L'ajustement est bon (`clc` = carat ln(carat)):

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1900.9      537.9   3.534 0.000473 ***
carat         7794.3      514.0  15.164  < 2e-16 ***
clc           7963.0      982.8   8.103 1.31e-14 ***

Residual standard error: 1015 on 305 degrees of freedom
Multiple R-Squared: 0.9115,     Adjusted R-squared: 0.911
F-statistic:  1572 on 2 and 305 DF,  p-value: < 2.2e-16
```

Dans le graphique suivant, cependant, on voit que l'ajustement est moins bon pour les diamants exceptionnellement gros. Il serait peut-être sage d'exclure ces diamants exceptionnels et de les traiter séparément. On observe également un certain degré d'hétéroscédasticité.



Par ailleurs, on peut tenter plutôt de transformer la variable *endogène*, utilisant la méthode de Box-Cox. Celle-ci suggère la transformation $y = \text{prix}^{0,3}$. Une relation linéaire semble adéquate, mais un polynôme quadratique présente une légère amélioration, comme le montre le graphique suivant :



L'analyse suivante confirme la qualité de l'ajustement

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.9419     0.1847   26.76  < 2e-16 ***
carat        13.7413     0.6535   21.03  < 2e-16 ***
carat2       -3.0123     0.5072   -5.94 7.78e-09 ***

Residual standard error: 0.5939 on 305 degrees of freedom
Multiple R-squared: 0.956,      Adjusted R-squared: 0.9557
F-statistic:  3313 on 2 and 305 DF,  p-value: < 2.2e-16
```

C'est le modèle qu'on adopterait pour le moment, malgré le léger inconfort suscité par le fait que la transformation a été choisie par des moyens purement empiriques qui ne se justifient pas a priori.

5.6 Évaluations du modèle

Résidus standardisés

L'examen des résidus

$$\varepsilon_{i(cr)} = y_i - \hat{y}_i$$

peut être faussé par le fait que les $\varepsilon_{i(cr)}$ n'ont pas tous la même variance. Puisque le vecteur des résidus $\boldsymbol{y} - \hat{\boldsymbol{y}} = (\mathbf{I} - \mathbf{H})\boldsymbol{y}$ a pour matrice de covariance $\sigma^2(\mathbf{I} - \mathbf{H})$, on a

$$\mathrm{Var}(\varepsilon_{i(cr)}) = \sigma^2(1 - h_{ii}),$$

où h_{ii} est l'élément $(i\,;\,i)$ de la diagonale de $\mathbf{H}$. Il y a deux façons de standardiser le résidu r_i. Puisque $h_{ii} = \boldsymbol{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{x}_i$, où $\boldsymbol{x}_i'$ est la i^e ligne de $\mathbf{X}$, une définition possible est

$$\varepsilon_{i(cr)} = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - \boldsymbol{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{x}_i}}.$$

Résidus de Student

L'inconvénient de cette façon de faire est que $\hat{y}_i$ et l'estimation de σ sont faites sous la supposition que le modèle est bon et que l'observation i satisfait les conditions du modèle — alors même qu'on tente d'évaluer le modèle et la place de l'observation i dans le modèle. Une meilleure façon de calculer le résidu et de le standardiser consisterait à faire tous les calculs *sans* l'observation i. Si on exclut l'observation i des données et on dénote par $\hat{y}_{i(i)}$ la prévision de y_i à partir des données réduites, et par $\hat{\sigma}_{(i)}$ l'estimation de l'écart-type à partir des données réduites, alors le résidu standardisé — que nous appellerons *résidu de Student* — peut s'écrire

$$t_i = \textit{résidu de Student} = \frac{y_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{1 + \boldsymbol{x}_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\boldsymbol{x}_i}}$$

où $\mathbf{X}_{(i)}$ est la matrice $\mathbf{X}$ avec la i^e enlevée. La distribution d'un résidu de Student est facile à déterminer sous l'hypothèse de normalité :

$$t_i \sim t_{n-q-1}$$

Le nombre de degrés de liberté est celui de $\hat{\sigma}_{(i)}$, qui est $(n\text{-}1)\text{-}q$, puisqu'il y a une donnée enlevée.

Le calcul des résidus de Student serait très laborieux s'il fallait vraiment faire n régressions pour obtenir les n résidus. Mais cela n'est pas nécessaire, car on peut montrer que

$$t_i = \frac{r_i}{\hat{\sigma}_{(i)}\sqrt{1 + h_{ii}}}$$

et que

$$\hat{\sigma}^2_{(i)} = \frac{n-q}{n-q-1}\hat{\sigma}^2 - \frac{1}{n-q-1}\,\frac{r_i^2}{1-h_{ii}}.$$

La distance D de Cook

Certaines observations peuvent avoir une grande influence sur les résultats d'une régression, et mériteraient, par conséquent d'être surveillées de près — elles pourraient être erronées, par exemple; ou elles pourraient être exceptionnelles au point de rendre leur exclusion légitime. La distance D de Cook est une mesure de l'influence d'une observation. L'influence D_i de l'observation i est définie par

$$D_i = \frac{(\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})'(\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})/q}{\hat{\sigma}^2} = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})/q}{\hat{\sigma}^2}.$$

Dans la première expression, $\hat{\boldsymbol{y}}_{(i)}$ est le vecteur dont la i^e composante est la prédiction de y_i à partir d'une régression déterminée par toutes les observations sauf la i^e; et $\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ est le vecteur des valeurs estimées. Dans la deuxième forme, $\hat{\boldsymbol{\beta}}_{(i)}$ est l'estimateur de $\boldsymbol{\beta}$ à partir de toutes les observations sauf la i^e.

D_i peut être interprétée comme une mesure de l'écart entre l'ensemble des prédictions faites avec et sans l'observation i; ou encore, comme une mesure de la différence entre l'estimation des coefficients faite avec et sans l'observation i. À quoi comparer D_i et quand le déclare-t-on excessivement grand? Habituellement, on considère D_i comme si c'était une statistique de loi $\mathcal{F}_{q;n-q}$: si $\hat{\boldsymbol{\beta}}_{(i)}$ était un vecteur fixe, D_i serait la statistique utilisée pour tester l'hypothèse que $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{(i)}$.

Exemple [*Les analyses ci-dessous ne sont pas toutes indispensables. Certaines ne sont présentées que pour illustrer quelques procédures du logiciel*]

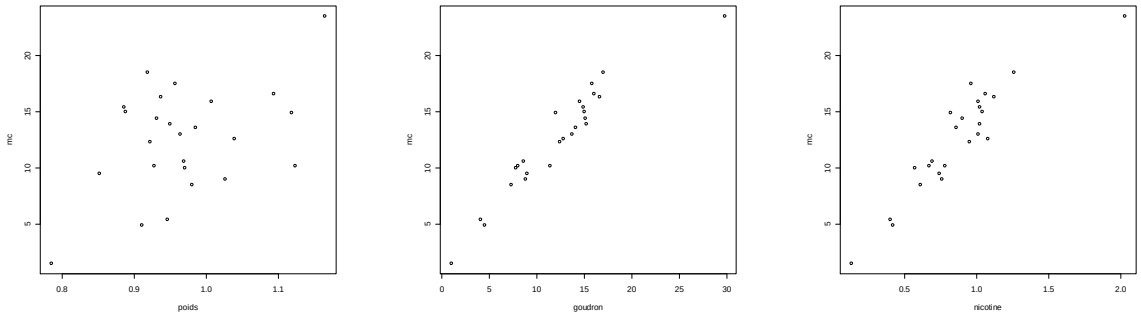
Le tableau suivant présente des données sur 25 marques de cigarettes [*Source* : Mendenhall, William, and Sincich, Terry (1992), _Statistics for Engineering and the Sciences_ (3rd ed.), New York: Dellen Publishing Co. Source originale: Federal Trade Commission, USA)]. Le but des analyses suivantes est de déterminer les facteurs qui contribuent à l'émission de monoxyde de carbone. Les variables observées sont :

`goudron` quantité de goudron, en mg
`nicotine` quantité de nicotine, en mg
`poids` poids de la cigarette, en g
`mc` monoxyde de carbone, en mg

Marque	goudron	nicotine	poids	mono
Alpine	14,1	0,86	0,9853	13,6
Benson&Hedges	16,0	1,06	1,0938	16,6
BullDurham	29,8	2,03	1,1650	23,5
CamelLights	8,0	0,67	0,9280	10,2
Carlton	4,1	0,40	0,9462	5,4
Chesterfield	15,0	1,04	0,8885	15,0
GoldenLights	8,8	0,76	1,0267	9,0
Kent	12,4	0,95	0,9225	12,3
Kool	16,6	1,12	0,9372	16,3
L&M	14,9	1,02	0,8858	15,4
LarkLights	13,7	1,01	0,9643	13,0
Marlboro	15,1	0,90	0,9316	14,4
Merit	7,8	0,57	0,9705	10,0
MultiFilter	11,4	0,78	1,1240	10,2
NewportLights	9,0	0,74	0,8517	9,5
Now	1,0	0,13	0,7851	1,5
OldGold	17,0	1,26	0,9186	18,5
PallMallLight	12,8	1,08	1,0395	12,6
Raleigh	15,8	0,96	0,9573	17,5
SalemUltra	4,5	0,42	0,9106	4,9
Tareyton	14,5	1,01	1,0070	15,9
True	7,3	0,61	0,9806	8,5
ViceroyRichLight	8,6	0,69	0,9693	10,6
VirginiaSlims	15,2	1,02	0,9496	13,9
WinstonLights	12,0	0,82	1,1184	14,9

On obtient des graphiques qui montrent l'ensemble des relations entre les différentes variables :

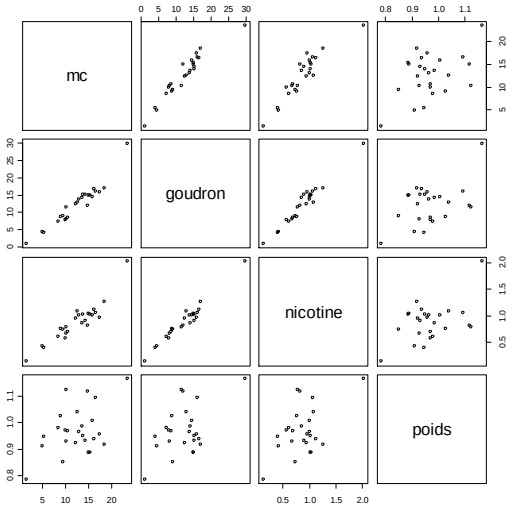
```
> plot(mc~goudron+nicotine+poids)
```



Il est évident que les variables `goudron` et `nicotine` sont fortement liées à la variable endogène `mc`; la relation est beaucoup moins forte avec le `poids`, mais elle est presque certainement significative.

Il importe aussi de connaître les relations entre les variables exogènes. Voici une commande qui fournit tous les croisements possibles :

```
> plot(data.frame(cbind(mc,goudron,nicotine,poids)))
```



Il y a de fortes dépendances entre les variables exogènes, en particulier entre nicotine et goudron. Une quantification de ces dépendances est donnée par les coefficients de corrélation :

```
> round(cor(cbind(mc,goudron,nicotine,poids)),2)
           mc goudron nicotine poids
mc       1.00    0.96     0.93  0.46
goudron  0.96    1.00     0.98  0.49
nicotine 0.93    0.98     1.00  0.50
poids    0.46    0.49     0.50  1.00
```

Ces corrélations confirment ce que les graphiques révèlent. Il est possible que ces dépendances nous causent des difficultés quand on voudra déterminer les effets de chacune sur la variable endogène. Voici donc une première analyse dans laquelle les trois variables exogènes sont présentes.

```
> reggnp<-lm(mc~goudron+nicotine+poids)
> reggnp
(Intercept)      goudron     nicotine        poids
     3.2022       0.9626      -2.6317      -0.1305
```

```
> summary(reggnp)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2022     3.4618   0.925 0.365464
goudron       0.9626     0.2422   3.974 0.000692 ***
nicotine     -2.6317     3.9006  -0.675 0.507234
poids        -0.1305     3.8853  -0.034 0.973527
Residual standard error: 1.446 on 21 degrees of freedom
Multiple R-Squared: 0.9186,     Adjusted R-squared: 0.907
F-statistic: 78.98 on 3 and 21 DF,  p-value: 1.329e-11
```

Globalement, la dépendance est fortement significative et le coefficient de détermination est très élevé. Les autres variables, dont `nicotine`, ne semblent nullement significatives. Pourtant `nicotine` a un coefficient de corrélation avec `mc` de 0,93, et sa relation avec `mc` est fortement significative lorsqu'elle est seule variable exogène (une régression avec `nicotine` seulement comme variable exogène donne un niveau de signification de l'ordre de 10-13). C'est évidemment la dépendance entre `nicotine` et `goudron` qui explique ce paradoxe.

La procédure `anova` met en évidence ces anomalies : ses résultats dépendent de l'ordre dans lequel les variables endogènes sont introduites dans le modèle. Voici ce que l'on obtient pour les différents ordres possibles (les 3 dernières lettres du nom indiquent l'ordre; par exemple, `reggnp` introduit le `goudron`, puis la `nicotine`, puis le `poids`)

```
> anova(reggnp)
Response: mc
          Df   Sum Sq  Mean Sq  F value    Pr(>F)
goudron    1   494.28   494.28 236.4843 6.651e-13 ***
nicotine   1     0.97     0.97   0.4661    0.5023
poids      1 0.002357 0.002357   0.0011    0.9735
Residuals 21    43.89     2.09
```

```
> anova(regnpg)
Response: mc
          Df    Sum Sq   Mean Sq  F value   Pr(>F)
nicotine   1    462.26    462.26 221.1620 1.27e-12 ***
poids      1 0.0004792 0.0004792   0.0002 0.988062
goudron    1     33.00     33.00  15.7892 0.000692 ***
Residuals 21     43.89      2.09
```

```
> anova(regpng)
Analysis of Variance Table

Response: mc
          Df Sum Sq Mean Sq F value    Pr(>F)
poids      1 116.06  116.06  55.526 2.522e-07 ***
nicotine   1 346.20  346.20 165.636 1.982e-11 ***
goudron    1  33.00   33.00  15.789  0.000692 ***
Residuals 21  43.89    2.09
```

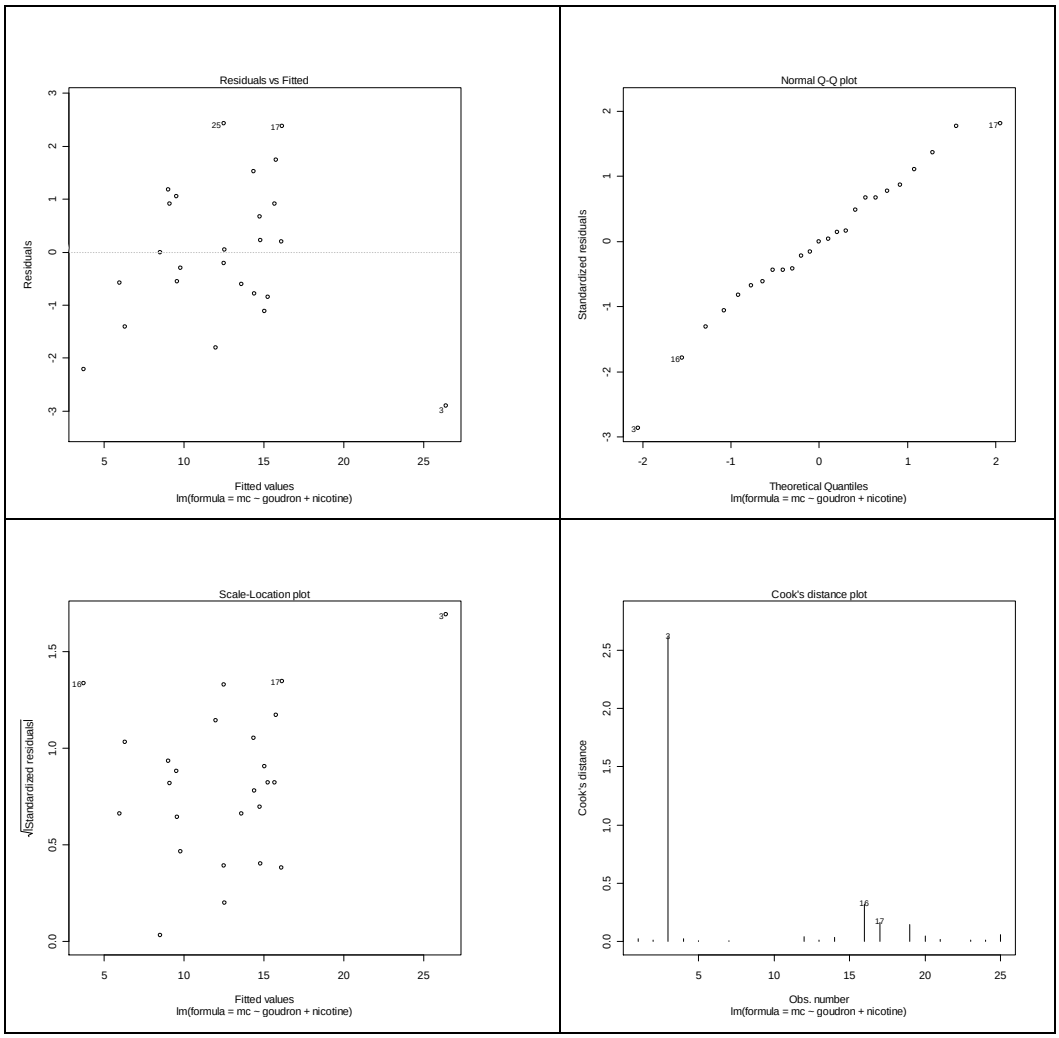
La première variable introduite est toujours significative (même la variable `poids` qui dans les autres analyses ne faisait pas … le poids.)

La variable poids, cependant, devra être éliminée. Voici ce qu'on obtient si l'on s'en tient à `goudron` et `nicotine`.

```
> reggn<-lm(mc~goudron+nicotine)
> summary(reggn)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0896     0.8438   3.662 0.001371 **
goudron       0.9625     0.2367   4.067 0.000512 ***
nicotine     -2.6463     3.7872  -0.699 0.492035

Residual standard error: 1.413 on 22 degrees of freedom
Multiple R-Squared: 0.9186,     Adjusted R-squared: 0.9112
F-statistic: 124.1 on 2 and 22 DF,  p-value: 1.042e-12
```

La variable `nicotine` est non seulement non significative : son coefficient négatif n'a pas de sens : il n'y a pas de raison scientifique de croire que le monoxyde de carbone décroît avec la nicotine. Il faudrait donc considérer le retrait de la variable `nicotine`. Avant d'examiner cette possibilité, cependant, nous montrons ce que le logiciel R peut produire en termes de graphiques diagnostiques :

```
> plot(reggn)
```



Ce qui frappe surtout c'est la présence d'une donnée aberrante, la troisième, Bull Durham. Elle s'éloigne beaucoup de la droite (résidus importants) et elle est très influente (selon le *D* de Cook). Il s'agit donc soit d'une erreur de mesure, soit d'une marque de cigarettes vraiment exceptionnelle. Peut-être y a-t-il

aussi un signe de non linéarité, puisque les résidus ont tendance à être négatifs pour les petites valeurs de cm et positives pour les grandes valeurs. En revanche, l'hypothèse de normalité ne semble pas invalidée par les données.

On peut obtenir les données nécessaires pour construire les graphiques ci-dessus à l'aide de la commande (nous n'en montrons pas les résultats)

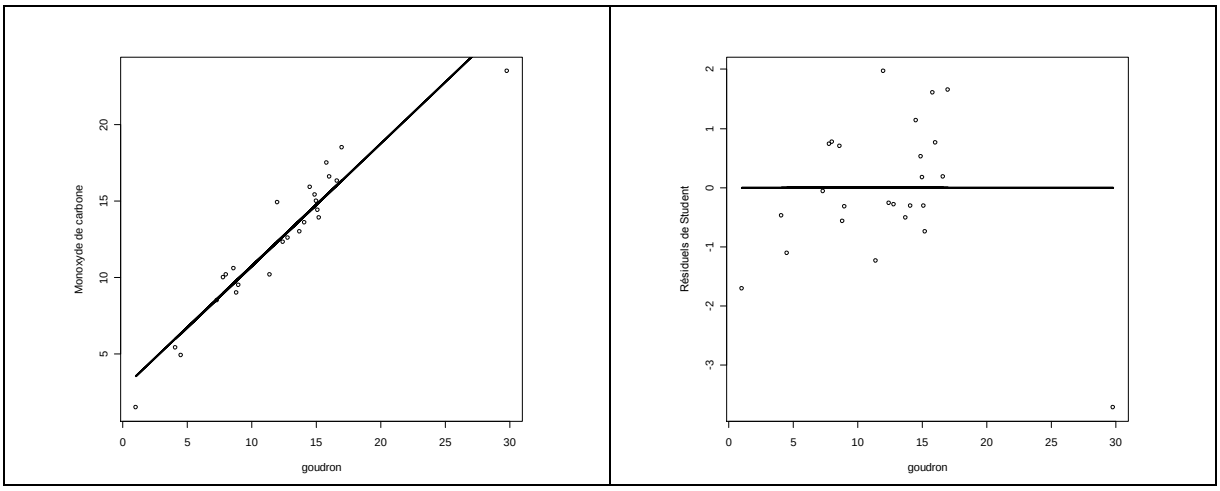
```
> ls.diag(reggn)
```

Considérons maintenant la possibilité d'éliminer la variable `nicotine`. Voici donc la régression avec `goudron` comme seule variable exogène :

```
> regg<-lm(mc~goudron)
> summary(regg)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.74328    0.67521   4.063 0.000481 ***
goudron      0.80098    0.05032  15.918 6.55e-14 ***
Multiple R-Squared: 0.9168,     Adjusted R-squared: 0.9132
F-statistic: 253.4 on 1 a  2.74328nd 23 DF,  p-value: 6.552e-14
```

Le coefficient de détermination a passé de 0,9186 (avec `goudron` et `nicotine`) à 0,9168 (avec `goudron` seul). Ce qui veut dire que `nicotine` ne contribue presque rien de plus que `goudron` à la prédiction de `mc`.

Examinons donc la qualité du modèle de régression linéaire simple. Le nuage de points ne révèle rien d'anormal (à l'exception de l'observation déjà signalée, qui correspond à un résidu important en valeur absolue.)



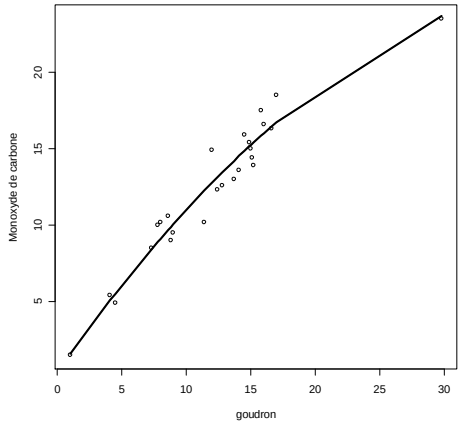
Le graphique des résidus de Student continue à présenter certains signes de non linéarité. Mais il est impossible d'en déduire une forme particulière autre que la droite. Nous considérerons plus loin une régression polynomiale.

```
Response: mc
          Df Sum Sq Mean Sq  F value    Pr(>F)
goudron    1 494.28  494.28 375.3049 7.086e-15 ***
goudron2   1  17.08   17.08  12.9722  0.001677 **
goudron3   1   0.13    0.13   0.0966  0.758997
Residuals 21  27.66    1.32
```

Il semble bien qu'un terme quadratique soit utile mais le terme cubique non. Voici les résultats d'une régression quadratique :

```
 (Intercept)      goudron     goudron2
   0.40667      1.19897     -0.01403
          Df Sum Sq Mean Sq F value    Pr(>F)
goudron    1 494.28  494.28 391.376 1.673e-15 ***
goudron2   1  17.08   17.08  13.528  0.001318 **
Residuals 22  27.78    1.26
```

Le graphique suivant fait soupçonner que c'est plutôt la valeur aberrante qui force une régression polynomiale. Le terme quadratique serait-il encore nécessaire si cette donnée était absente?



Lorsqu'on élimine la donnée aberrante on découvre que le terme quadratique n'est pas nécessaire :

```
Analysis of Variance Table
Response: mc[-3]
             Df Sum Sq Mean Sq  F value    Pr(>F)
goudron[-3]   1 386.22  386.22 298.2111 6.912e-14 ***
goudron2[-3]  1   0.33    0.33   0.2568    0.6176
```

Si on l'élimine, on obtient on trouve qu'une régression linéaire parfaitement adéquate :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.41285    0.64822    2.18   0.0403 *
goudron            0.92813    0.05283   17.57 1.96e-14 ***
Residual standard error: 1.119 on 22 degrees of freedom
Multiple R-Squared: 0.9335,     Adjusted R-squared: 0.9304
F-statistic: 308.6 on 1 and 22 DF,  p-value: 1.964e-14
```

Le critère d'information d'Akaike (AIC)

Le critère d'Akaike est une fonction de la vraisemblance maximisée L qui tient compte du nombre de paramètres dans le modèle. En voici la définition :

$$\text{AIC}(L) = -2\ln L + 2(\text{nombre de paramètres})$$

Dans un modèle linéaire où l'on suppose la normalité du vecteur $\boldsymbol{y}$, la fonction de vraisemblance, évaluée en son maximum est

$$L = \frac{1}{(2\pi)^{n/2}(\tilde{\sigma}^2)^{n/2}} e^{-(1/2)(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}})/\tilde{\sigma}^2}$$

où $\hat{\boldsymbol{\beta}}$ est l'estimateur usuel de $\boldsymbol{\beta}$ et $\tilde{\sigma}^2 = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n = (n-q)\hat{\sigma}^2/n$ est l'estimateur du maximum de vraisemblance de σ^2. Le nombre de paramètres (incluant σ^2) est q+1. Alors

$$\text{AIC}(L) = n\ln 2\pi + n\ln\tilde{\sigma}^2 + n + 2(q+1)$$

AIC étant une fonction décroissante de la vraisemblance estimée, on souhaite une valeur *petite*.

Considérons les données du tableau 4.4.1, qui présente le taux de cholestérol en fonction de l'âge. Une régression linéaire simple donne ceci :

```
> reglin<-lm(Cholest~Âge)
> summary(reglin)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  91.5716    25.6526   3.570  0.00131 **
Âge           2.5140     0.5014   5.013 2.67e-05 ***

Residual standard error: 44.14 on 28 degrees of freedom
Multiple R-squared: 0.473,      Adjusted R-squared: 0.4542
F-statistic: 25.13 on 1 and 28 DF,  p-value: 2.673e-05
```

La relation n'est pas très forte, mais elle est nettement significative, la valeur *p* étant de l'ordre de 10^{-5}. Maintenant considérons une régression polynomiale où l'on ajoute un terme quadratique en `Âge`, `Âge2`= `Âge`^{2}.

```
> reglin2<-lm(Cholest~Âge+Âge2)
> summary(reglin2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 163.71225   60.41116   2.710   0.0115 *
Âge          -0.89425    2.63805  -0.339   0.7372
Âge2          0.03568    0.02713   1.315   0.1995

Residual standard error: 43.58 on 27 degrees of freedom
Multiple R-squared: 0.5048,     Adjusted R-squared: 0.4681
F-statistic: 13.76 on 2 and 27 DF,  p-value: 7.585e-05
```

Le coefficient de détermination a en effet augmenté (de 0,473 pour le modèle linéaire à 0,5048 pour le modèle quadratique) mais une augmentation du coefficient de détermination est inévitable lorsqu'on augmente le nombre de paramètres. La mesure d'Akaike tient compte du nombre de paramètres et montre bien que le modèle quadratique n'est pas meilleur :

```
> AIC(reglin)
[1] 316.3123
> AIC(reglin2)
[1] 316.4491
```

La valeur de AIC *baisse* lorsqu'on ajoute le terme quadratique, un indice d'une qualité moindre. On n'a donc pas intérêt à l'inclure. En passant, le coefficient R^2 *ajusté* est censé tenir compte du nombre de paramètres, mais dans cet exemple il fait encore croire que le modèle quadratique est meilleur.

Remarquez par ailleurs que les valeurs *p* pour les coefficients de `Âge` et de `Âge2` sont tous deux élevés, alors qu'en fait, chacune des variables exogènes `Âge` et `Âge2` est utile à la prédiction de la variable endogène. Ce genre de phénomène se produit souvent avec une régression polynomiale, à cause de la corrélation élevée entre une variable et son carré (0,98 ici). Les résultats nous disent que si `Âge` est dans la régression, `Âge2` n'y ajoute pas grand-chose; et inversement, que si `Âge2` y est, il est inutile que `Âge2` y soit. La procédure `summary()` teste chaque variable exogène en présence des autres. La procédure `anova()`, en revanche, introduit les variables exogènes dans l'ordre présenté. Ainsi donc dans l'analyse suivante, la variable Âge est testée dans un modèle qui ne comprend pas `Âge2`. `Âge2` est ensuite testé dans le modèle complet :

```
> anova(lm(Cholest~Âge+Âge2))
Analysis of Variance Table

Response: Cholest
          Df Sum Sq Mean Sq F value    Pr(>F)
Âge        1  48976   48976   25.79 2.467e-05 ***
Âge2       1   3285    3285    1.73    0.1995
Residuals 27  51275    1899
```

Le C_p de Mallows

Considérons un modèle linéaire de q paramètres et un certain nombre restreint $p < q$ de paramètres constituant un modèle réduit. Une mesure de la perte causée par la réduction est l'indice C_p proposée par Mallows et défini par

$$C_p = \frac{\text{SCR}(p)}{\hat{\sigma}^2} + 2p - n$$

où SCR(p) est la somme de carrés résiduelle dans le modèle restreint et $\hat{\sigma}^2$ est l'estimateur de la variance dans le modèle complet.

Supposons que le modèle complet est $\boldsymbol{y} = \mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{X}_2\boldsymbol{\beta}_2+ \boldsymbol{\varepsilon}$, où $\mathbf{X}_1$ est $n\times p$ et $\mathbf{X}_2$ est $n\times(q-p)$. Si on utilise plutôt le modèle $\boldsymbol{y} = \mathbf{X}_1\boldsymbol{\theta}_1 + \boldsymbol{\varepsilon}$, la prédiction de y_i sera $\hat{y}_i = \boldsymbol{x}_{i_1}'\hat{\boldsymbol{\theta}}_1 = \boldsymbol{x}_{i_1}'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\boldsymbol{y}$. L'erreur quadratique moyenne est

$\text{EQM}(\hat{y}_i) = \sigma^2\boldsymbol{x}_{i_1}'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\boldsymbol{x}_{i_1} + \left[\boldsymbol{x}_{i_1}'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 - \boldsymbol{x}_{i_2}'\boldsymbol{\beta}_2\right]^2$.

La somme de ces EQM est $\sigma^2 p + \boldsymbol{\beta}_2'\mathbf{X}_2'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{X}_2\boldsymbol{\beta}_2$.

Étant donné que $\text{E}[\text{SCR}(p)] = (n-p)\sigma^2 + \boldsymbol{\beta}_2'\mathbf{X}_2'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{X}_2\boldsymbol{\beta}_2$, on a que

$$\frac{\sum_{i=1}^{n}\text{EQM}(\hat{y}_i)}{\sigma^2} = \frac{\text{E}[\text{SCR}(p)]}{\sigma^2} + 2p - n$$

On obtient C_p lorsqu'on remplace $\text{E}[\text{SCR}(p)]$ par $\text{SCR}(p)$ et σ^2 par $\hat{\sigma}^2$.

$C_p \le p \Leftrightarrow \text{SCR}(p) \le \text{SCR}$.

$C_p \approx p$ signifie $\text{MCR}(p) \approx \text{MCR}$.

5.7 Régression avec des données catégoriques

L'analyse de variance (dans le sens d'une analyse de résultats expérimentaux) est un cas particulier du modèle linéaire général et peut être traitée à l'aide d'une procédure de régression. Considérons les données suivantes. Elles représentent les pertes d'humidité sur le sol des forêts après la coupe du bois. Les observations sont classées selon trois traitements, dépendant de la quantité de copeaux laissés sur le sol. Les données sont présentées dans le tableau suivant:

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Pertes	1,52	1,38	1,29	1,48	1,63	1,45	1,63	1,82	1,35	1,03	2,30	2,78	2,56	3,32	2,76	2,63	2,12
Traitement	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3

Les trois moyennes μ_1, μ_2 et μ_3 peuvent être reparamétrisées ainsi: $\mu_1 = \mu$, $\mu_2 = \mu + \alpha_2$, $\mu_3 = \mu + \alpha_3$. En d'autres termes

$\text{E}(y_i)$ = μ si l'observation i a subi le traitement 1
= $\mu+\alpha_2$ si l'observation i a subi le traitement 2
= $\mu+\alpha_3$ si l'observation i a subi le traitement 3

Posons $\boldsymbol{\beta} = (\mu, \alpha_2, a_3)'$. Alors le vecteur moyenne de $\boldsymbol{y}$ est $\mathbf{X}\boldsymbol{\beta}$ où $\mathbf{X} =$

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

Nous définissons donc deux variables exogènes, x_2, et x_3, des variables dont les seules valeurs, 0 et 1, indiquent si une observation donnée appartient ou non à une classe donnée. Leurs valeurs sont celles des deux dernières colonnes ci-dessus.

ID	Pertes	Trait	alpha1	alpha2	alpha3
1	1,52	1	1	0	0
2	1,38	1	1	0	0
3	1,29	1	1	0	0
4	1,48	1	1	0	0
5	1,63	1	1	0	0
6	1,45	1	1	0	0
7	1,63	2	0	1	0
8	1,82	2	0	1	0
9	1,35	2	0	1	0
10	1,03	2	0	1	0
11	2,30	2	0	1	0
12	2,78	2	0	1	0
13	2,56	3	0	0	1
14	3,32	3	0	0	1
15	2,76	3	0	0	1
16	2,63	3	0	0	1
17	2,12	3	0	0	1

Il suffit alors de déterminer une régression multiple avec x_2 et x_3 comme variables exogènes. Voici les résultats :

```
Predictor     Coef    Stdev       t-ratio             p
Constant     1.4583   0.1841          7.92         0.000
alpha2       0.3600   0.2604          1.38         0.188
alpha3       1.2197   0.2731          4.47         0.001
s = 0.4510      R-sq = 59.6%     R-sq(adj) = 53.8%
Analysis of Variance
SOURCE       DF          SS           MS            F            p
Regression   2           4.2038       2.1019      10.33        0.002
Error        14          2.8478       0.2034
Total        16          7.0516
```

Quelle est l'hypothèse testée par la valeur $p = 0{,}002$ dans le deuxième tableau ? En régression, la table d'analyse de variance teste l'hypothèse que les coefficients des variables exogènes (ici, α_2 et α_3) sont nuls. Or l'hypothèse que $\alpha_2 = 0$ et $\alpha_3 = 0$ est précisément celle que nous voulons tester, puisqu'elle est équivalente à l'hypothèse que les trois moyennes sont égales.

5.8 Analyse de covariance

L'analyse de covariance est l'analyse d'un modèle linéaire dans lequel certaines des variables exogènes sont qualitatives, d'autres quantitatives. Dans un bon nombre des applications, l'objectif est de comparer des traitements, mais les données doivent être ajustées pour tenir compte de certaines différences initiales dans les unités expérimentales. L'exemple suivant illustre cette préoccupation.

Exemple [Morrison, Donald F. , *Applied Linear Statistical Methods*, Prentice-Hall, Englewood Cliffs, New Jersey, 1983, p. 462]. On effectue une expérience afin de déterminer l'effet d'un certain traitement sur le poids de la thyroïde des cobayes. Seize cobayes ont été répartis au hasard en deux groupes de taille 8. Le groupe expérimental a reçu un certain traitement pendant 7 jours, alors que le groupe témoin n'a reçu que de l'eau. Il a semblé nécessaire de tenir compte du poids de l'animal lui-même dans la comparaison des thyroïdes. On a donc pris note des poids des cobayes. Les données sont présentées dans le tableau suivant.

Groupe Contrôle		*Groupe Expérimental*	
Thyroïde (y, mg)	*Corps (g)*	*Thyroïde (y, mg)*	*Corps (g)*
16,0	203	17,0	210
13,0	180	16,4	199
12,0	183	14,0	189
15,4	191	15,8	186
17,4	204	17,2	216
13,0	178	18,0	209
12,0	182	14,6	191
13,8	187	17,8	211

Soit y le poids de la thyroïde, x_2 le poids du corps, et x_1 une variable définie par

$$x_{i1} = \begin{cases} 0 & \text{si l'animal est dans le groupe temoin} \\ 1 & \text{si l'animal est dans le groupe experimental} \end{cases}$$

Définissons en plus la variable

$$x_3 = x_1 x_2$$

Le modèle le plus général est $E(y_{ijk}) = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$. Mais nous examinerons le sens de plusieurs autres modèles, avant d'examiner celui-ci, dans le tableau suivant.

	Moyennes		
Espérance de y_{ijk}	*Groupe contrôle*	*Groupe expérimental*	*Signification du modèle*
$\beta_o + \beta_1 x_{i1}$	β_o	$\beta_o + \beta_1$	*Le poids du corps n'entre pas en ligne de compte. Il y a une différence, β_1, entre les deux groupes quant au poids de la thyroïde.*
$\beta_o + \beta_2 x_{i2}$	$\beta_o + \beta_2 x_2$	$\beta_o + \beta_2 x_2$	*Pour les deux groupes, le poids moyen de la thyroïde est fonction linéaire du poids du corps, avec même constante et même pente pour les deux groupes.*
$\beta_o + \beta_1 x_{i1} + \beta_2 x_{i2}$	$\beta_o + \beta_2 x_2$	$(\beta_o + \beta_1) + \beta_2 x_2$	*Le poids moyen de la thyroïde est fonction linéaire du poids du corps pour les deux groupes. La pente de la droite est la même dans les deux groupes (les droites sont donc parallèles); mais la constante diffère de β_1. Ce qui veut dire que pour un poids (du corps) donné, il y a une différence β_1 dans le poids de la thyroïde.*
$\beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$	$\beta_o + \beta_2 x_2$	$(\beta_o + \beta_1) + (\beta_2 + \beta_3) x_2$	*Le poids moyen de la thyroïde est fonction linéaire du poids du corps pour les deux groupes. La constante diffère de b_1 et la pente de b_3. Ce qui veut dire que pour un poids (du corps) donné, x_2, il y a une différence de $\beta_1 + \beta_3 x_2$ dans le poids de la thyroïde. La différence entre les deux groupes dépend donc du poids du corps. C'est la notion d'interaction entre deux variables lorsque l'une est qualitative et l'autre quantitative.*

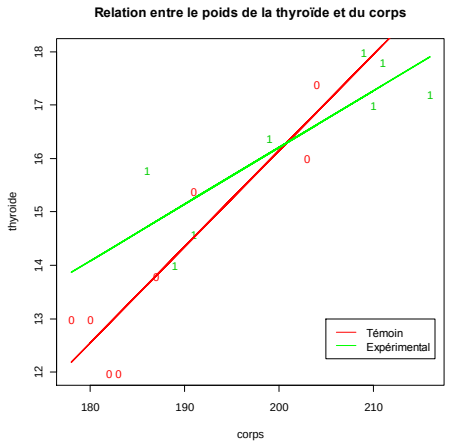
En un premier temps, considérons la régression de y sur x_1, x_2 et x_3. Voici les résultats:

```
> lm.thyro<-lm(thyroide~traitement*corps)
> summary(lm.thyro)
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -19.84972    5.97937  -3.320 0.006114 **
traitement        14.78742    8.16681   1.811 0.095283 .
corps              0.17997    0.03168   5.681 0.000102 ***
traitement:corps  -0.07364    0.04201  -1.753 0.105068

Residual standard error: 0.8465 on 12 degrees of freedom
Multiple R-squared: 0.8637,     Adjusted R-squared: 0.8296
F-statistic: 25.34 on 3 and 12 DF,  p-value: 1.770e-05
```

```
> anova(lm.thyro)
Analysis of Variance Table

Response: thyroide
                Df  Sum Sq Mean Sq F value    Pr(>F)
traitement       1 20.7025 20.7025 28.8885 0.0001668 ***
corps            1 31.5729 31.5729 44.0573 2.403e-05 ***
traitement:corps 1  2.2025  2.2025  3.0734 0.1050683
Residuals       12  8.5996  0.7166
```

Voici le graphique correspondant à ce modèle:

```
> L1<-predict(lm.thyro,data.frame(corps=corps,traitement=0))
> L2<-predict(lm.thyro,data.frame(corps=corps,traitement=1))
> plot(lm.thyro~corps,pch=as.character(traitement),col=traitement+1)
> legend(205,13,c("Témoin","Expérimental"),col=c("red","green"),lwd=.1)
> title(main="Relation entre le poids de la thyroïde et du corps",pch=.3)
```

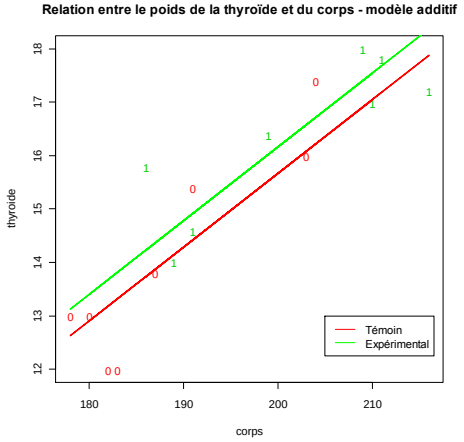


Bien que les droites se croisent, on trouve que la variable x_{12} n'est pas significative, et donc que les « vraies » droites sont parallèles. Nous allons donc supposer que les droites sont parallèles, ce qui facilite la description et l'interprétation des résultats. Nous allons donc refaire la régression, en éliminant la variable x_{12}. Voici les résultats :

```
> lm.additif<-lm(thyroide~traitement+corps)
> summary(lm.additif)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.9538     4.2349  -2.823   0.0144 *
traitement    0.4972     0.5394   0.922   0.3734
corps         0.1381     0.0224   6.164 3.41e-05 ***

Residual standard error: 0.9116 on 13 degrees of freedom
Multiple R-squared: 0.8287,     Adjusted R-squared: 0.8024
F-statistic: 31.46 on 2 and 13 DF,  p-value: 1.044e-05
```

Voici le graphique qui correspond à ce modèle:

```
> plot(thyroide~corps,pch=as.character(traitement),col=traitement+2)
> lines(L1~corps,col="red",lwd=2)
> lines(L2~corps,col="green",lwd=2)
> legend(205,13,c("Témoin","Expérimental"),col=c("red","green"),lwd=.1)
> title(main="Relation entre le poids de la thyroïde et du corps - modèle additif",pch=.3)
```



La variable x_1 n'est pas significative : malgré le graphique, les deux droites sont confondues: il n'y a pas de différence entre les deux groupes. La variable x_2, cependant, est significative : il était effectivement nécessaire de tenir compte du poids du corps lorsqu'on compare les poids de la thyroïde.

Il est intéressant de noter que la différence entre les deux groupes aurait été trouvée significative si on n'avait pas tenu compte des poids du corps. Les résultats d'un test comparant les traitements, sans ajustement, le montrent :

```
> pairwise.t.test(thyroide,traitement,pooled.sd=T,paired=F)
        Pairwise comparisons using t tests with pooled SD
data:  thyroide and traitement
  0
1 0.020
```

C'est aussi ce que nous aurions conclu si notre analyse de variance avait placé le traitement avant le poids corporel :

```
> anova(lm.additif)
Analysis of Variance Table

Response: thyroide
           Df  Sum Sq Mean Sq F value    Pr(>F)
traitement  1 20.7025 20.7025  24.915 0.0002466 ***
corps       1 31.5729 31.5729  37.997 3.409e-05 ***
Residuals  13 10.8021  0.8309
```

5.9 Test d'ajustement à une droite

Revenons à la régression linéaire. Dans la plupart des cas, l'hypothèse que l'espérance $E(y_i)$ est une fonction linéaire $\beta_o + \beta_1 x_i$ des x_i est supposée vraie sans autre évidence que le nuage des points. Dans certains cas, cependant, il est possible de soumettre cette supposition à un test statistique. C'est le cas où nous avons des observations répétées pour une même valeur de x. Supposons que l'ensemble des valeurs de x est

$$\begin{aligned} x_{11} = \ldots = x_{1n_1} &= x_1 \\ x_{21} = \ldots = x_{2n_2} &= x_2 \\ \ldots \quad \ldots \quad \ldots \quad &\ldots \\ x_{k1} = \ldots = x_{kn_k} &= x_k \end{aligned}$$

Il y a en tout $n = \Sigma n_i$ observations, mais seulement k valeurs distinctes $x_1, \ldots, x_k$ de x. Pour maintenir une notation cohérente nous affectons les y d'un double indice, comme les x correspondants, sauf que, contrairement aux x, nous ne supposons pas que des y avec le même premier indice sont égaux. Avec ces doubles indices, le modèle de régression s'écrit

$$\mathcal{M}_o: \; y_{ij} = \beta_o + \beta_1 x_{ij} + \varepsilon_{ij}$$

Mais pour tester cette supposition de linéarité, nous commençons par un modèle plus général, dans lequel nous n'imposons pas la linéarité, soit

$$\mathcal{M}: \; y_{ij} = \mu_i + \varepsilon_{ij}$$

Or le modèle $\mathcal{M}$ est le modèle d'analyse de variance à un facteur et le modèle $\mathcal{M}_o$ est le modèle de régression linéaire simple. Nous pouvons, *dans le modèle* $\mathcal{M}$, tester l'hypothèse que le modèle $\mathcal{M}_o$ s'applique, c'est-à-dire, l'hypothèse linéaire

$$\mu_i = \beta_o + \beta_1 x_i$$

Le rapport F pour tester cette hypothèse aura pour numérateur une somme de carrés expliquée exprimée comme la différence de deux somme de carrés résiduelles, $SCE = SCR_o - SCR$, où SCR est simplement la somme des carrés résiduelle dans le modèle $\mathcal{M}$ et SCR_o est la somme des carrés résiduelle dans le modèle réduit par l'hypothèse nulle, soit le modèle $\mathcal{M}_o$. Nous avons déjà des formules pour ces sommes de carrés :

$$SCR = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

$$SCR_o = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \hat{\beta}_o - \hat{\beta}_1 x_i)^2$$

où $\hat{\beta}_o$ et $\hat{\beta}_1$ sont les estimateurs de β_o et β_1 définis au chapitre 4. Quelques manipulations algébriques permettent d'écrire la différence $SCR_o - SCR$ de la manière instructive suivante :

$$SCR_o - SCR = \sum_{i=1}^{k} n_i (\bar{y}_i - \hat{\beta}_o - \hat{\beta}_1 x_i)^2$$

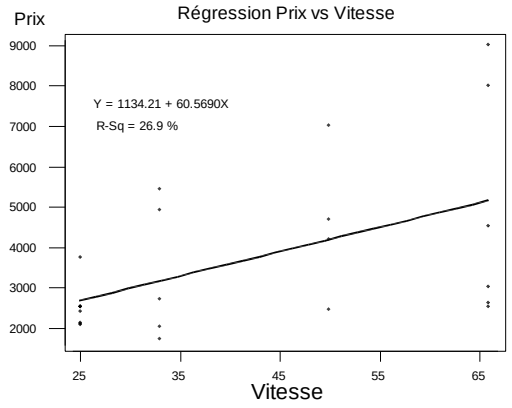
Cette somme de carrés devrait être petite si l'hypothèse $\mu_i = \beta_o + \beta_1 x_i$ est vraie, car $\bar{y}_i$ estime μ_i et $\hat{\beta}_o + \hat{\beta}_1 x_i$ estime $\beta_o + \beta_1 x_i$. Le nombre de degrés de liberté est n-k pour SCR et n-2 pour SCR_o. Donc $SCR_o - SCR$ a k-2 degrés de liberté (ce qui s'explique: la somme a k termes et 2 paramètres estimés). La statistique F est donc

$$F = \frac{[SCR_o - SCR]/(k-2)}{SCR/(n-k)} \sim F_{k-2;n-k}$$

Exemple On a prélevé les données suivantes sur 24 ordinateurs afin d'analyser la relation entre la vitesse de l'ordinateur et sont prix.

ID	*Vitesse (mhz)*	*Prix ($)*	*ID*	*Vitesse (mhz)*	*Prix ($)*
1	25	2 045 $	13	33	4 898 $
2	25	2 069 $	14	33	5 428 $
3	25	2 100 $	15	50	2 432 $
4	25	2 394 $	16	50	4 178 $
5	25	2 499 $	17	50	4 678 $
6	25	2 499 $	18	50	6 995 $
7	25	2 499 $	19	66	2 495 $
8	25	2 515 $	20	66	2 600 $
9	25	3 720 $	21	66	2 999 $
10	33	1 708 $	22	66	4 499 $
11	33	1 999 $	23	66	7 995 $
12	33	2 699 $	24	66	8 999 $

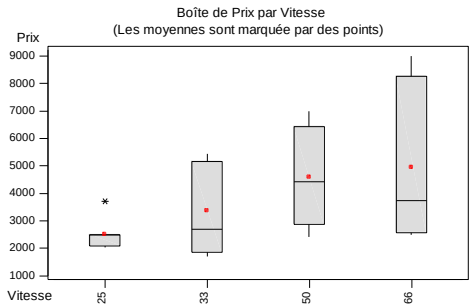
Le graphique suivant montre qu'il y a une certaine relation. Elle est plutôt faible, mais elle existe. La relation est-elle linéaire ?



Nous faisons d'abord une analyse descriptive parallèle au raisonnement du test formel qui suivra. Nous commençons par adopter une modèle qui fait le moins d'hypothèses possibles, c'est-à-dire, on suppose seulement que les moyennes des 4 groupes (25 mhz, 33 mhz, 50 mhz, et 66 mhz) sont μ_1, μ_2, μ_3 et μ_4, sans aucune restriction sur les μ_i. On estime ces moyennes par les moyenne échantillonnales, qui sont

$$\bar{y}_1 = 2482{,}222;\ \bar{y}_2 = 3346{,}4;\ \bar{y}_3 = 4570{,}75;\ \bar{y}_4 = 4931{,}167.$$

L'hypothèse de linéarité est l'hypothèse que $\mu_i = \beta_o + \beta_1 x_i$, $i = 1, 2, 3, 4$, c'est-à-dire, que les 4 moyennes se situent sur une droite. Le graphique suivant présente les moyenne $\bar{y}_i$ ainsi qu'une boîte qui résume les données dans chaque classe :



Le modèle de régression linéaire suppose que $\mu_i = \beta_o + \beta_1 x_i$ et fournit une estimation des coefficients:

$$\text{Prix} = \hat{\beta}_0 + \hat{\beta}_1 x_i = 1134 + 60{,}6x_i$$

Dans ce modèle, l'estimation des moyennes des 4 groupes est:

$\hat{\mu}_1 = \hat{\beta}_o + \hat{\beta}_1 x_1 = 1134+60{,}6(25) = 2648{,}432;$

$\hat{\mu}_2 = \hat{\beta}_o + \hat{\beta}_1 x_2 = 1134+60{,}6(33) = 3132{,}984;$

$\hat{\mu}_3 = \hat{\beta}_o + \hat{\beta}_1 x_3 = 1134+60{,}6(50) = 4162{,}657;$

$\hat{\mu}_4 = \hat{\beta}_o + \hat{\beta}_1 x_4 = 1134+60{,}6(66) = 5131{,}760.$

Nous devons donc comparer les deux séries d'estimation, celles basées sur le modèle d'analyse de variance (les $\bar{y}_i$) et celles basées sur le modèle de régression (les $\hat{\beta}_o + \hat{\beta}_1 x_i$):

Vitesse (i)	25 *mhz*	33 *mhz*	50 *mhz*	66 *mhz*
Effectif (n_i)	9	5	4	6
Modèle d'anova ($\bar{y}_i$)	2482,222	3346,4	4570,75	4931,167
Modèle de régression ($\hat{\beta}_o + \hat{\beta}_1 x_i$)	2648,432	3132,984	4162,657	5131,760

La statistique pour tester l'hypothèse de linéarité est

$$F = \frac{[\text{SCR}_o - \text{SCR}]/(k-2)}{\text{SCR}/(n-k)}$$

qui suit une loi $F_{k\text{-}2;n\text{-}k}$ sous l'hypothèse de linéarité. Appliquant l'une des formules de SCR_o-SCR, nous obtenons

$\text{SCR}_o - \text{SCR} = \sum_{i=1}^{k} n_i(\bar{y}_i - \hat{\beta}_o - \hat{\beta}_1 x_i)^2$
$= 9(2482{,}22\text{-}2648{,}43)^2 + 5(3346{,}4\text{-}3132{,}98)^2 + 4(4570{,}75\text{-}4162{,}66)^2 + 6(4931{,}17\text{-}5131{,}76)^2$
$= 1\ 383\ 951.$

Le nombre de degrés de liberté est k-2 = 4 -2 = 2.

Quant à SCR, c'est la somme des carrés des écarts entre les observations et leur moyenne estimée, soit
$\text{SCR} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$

$$= \sum_{i=1}^{9}(y_{1j} - \bar{y}_1)^2 + \sum_{i=1}^{5}(y_{2j} - \bar{y}_2)^2 + \sum_{i=1}^{4}(y_{3j} - \bar{y}_3)^2 + \sum_{i=1}^{6}(y_{4j} - \bar{y}_4)^2$$

$= 2\ 049\ 806 + 11\ 659\ 489 + 10\ 616\ 995 + 41\ 223\ 625 = 65\ 549\ 914$

Le nombre de degrés de liberté de SCR est n-k = 24-4 = 20.
Donc la valeur de F est

$$F = \frac{[\mathrm{SCR}_o - \mathrm{SCR}]/(k-2)}{\mathrm{SCR}/(n-k)} = \frac{[66933866 - 65549914]/(4-2)}{65549914/(24-4)} = 0{,}21,$$

ce qui, à 2 et 20 degrés de liberté, est non significatif : on ne rejette pas l'hypothèse de linéarité.
Ceci complète le test ■

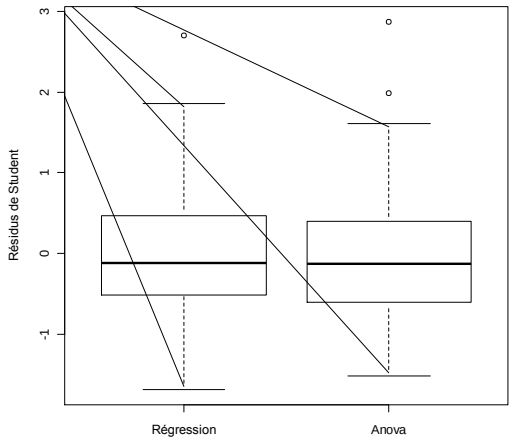
Remarque. *Il existe une autre façon d'interpréter la statistique* SCR_o - SCR. *Nous savons que*.

$$\mathrm{SCR}_o - \mathrm{SCR} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \hat{\beta}_o - \hat{\beta}_1 x_i)^2 - \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

Ceci explique pourquoi cette différence figure au numérateur de la statistique F. Si le modèle de régression est incorrect, les estimations $\hat{\beta}_o + \hat{\beta}_1 x_i$ *s'éloigneraient des observations plus que ne le feraient les* $\bar{y}_i$ *; par conséquent les résidus seraient importants et* SCR_o *serait bien plus grand que* SCR. *Mais on peut aussi montrer que*

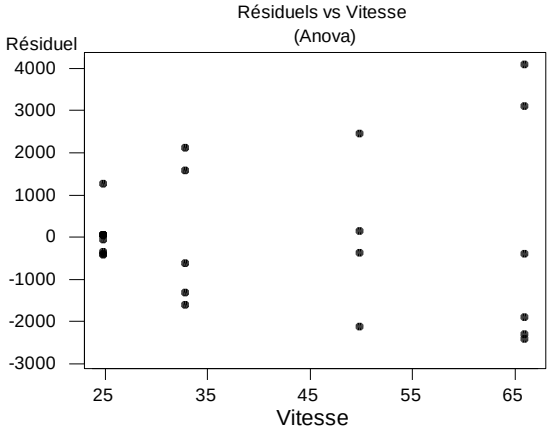
$$\mathrm{SCR}_o - \mathrm{SCR} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left((y_{ij} - \hat{\beta}_o - \hat{\beta}_1 x_i) - (y_{ij} - \bar{y}_i)\right)^2$$

Ceci signifie que le test est basé sur une comparaison de sommes résiduelles : la somme sous le modèle de régression et celle sous un modèle d'analyse de variance. Une comparaison visuelle montre pourquoi on ne rejette pas l'hypothèse de linéarité : les résidus des deux modèles sont comparables :



■

Remarque *Si la figure ci-dessus ne présente pas d'évidence de non linéarité, elle suggère en revanche que les variances varient avec la vitesse. Le graphique suivant le montre encore. On verra maintenant que dans le cas présent il est possible de tester l'hypothèse d'homoscédasticité.*


Résiduels vs Vitesse
(Anova)
Résiduel
4000
3000
2000
1000
0
-1000
-2000
-3000
25
35
45
55
65
Vitesse