

Chapitre I

Régression linéaire simple

Licence 3 MIA SHS - Université de Bordeaux

Marie Chavent

On cherche à **modéliser** la relation entre le **prix d'un appartement** et sa **surface**.

On pose :

- y = prix en euros/1000,
- x = surface en m^2 .

On suppose que cette **relation est linéaire** de la forme :

$$y = \beta_0 + \beta_1 x$$

- On veut **estimer** cette relation appelée **droite de régression théorique**.
- On utilise un **échantillon** de n appartements dont on connaît le prix et la surface

```

#=====
# Prix des appartements en fonction de la surface
#=====
prix<-c(130,280,268,500,320,250,378,250,350,300,155,245,200,325,85,78,375,200,270,85)
surface<-c(28,50,55,110,60,48,90,35,86,65,32,52,40,70,28,30,105,52,80,20)
apparts <- data.frame(prix,surface)
apparts

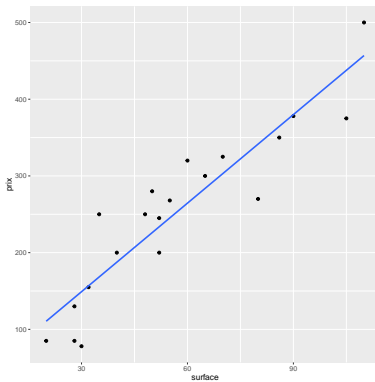
```

```

##      prix surface
## 1     130      28
## 2     280      50
## 3     268      55
## 4     500     110
## 5     320      60
## 6     250      48
## 7     378      90
## 8     250      35
## 9     350      86
## 10    300      65
## 11    155      32
## 12    245      52
## 13    200      40
## 14    325      70
## 15     85      28
## 16     78      30
## 17    375     105
## 18    200      52
## 19    270      80
## 20     85      20

```

```
library(ggplot2)
ggplot(apparts, aes(x=surface, y=prix)) +
  geom_point() +
  geom_smooth(method=lm, se=FALSE)
```



```
coef(lm(prix ~ surface))
```

```
## (Intercept)    surface
##          33.6         3.8
```

On cherche à modéliser la relation entre **deux variables quantitatives continues**.

Un **modèle de régression linéaire simple** est de la forme suivante :

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (1)$$

où :

- y est la **variable à expliquer** (à valeurs dans \mathbb{R});
- x est la **variable explicative** (à valeurs dans \mathbb{R});
- ε est le **terme d'erreur aléatoire** du modèle;
- β_0 et β_1 sont deux paramètres à estimer.

Commentaires :

- La désignation “**simple**” fait référence au fait qu’il n’y a qu’une seule variable explicative x pour expliquer y .
- La désignation “**linéaire**” correspond au fait que le modèle (1) est linéaire en β_0 et β_1 .

Pour n observations, on peut écrire le modèle de régression linéaire simple sous la forme :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (2)$$

Dans ce chapitre, on suppose que :

- ε_i est une variable *aléatoire*, non observée,
- x_i est observée et *non aléatoire*,
- y_i est observée et *aléatoire*.

On fait les trois **hypothèses additionnelles** suivantes :

$$(A1) \mathbb{E}[\varepsilon_i] = 0, \forall i = 1, \dots, n,$$

ou de manière équivalente :

$$\mathbb{E}[y_i] = \beta_0 + \beta_1 x_i, \forall i = 1, \dots, n.$$

Commentaire sur l'hypothèse (A1) : elle indique que **les erreurs sont centrées** ce qui implique que y_i dépend seulement de x_i et que les autres sources de variations de y_i sont aléatoires.

$$(A2) \quad \mathbb{V}(\varepsilon_i) = \sigma^2, \quad \forall i = 1, \dots, n,$$

ou de manière équivalente :

$$\mathbb{V}(y_i) = \sigma^2, \quad \forall i = 1, \dots, n.$$

Commentaires sur l'hypothèse (A2) :

- On parle d'hypothèse d'**homoscédasticité** (\simeq homogénéité des variances).
- Cette variance est supposée **constante et indépendante de x_i** .
- Cette variance σ^2 est un **paramètre du modèle qu'il faudra estimer**.

$$(A3) \quad \text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad \forall i \neq j$$

ou de manière équivalente :

$$\text{Cov}(y_i, y_j) = 0, \quad \forall i \neq j.$$

Commentaire sur l'hypothèse (A3) :

- Sous cette hypothèse, **les termes d'erreur ε_i sont non corrélés** .
- Lorsque l'on rajoutera une **hypothèse de normalité** sur les ε_i , les erreurs ε_i seront alors **indépendantes**.

On peut écrire **matriciellement** le modèle (2) de la manière suivante :

$$Y = X\beta + \varepsilon \quad (3)$$

où

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \text{et} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

- Y désigne le vecteur à expliquer de taille $n \times 1$,
- X la matrice explicative de taille $n \times 2$,
- ε le vecteur d'erreurs de taille $n \times 1$.

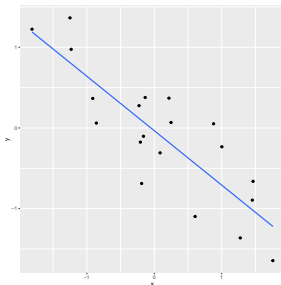
Exercice : Trouver X et Y pour les données sur les appartements.

Exemple. Données simulées à partir du modèle $y = -x + \varepsilon$.

```
library(ggplot2)
n <- 20; sigma2 <- 0.5; eps <- rnorm(n,0,sigma2)
x <- rnorm(n,0,1)
y <- -x+eps
data.frame(x,y)
```

```
##           x           y
## 1    1.454 -0.895
## 2    0.086 -0.309
## 3   -0.913  0.366
## 4   -1.233  0.975
## 5   -0.188 -0.688
## 6   -1.250  1.364
## 7    1.003 -0.234
## 8    0.249  0.068
## 9   -0.135  0.378
## 10  -0.161 -0.103
## 11  0.879  0.052
## 12 -0.206 -0.176
## 13 -0.226  0.276
## 14  0.606 -1.097
## 15  1.759 -1.645
## 16  1.276 -1.363
## 17 -0.859  0.060
## 18  1.466 -0.662
## 19  0.218  0.370
## 20 -1.813  1.224
```

```
library(ggplot2)
ggplot(data.frame(x,y), aes(x=x, y=y)) + geom_point() + geom_smooth(method=lm,se=FALSE)
```



```
coef(lm(y ~ x))
```

```
## (Intercept)          x
##      -0.034      -0.674
```

Exercice : tapez ce code dans R et regardez ce qui se passe

- lorsque vous simulez plusieurs échantillons. Représentez alors les valeurs estimées de β_0 et β_1 dans des boxplots.
- lorsque la valeur de σ^2 varie, lorsque la valeur de n varie.

2. Estimation des paramètres β_0 , β_1 et σ^2

A partir de l'échantillon (aléatoire) de n observations

$$\{(x_i, y_i), i = 1, \dots, n\},$$

on veut **estimer** les paramètres

$$\beta_0, \beta_1 \text{ et } \sigma^2.$$

- Pour estimer β_0 et β_1 , on peut utiliser la **méthode des moindres carrés** qui ne nécessite pas d'hypothèse supplémentaire sur la distribution de ε_i (ou de y_i), contrairement à la **méthode du maximum de vraisemblance** (que l'on peut aussi utiliser) qui est fondée sur la **normalité** de ε_i (ou de y_i).
- La méthode des moindres carrés **ne fournit pas** un estimateur de σ^2 .

Estimation de β_0 et β_1 par les moindres carrés

On cherche $\hat{\beta}_0$ et $\hat{\beta}_1$ qui minimisent la somme des **erreurs quadratiques**

$$\epsilon_i^2 = (y_i - \beta_0 - \beta_1 x_i)^2$$

On doit donc résoudre le **problème d'optimisation** suivant :

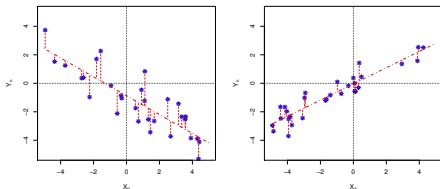
$$(\hat{\beta}_0, \hat{\beta}_1) = \text{Arg} \min_{(\beta_0, \beta_1) \in \mathbb{R}^2} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2. \quad (4)$$

Vocabulaire :

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ est appelé la **valeur prédite**.
- $\hat{\epsilon}_i = y_i - \hat{y}_i$ est appelé le **résidu**.

Interprétation graphique

Graphiquement, $\hat{\beta}_0$ et $\hat{\beta}_1$ sont construits pour minimiser les distances verticales entre les observations (y_n) et la **droite de régression théorique** $y = \beta_0 + \beta_1 x$. Nous avons représenté ces distances sur les figures ci-dessous.



La droite d'équation $y = \hat{\beta}_0 + \hat{\beta}_1 x$ est la **droite de régression estimée** sur le nuage de points

Résolution du problème d'optimisation

Le problème d'optimisation est :

$$\min_{(\beta_0, \beta_1)} F(\beta_0, \beta_1),$$

avec $F(\beta_0, \beta_1) = \sum_{i=1}^n \{y_i - (\beta_0 + \beta_1 x_i)\}^2$.

Le minimum est atteint pour

$$\begin{cases} \left. \frac{\partial F(\beta_0, \beta_1)}{\partial \beta_0} \right|_{\beta_0 = \hat{\beta}_0, \beta_1 = \hat{\beta}_1} = 0, \\ \left. \frac{\partial F(\beta_0, \beta_1)}{\partial \beta_1} \right|_{\beta_0 = \hat{\beta}_0, \beta_1 = \hat{\beta}_1} = 0, \end{cases}$$

soit après quelques calculs :

$$\begin{cases} -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \\ -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0. \end{cases}$$

Solution du problème d'optimisation

On en déduit après quelques manipulations :

$$\left\{ \begin{array}{l} \hat{\beta}_1 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2} = \frac{c_{x,y}}{s_x^2}, \\ \hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n. \end{array} \right.$$

où $c_{x,y}$ est la covariance empirique entre les x_i et les y_i et s_x^2 est la variance empirique des x_i .

Commentaires

- Le minimum de F est égal à $\sum_{i=1}^n \hat{\epsilon}_i^2$. Ce minimum est appelé la **somme des carrés des résidus (SCR)**.
- La valeur prédite \hat{y}_i estime $\mathbb{E}[y_i] = \beta_0 + \beta_1 x_i$ et non pas y_i . Une meilleure notation serait $\mathbb{E}[y_i]$.
- Aucune des hypothèses (A1), (A2) et (A3) n'a été utilisée ici pour obtenir les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$.

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

Sous les hypothèses (A1), (A2) et (A3), on peut montrer que

- $\mathbb{E}[\hat{\beta}_0] = \beta_0$,
- $\mathbb{E}[\hat{\beta}_1] = \beta_1$,
- $\mathbb{V}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x}_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \right)$,
- $\mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}$.

Commentaires

- Les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ sont **sans biais**.
- Ils sont aussi **de variance minimale** parmi tous les estimateurs linéaires (par rapport à y_1, \dots, y_n) sans biais (propriété dite de Gauss-Markov).

Estimation de σ^2

Le paramètre σ^2 est défini par

$$\sigma^2 = \mathbb{V}(\varepsilon_i) = \mathbb{V}(y_i) = \mathbb{E} [(y_i - \mathbb{E}[y_i])^2].$$

En prenant $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ comme estimateur de $\mathbb{E}[y_i]$, il apparaît naturel d'**estimer** σ^2 par

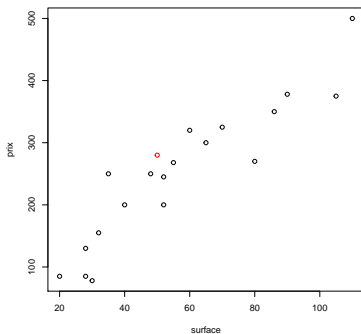
$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^n (\hat{\varepsilon}_i)^2}{n-2} = \frac{SCR}{n-2}.$$

Commentaires

- s^2 est un estimateur **sans biais** de σ^2
- La perte de deux degrés de liberté dans l'expression de s^2 est le "coût" de l'estimation de β_0 et de β_1 nécessaire pour obtenir les \hat{y}_i .

Exemple de données réelles : les appartements Parisiens.

```
x <- appart$surface  
y <- appart$prix  
plot(x,y,xlab='surface',ylab='prix',col=ifelse(x==50, "red", "black"))
```



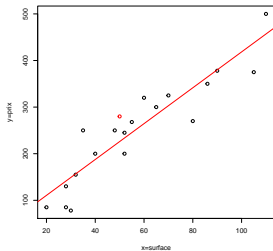
Sorties R

```
mod <- lm(y ~ x) #fonction linear model
names(mod)

## [1] "coefficients" "residuals" "effects" "rank" "fitted.values" "assign"
## [7] "qr" "df.residual" "xlevels" "call" "terms" "model"

summary(mod)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
## Min 1Q Median 3Q Max
## -71.47 -27.63 4.75 24.96 81.68
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.644 24.445 1.38 0.19
## x 3.848 0.392 9.81 1.2e-08 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45 on 18 degrees of freedom
## Multiple R-squared: 0.842, Adjusted R-squared: 0.834
## F-statistic: 96.3 on 1 and 18 DF, p-value: 1.2e-08
```



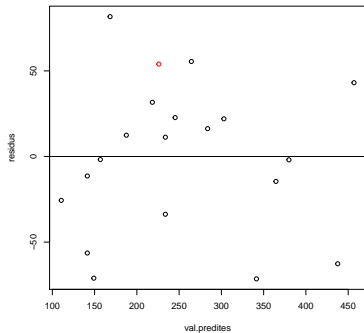
```
round(data.frame(y,val.predites=fitted(mod),residus=resid(mod))[1:5,],digit=2)
```

```
##      y val.predites residus
## 1 130          141      -11
## 2 280          226       54
## 3 268          245       23
## 4 500          457       43
## 5 320          265       55
```

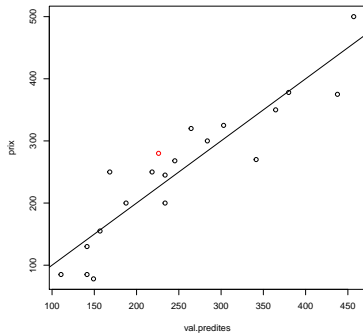
```
n <- 20
sqrt(sum(resid(mod)^2)/(n-2)) #residual standard error (square root of SCR)
```

```
## [1] 45
```

Graphique croisant les valeurs prédites \hat{y}_i et les résidus $\hat{\varepsilon}_i = y_i - \hat{y}_i$



Graphique croisant les valeurs prédites \hat{y}_i et les valeurs observées y_i



3. Test d'hypothèses et intervalle de confiance pour β_1

Typiquement, les hypothèses portant sur β_1 ont plus d'intérêt que celles portant sur β_0 . On va donc se limiter à tester la nullité de la pente β_1 (**absence de liaison linéaire entre x et y**) :

$$\mathcal{H}_0 : \beta_1 = 0 \quad \text{contre} \quad \mathcal{H}_1 : \beta_1 \neq 0$$

Pour faire ce test, il est nécessaire de faire une **hypothèse supplémentaire** :

$$(A4) \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

ou de manière équivalente

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

Commentaire. L'unique "nouveau" ici est la **normalité**.

Nouvelles propriétés pour les estimateurs $\hat{\beta}_1$ et s^2

Sous les hypothèses (A1)-(A4), on a :

$$(a) \hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\right);$$

$$(b) \frac{(n-2)s^2}{\sigma^2} \sim \chi^2(n-2);$$

(c) $\hat{\beta}_1$ et s^2 sont indépendants.

Commentaires. La propriété (a) est facile à établir. Les propriétés (b) et (c) seront démontrées ultérieurement.

Un rappel de probabilité

Si $U \sim \mathcal{N}(0, 1)$, $V \sim \chi^2(\nu)$ et U est indépendant de V , alors $\frac{U}{\sqrt{\frac{V}{\nu}}} \sim T(\nu)$.

On déduit alors des propriétés (a)-(c) que

$$\frac{\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}}}{\sqrt{\frac{(n-2)s^2}{\sigma^2}}}{n-2} = \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2}} \sim T(n-2).$$

Commentaire. On peut remarquer que le dénominateur $s/\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2}$ est un estimateur de $\sqrt{\mathbb{V}(\hat{\beta}_1)}$, l'écart-type de $\hat{\beta}_1$.

On utilisera la **statistique** suivante :

$$T_n = \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2}},$$

qui est distribuée selon **une loi de Student** à $n - 2$ degrés de liberté.

Test de \mathcal{H}_0 contre \mathcal{H}_1

Sous l'hypothèse $\mathcal{H}_0 : \beta_1 = 0$, on a

$$T_n = \frac{\hat{\beta}_1}{s / \sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2}} \sim T(n-2). \quad (5)$$

Pour une hypothèse alternative $\mathcal{H}_1 : \beta_1 \neq 0$ bilatérale, on rejette \mathcal{H}_0 avec un risque $0 \leq \alpha \leq 1$ si

$$|t| \geq t_{n-2, 1-\alpha/2}$$

où t est la réalisation de T_n et $t_{n-2, 1-\alpha/2}$ est le fractile d'ordre $1 - \alpha/2$ de la loi $T(n-2)$.

Commentaire. Pour réaliser ce test, on peut également regarder la *p-valeur* aussi appelée niveau de signification du test : si *p-valeur* $\leq \alpha$, on rejette \mathcal{H}_0 . Dans le cas d'un test bilatéral ($\mathcal{H}_1 : \beta_1 \neq 0$), on a :

$$p\text{-valeur} = \mathbb{P}(|T_n| > |t| / \mathcal{H}_0). \quad (6)$$

On rejette \mathcal{H}_0 si *p-valeur* $\leq \alpha$

Intervalle de confiance pour β_1 au niveau de confiance $1 - \alpha$:

L'intervalle de confiance de β_1 est :

$$[\hat{\beta}_1 \pm t_{n-2, 1-\alpha/2} \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}].$$

Commentaire. On **rejette** \mathcal{H}_0 si 0 n'appartient pas à cet intervalle.

Exemple des données appartements.

```
summary(mod)$coefficients

##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)    33.6      24.44    1.4 1.9e-01
## x              3.8       0.39    9.8 1.2e-08

qt(0.975,18) # quantile loi Student

## [1] 2.1

confint(mod)

##           2.5 % 97.5 %
## (Intercept)  -18  85.0
## x            3   4.7
```

Table d'analyse de la variance (ANOVA) : On complète souvent l'étude en construisant la table d'ANOVA.

Source de variation	Somme des carrés	ddl	carré moyen	F
régression (expliquée)	$SCE = \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2$	1	$\sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2$	$\frac{SCE}{SCR/(n-2)}$
Résiduelle	$SCR = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	n-2	$\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$	
Totale	$SCT = \sum_{i=1}^n (y_i - \bar{y}_n)^2$	n-1	$\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y}_n)^2$	

Commentaire. La statistique F , dite **statistique de Fisher**, permet de tester $\mathcal{H}_0 : \beta_1 = 0$ contre $\mathcal{H}_1 : \beta_1 \neq 0$.

On rejette \mathcal{H}_0 si

$$F > f_{1, n-2, 1-\alpha}$$

où $f_{1, n-2, 1-\alpha}$ est le fractile d'ordre $1 - \alpha$ d'une loi $F(1, n - 2)$.

Commentaires.

- Le carré d'une variable de Student à ν degrés de libertés est une variable de Fisher à $(1, \nu)$ degrés de libertés.
- En régression linéaire simple, le test de Fisher issu de l'ANOVA est donc le même que le test de student pour tester la nullité de β_1 .
- En régression linéaire multiple, la table d'ANOVA et le test de Fisher permettront de tester la nullité simultanée des p coefficients des p variables explicatives soit $\mathcal{H}_0 : \beta_1 = \dots = \beta_p = 0$.

Exemple des données appartements.

```
anova(mod)

## Analysis of Variance Table
##
## Response: y
##          Df Sum Sq Mean Sq F value Pr(>F)
## x          1 195068   195068   96.3 1.2e-08 ***
## Residuals 18  36477    2026
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. Coefficient de détermination

Le coefficient de détermination R^2 est défini par

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2}{\sum_{i=1}^n (y_i - \bar{y}_n)^2} = \frac{\text{variabilité expliquée (SCE)}}{\text{variabilité totale (SCT)}} = 1 - \frac{SCR}{SCT}$$

Remarque. On a la formule “classique” de l’analyse de la variance nous donnant la décomposition suivante :

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

variabilité totale = variabilité résiduelle + variabilité expliquée

Commentaire. Le coefficient R^2 donne la proportion de variabilité de y qui est expliquée par le modèle. Plus le R^2 est proche de 1, meilleure est l’adéquation du modèle aux données.

```
summary(mod)$r.squared
```

```
## [1] 0.84
```

5. Prédiction d'une valeur ultérieure

On désire **prévoir** à l'aide du modèle la valeur de la variable y **pour une valeur non observé** x_0 de x .

D'après le modèle on a $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$, où y_0 et ε_0 sont des variables aléatoires. La prédiction naturelle est alors :

$$\hat{y}_0 = \widehat{\mathbb{E}[y_0]} = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

L'erreur de prédiction est définie par $\hat{y}_0 - y_0$ et on peut montrer que sous les hypothèses du modèle (incluant l'hypothèse de normalité), on a :

$$\hat{y}_0 - y_0 \sim \mathcal{N} \left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \right) \right). \quad (7)$$

On en déduit que :

$$\frac{y_0 - \hat{y}_0}{\sigma \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}} \sim \mathcal{N}(0, 1).$$

On peut montrer que :

$$\frac{y_0 - \hat{y}_0}{s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}} \sim T(n - 2).$$

On utilise ce résultat pour construire un **intervalle de prédiction** pour y_0 , c'est à dire l'intervalle $[A, B]$ tel que

$$\mathbb{P}(A \leq y_0 \leq B) = 1 - \alpha.$$

Ici, y_0 est une variable aléatoire et non pas un paramètre. L'intervalle de prédiction est donc un **intervalle dans lequel une future observation y_0 va tomber avec une certaine probabilité** (différent d'un intervalle de confiance).

On en déduit l'**intervalle de prédiction** pour y_0 au niveau de confiance $1 - \alpha$ suivant :

$$\left[\hat{y}_0 \pm t_{n-2, 1-\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}} \right]$$

Commentaires. La variance de l'erreur de prévision dépend

- de la variabilité intrinsèque σ^2 de la variable (aléatoire) y_0 ,
- de la variabilité due à "l'imprécision" des estimations de β_0 et β_1 dans la formule de régression.

Cette source de variabilité peut être réduite (en augmentant la taille de l'échantillon par exemple), contrairement à la première source de variabilité.

On peut aussi construire un **intervalle de confiance** de la valeur moyenne

$$\mathbb{E}[y_0] = \beta_0 + \beta_1 x_0,$$

qui est cette fois un paramètre. On va donc chercher l'**intervalle aléatoire** $[A, B]$ tel que

$$\mathbb{P}(A \leq \mathbb{E}[y_0] \leq B) = 1 - \alpha.$$

Pour construire cet intervalle, on montre que :

$$\hat{y}_0 \sim \mathcal{N} \left(\beta_0 + \beta_1 x_0, \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \right) \right), \quad (8)$$

$$\frac{\hat{y}_0 - \beta_0 + \beta_1 x_0}{s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}} \sim T(n-2). \quad (9)$$

On en déduit l'**intervalle de confiance** de $\mathbb{E}[y_0]$ suivant :

$$\left[\hat{y}_0 \mp t_{n-2, 1-\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}} \right].$$

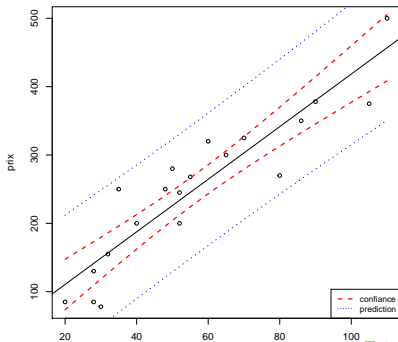
Exemple des données appartements.

```
x0 <- 50  
predict(mod,data.frame(x=x0),interval="prediction")
```

```
##   fit lwr upr  
## 1 226 129 323
```

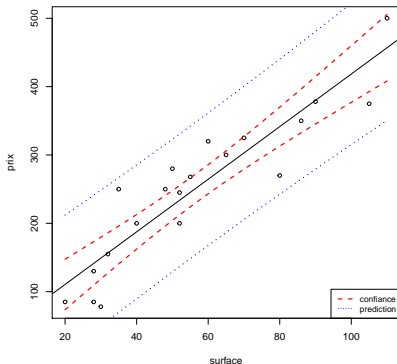
```
predict(mod,data.frame(x=x0),interval="confidence")
```

```
##   fit lwr upr  
## 1 226 204 248
```



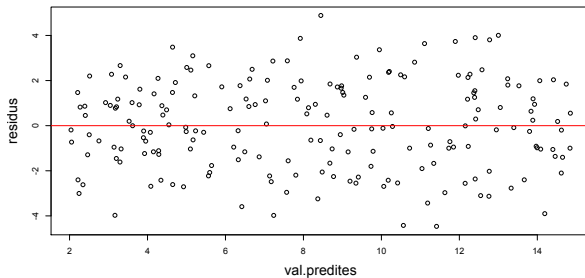
Code R pour obtenir ce graphique.

```
seqx <- seq(min(x),max(x),length=50)
intpred <- predict(mod,data.frame(x=seqx),interval="prediction")[,c("lwr","upr")]
intconf <- predict(mod,data.frame(x=seqx),interval="confidence")[,c("lwr","upr")]
plot(y~x,xlab="surface",ylab="prix",cex=0.8)
abline(mod)
matlines(seqx,cbind(intconf,intpred),lty=c(2,2,3,3), col=c("red","red","blue","blue"),lwd=c(2,2))
legend("bottomright",lty=c(2,3),lwd=c(2,1), c("confiance","prediction"),col=c("red","blue"),cex=0.8)
```

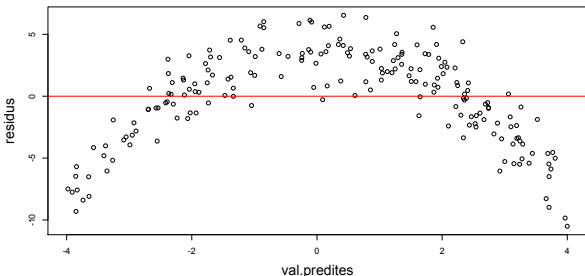


Quelques graphiques permettant de “vérifier visuellement” des hypothèses sous-jacentes.

- Graphique croisant les valeurs prédites \hat{y}_i et les résidus $\hat{\varepsilon}_i = y_i - \hat{y}_i$:



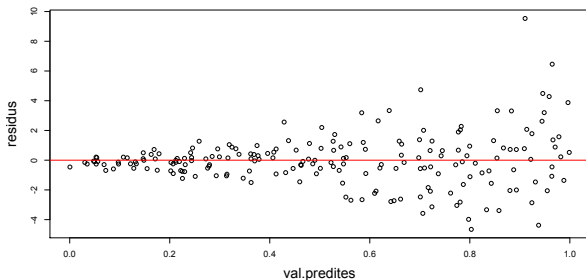
On observe un “comportement aléatoire” et “une variance constante”.



On observe un **"structure évidente"** dans les résidus (qui ne sont plus vraiment aléatoires).

↔ Il faut "changer" de modèle pour essayer de prendre en compte cette structure.

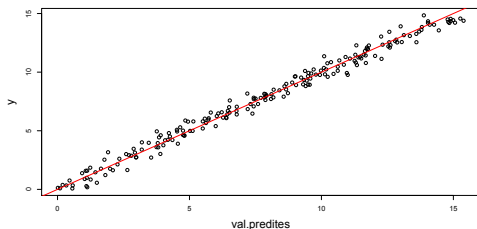
(Par exemple rajouter un terme quadratique x^2 dans la partie explicative du modèle).



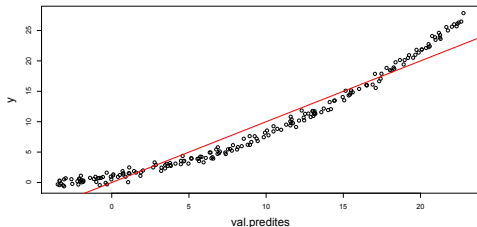
On observe que "la variance des résidus n'est pas constante", elle augmente clairement en fonction de \hat{y}_i (elle dépend donc des x_i). Il n'y a donc pas homoscedasticité.

↪ Il faut "changer" de modèle pour prendre en compte cette hétéroscédasticité.

- Graphique croisant les valeurs prédites \hat{y}_i et les valeurs observées y_i :



Les points s'alignent sur la première bissectrice :
l'adéquation du modèle aux données est correcte.



On voit ici clairement apparaître une structure non linéaire :
il y a une mauvaise adéquation du modèle.

↪ Il faut changer de modèle.

Normalité des résidus.

La **théorie** sous-jacente à l'inférence du modèle (tests d'hypothèses , intervalles de confiance et de prédiction) suppose **la normalité du terme d'erreur ε_i** .

Il convient donc de tester cette hypothèse *a posteriori* en utilisant les résidus du modèle : $\{\hat{\varepsilon}_i, i = 1, \dots, n\}$. Pour cela, on peut faire un **test de normalité de Shapiro-Wilk**.

```
residus <- resid(mod)
shapiro.test(residus)

##
## Shapiro-Wilk normality test
##
## data:  residus
## W = 1, p-value = 0.7
```

Dans l'exemple des appartements, en prenant un risque de première espèce de 5%, on accepte la normalité des résidus ($p\text{-value}=0.5177 > \alpha = 5\%$). Les tests d'hypothèses sont donc "valides" ainsi que les intervalles de confiance.

On peut aussi faire un **examen graphique** de la normalité des résidus.

Résidus **standardisés** : on divise $\hat{\epsilon}_i$ par son écart-type (estimé) :

$$\hat{\epsilon}_i^* = \frac{\hat{\epsilon}_i}{s\sqrt{1-h_{ii}}}$$

avec $h_{ii} = \frac{1}{n} + \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$

Parfois appelé résidus **studentisés** (interne) car comme $\hat{\epsilon}_i$ suit une loi normale, on peut montrer que $\hat{\epsilon}_i^* \sim T(n-2)$ et pour n assez grand on pourra considérer que $\hat{\epsilon}_i^* \sim \mathcal{N}(0, 1)$.

