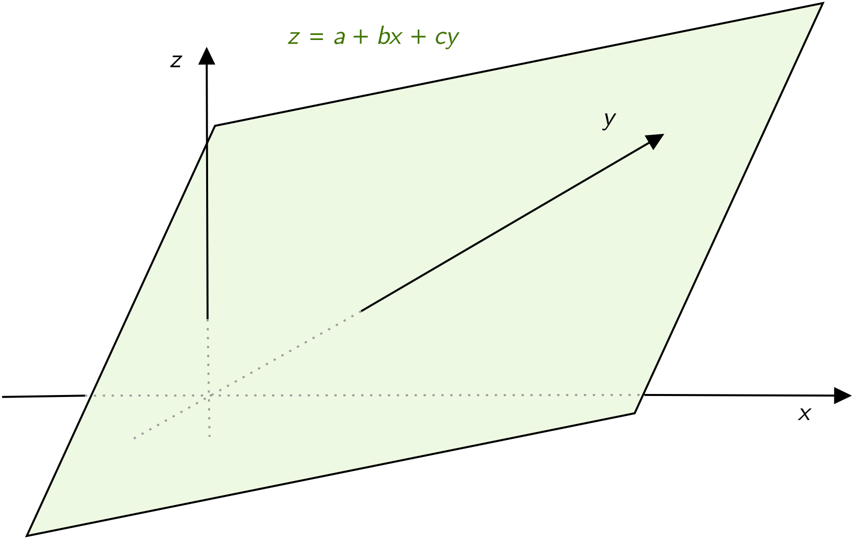
Statistiques pour les sciences (MAT-4681)

Arthur Charpentier

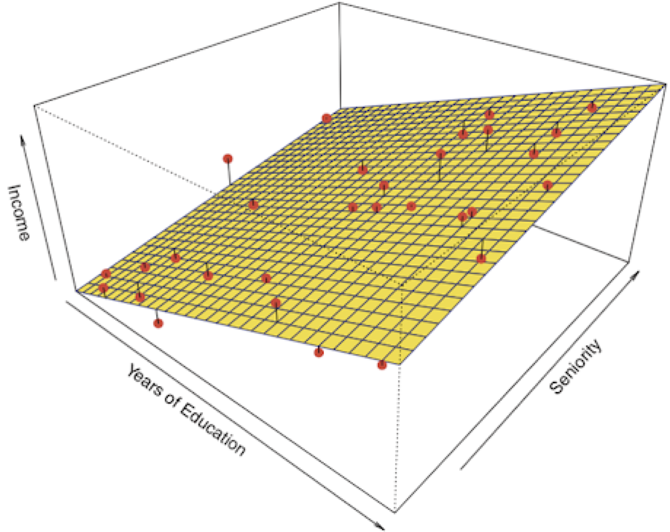
16 - Régression multiple

été 2022

Plan (dans l'espace, $\mathbb{R}^3$)



Plan (dans l'espace, $\mathbb{R}^3$)



Moindres carrés

Soit $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_{1,1}, x_{2,1}, y_1), \cdots, (x_{1,n}, x_{2,n}, y_n)\}$ un échantillon de trois variables. On suppose que

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$$

- y est la variable d'intérêt (que l'on veut prédire)
- x_1 et x_2 sont deux variables explicatives (possibles)

On va chercher le plan qui passe au mieux dans le nuage de points,

$$\min_{\alpha,\beta}\left\{\sum_{i=1}^n \varepsilon_i^2\right\} = \min_{\beta_0,\beta_1,\beta_2}\left\{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{1,i} - \beta_2 x_{2,i})^2\right\}$$

Moindres carrés

Plan de régression, moindres carrées (OLS)

Soit $(\boldsymbol{x},\boldsymbol{y}) = \{(x_{1,1},x_{2,1},y_1),\cdots,(x_{1,n},x_{2,n},y_n)\}$ un échantillon. Le plan de régression qui minimise la somme des carrés des erreurs est $y = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + +\widehat{\beta}_2 x_2$ où

$$(\widehat{\beta}_0,\widehat{\beta}_1,\widehat{\beta}_2) = \underset{\beta_0,\beta_1,\beta_2}{\text{argmin}}\left\{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{1,i} - \beta_2 x_{2,i})^2\right\}$$

Note il existe une unique solution à ce programme d'optimisation

Note pour donner les valeurs des paramètres $(\widehat{\beta}_j)$ on va devoir passer par une représentation matricielle (cf MAT105 - 201-NYC)

Matrices et vecteurs

Soient $m, n \geq 1$. Une matrice $\boldsymbol{A}$ de taille (m, n) à coefficients réels est un tableau de nombres réels ayant m lignes et n colonnes. On note également par $(\boldsymbol{A})_{ij}$ ou plus simplement A_{ij} l'élément sur la ligne i et sur la colonne j de $\boldsymbol{A}$.

Example:

$$\boldsymbol{A} = \begin{pmatrix} 1.5 & 2 & 3.1 & 8 \\ -1 & 4 & 5 & 6.5 \end{pmatrix}$$

$\boldsymbol{A}$ est de taille (2×4) et par exemple $A_{13} = 3.1$.
Une matrice ne contenant qu'une colonne est appelée un vecteur et une matrice ne contenant qu'une ligne est un vecteur ligne. Par exemple $\boldsymbol{x} = \begin{pmatrix} 1.5 \\ -1 \end{pmatrix}$ et $\boldsymbol{y} = (1.5\ 2\ 3.1\ 8)$ sont respectivement de taille $(2, 1)$ et $(1, 4)$.

Transposée

Soit $\boldsymbol{A}$ une matrice réelle de taille (m, n). La matrice transposée notée $\boldsymbol{A}^\top$ de taille (n, m) est définie par $(\boldsymbol{A}^\top)_{ij} = A_{ji}$ pour $i = 1, \dots, n$ et $j = 1, \dots, m$.
Et $(\boldsymbol{A}^\top)^\top = \boldsymbol{A}$.

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}^\top = \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix}$$

$$\|\boldsymbol{x}\|^2 = \boldsymbol{x}^\top \boldsymbol{x} = \sum_{i=1}^n x_i^2$$

$$\boldsymbol{X}^\top \boldsymbol{X} = \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top$$

```
> t(1:4) %*% rep(1,4)
     [,1]
[1,]   10
> (1:4) %*% t(rep(1,4))
     [,1] [,2] [,3] [,4]
[1,]    1    1    1    1
[2,]    2    2    2    2
[3,]    3    3    3    3
[4,]    4    4    4    4
> t(1:4) %*% (1:4)
     [,1]
[1,]   30
> (1:4) %*% t(1:4)
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    2    4    6    8
[3,]    3    6    9   12
[4,]    4    8   12   16
```

Transposée

Pour $\boldsymbol{a}$ et $\boldsymbol{b}$, de dimension n,

$$\boldsymbol{a}^\top \boldsymbol{b} = \boldsymbol{b}^\top \boldsymbol{a} = \sum_{i=1}^{n} a_i b_i$$

```
> a = c(1,2,3,4)
> b = c(7,2,4,1)
> t(a) %*% b
     [,1]
[1,]   27
```

L'espérance peut s'éecrire sous cette forme

$$\mathbb{E}[X] = \sum_{i=0}^{n} x_i p_i = \boldsymbol{x}^\top \boldsymbol{p}$$

```
> x = 0:6
> p = dbinom(0:6, 6, 1/3)
> t(x) %*% p
     [,1]
[1,]    2
```

$$(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top$$

$$(\boldsymbol{A}+\boldsymbol{B})^\top = \boldsymbol{A}^\top + \boldsymbol{B}^\top$$

Une matrice carrée $\boldsymbol{A}$ de taille (n, n) est dite symétrique si $\boldsymbol{A} = \boldsymbol{A}^\top$.

Produit ★★★

Si $\boldsymbol{A}$ et $\boldsymbol{B}$ sont (respectivement) des matrices $k \times m$ et $m \times n$,

$$\boldsymbol{C}_{ij} = \boldsymbol{A}_{i\cdot}^{\top} \boldsymbol{B}_{\cdot j} = A_{i1}B_{1j} + \cdots + A_{im}B_{mj} = \sum_{k=1}^{m} A_{ik}B_{kj},$$

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{pmatrix} \cdot \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1p} \\ B_{21} & B_{22} & \cdots & B_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ B_{m1} & B_{m2} & \cdots & B_{mp} \end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1p} \\ C_{21} & C_{22} & \cdots & C_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \cdots & C_{np} \end{pmatrix}$$

```
> A = matrix(1:6,2,3)
> B = matrix(1:12,3,4)
> A %*% B
     [,1] [,2] [,3] [,4]
[1,]   22   49   76  103
[2,]   28   64  100  136
```

Le produit matriciel n'est pas commutatif pour deux matrices quelconque de même taille: $\boldsymbol{AB} \neq \boldsymbol{BA}$

Produit et inverse ★★★

Soit $\mathbb{I}_n$ la matrice de taille (n, n) composée de 1 sur la diagonale et de 0 ailleurs. Alors, pour $\boldsymbol{A}$ de taille (n, n), $\mathbb{I}_n$ est l'élément neutre tel que $\boldsymbol{A}\mathbb{I}_n = \mathbb{I}_n\boldsymbol{A} = \boldsymbol{A}$.
Soient $\boldsymbol{A}$, $\boldsymbol{B}$ et $\boldsymbol{C}$ trois matrices réelles de dimension concordante, alors

- $(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC})$ (associativité du produit)
- $\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}$ (distributivité du produit)
- $(\boldsymbol{AB})^\top = \boldsymbol{B}^\top\boldsymbol{A}^\top$.

Inverse matricielle

Soit $\boldsymbol{A}$ une matrice carrée de taille (n, n) dont le déterminant est non nul, alors $\boldsymbol{A}$ est dite non singulière et il existe une matrice inverse (de même taille) notée $\boldsymbol{A}^{-1}$ vérifiant $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \mathbb{I}_n$

Produit et inverse ★★★

Note le déterminnant ne sera pas redéfini ici

Soient $\boldsymbol{A}$ et $\boldsymbol{B}$ deux matrices inversibles de taille (n, n) alors

- $(\boldsymbol{A}^{-1})^{\top} = (\boldsymbol{A}^{\top})^{-1}$

($\boldsymbol{A}^{-1}$ est symétrique ssi $\boldsymbol{A}$ est symétrique)

- $(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$.
- $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$
- $\det(\boldsymbol{A}^{-1}) = \dfrac{1}{\det(\boldsymbol{A})}$

```
> A = matrix(
  c(3,2,4,3),2,2)
> A
     [,1] [,2]
[1,]    3    4
[2,]    2    3
> solve(A)
     [,1] [,2]
[1,]    3   -4
[2,]   -2    3
> A %*% solve(A)
     [,1] [,2]
[1,]    1    0
[2,]    0    1
```

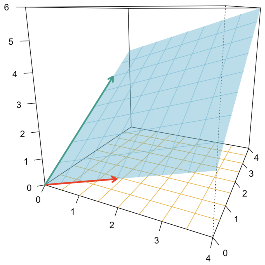
Espace vectoriel engendré ★★★

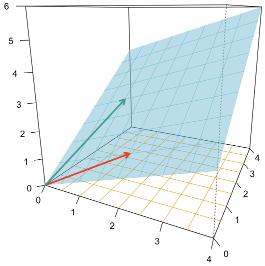
Soient $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \in \mathbb{R}^n$, on définit $\mathcal{V}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ comme

$$\left\{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} = \sum_{i=1}^{p} a_i \boldsymbol{x}_i = \boldsymbol{X}\mathbf{a}, \mathbf{a} \in \mathbb{R}^p\right\}$$

où $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p]$ est une matrice $n \times p$.

La dimension de $\mathcal{V}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ est le rang de $\boldsymbol{X}$.





Régression Linéaire

Nous supposons que les données collectées suivent le modèle suivant

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i,$$

où

- x_{ij} sont des nombres déterministes (connus). β_0 représente la constante (`intercept` dans les logiciels). On notera souvent $x_{i0} = 1$.
- β_j, $j = 0, 1, \ldots, k$ paramètres réels à estimer. On pose $p = k + 1$
- les variables ε_i sont des fluctuations aléatoires (erreur de mesures, mauvaise spécification du modèle,...).

On peut reformuler le modèle en:

$$\underbrace{\boldsymbol{y}}_{(n\times 1)} = \underbrace{\boldsymbol{X}}_{(n\times p)} \underbrace{\boldsymbol{\beta}}_{(p\times 1)} + \underbrace{\boldsymbol{\varepsilon}}_{(n\times 1)}$$

Régression Linéaire

On peut reformuler le modèle en: $\underbrace{\boldsymbol{y}}_{(n\times 1)} = \underbrace{\boldsymbol{X}}_{(n\times p)} \underbrace{\boldsymbol{\beta}}_{(p\times 1)} + \underbrace{\boldsymbol{\varepsilon}}_{(n\times 1)}$ où

$1 \leq p = 1 + k \leq n$ et

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$\boldsymbol{\beta} \in \mathbb{R}^p$ et $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})^\top$. Lorsque $p = 2$, ce modèle correspond au modèle de régression linéaire **simple**. On notera

- $\boldsymbol{x}_j$ le vecteur de taille $(n \times 1)$ des n observations de la $j$ème covariable.
- $\boldsymbol{x}_i^\top$ le vecteur de taille $(1 \times p)$ des valeurs des p covariables pour l'individu i.
- $\boldsymbol{y}$ le vecteur réponse; $\boldsymbol{\varepsilon}$ vecteur aléatoire (centré sans perte de généralité).

Moindres carrés ★★★

Formellement,

$\mathcal{H}_1$: La matrice de design $\boldsymbol{X}$ est de plein rang.

$p \leq n$, $\mathcal{H}_1 \Rightarrow \text{rang}(\boldsymbol{X}) = p$, $\boldsymbol{X}^\top\boldsymbol{X}$ de taille (p,p) est symétrique, définie positive et donc **inversible**.

$\mathcal{H}_2$: Les erreurs sont centrées, de même variance et non corrélées $\Leftrightarrow \mathbb{E}(\varepsilon_i) = 0$ et $\text{Var}(\varepsilon_i) = \sigma^2$.

$\mathcal{H}_2^{\mathcal{N}}$: Les erreurs sont indépendantes et de même loi $\mathcal{N}(0,\sigma^2)$

$\mathcal{H}_3$: La matrice de design $\boldsymbol{X}$ est telle que lorsque $n \to \infty$, $\frac{1}{n}(\boldsymbol{X}^\top\boldsymbol{X}) \to \boldsymbol{Q}$ où $\boldsymbol{Q}$ est une matrice définie positive

- $\mathcal{H}_1$: permet de démontrer l'existence de $\widehat{\boldsymbol{\beta}}$.
- $\mathcal{H}_2$: permet de démontrer des propriétés pour $\widehat{\boldsymbol{\beta}}$ (sans biais, calcul de variance).
- $\mathcal{H}_2^{\mathcal{N}}$: permet de faire des tests.
- $\mathcal{H}_3$: permet de démontrer la convergence de $\widehat{\boldsymbol{\beta}}$.

Moindres carrés

Plan de régression, moindres carrées (OLS)

Soit $(\boldsymbol{X},\boldsymbol{y}) = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ un échantillon. Le plan de régression qui minimise la somme des carrés des erreurs est $y = \boldsymbol{x}^\top\widehat{\beta}$ où

$$\widehat{\beta} = \underset{\beta_0,\beta_1,\ldots,\beta_k}{\text{argmin}} \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top\beta)^2$$

Sous l'hypothèse $\mathcal{H}_1$, l'estimateur des moindres carrées existe et vaut

$$\widehat{\beta} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}.$$

Note il existe une unique solution à ce programme d'optimisation

Coefficient de détermination R^2

R^2

Soit $(\boldsymbol{X},\boldsymbol{y}) = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ un échantillon, et $\widehat{y}_i$ la prévision par régression linéaire. Alors

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

R_a^2 (R^2 ajusté)

Soit $(\boldsymbol{X},\boldsymbol{y}) = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ un échantillon, et $\widehat{y}_i$ la prévision par régression linéaire. Alors

$$R_a^2 = 1 - \frac{\text{SCR}/(n-k-1)}{\text{SCT}/(n-1)} = 1 - \frac{n-1}{n-k-1}\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Coefficient de détermination R^2

R_a^2 et R^2

$R_a^2 \leq R^2$ et

$$R_a^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-k-1}$$

On pénalise ici les modèles trop complexes, avec trop de variables explicatives.

Prévision et résidus

Prévision et résidus

Soit $(\boldsymbol{X},\boldsymbol{y}) = \{(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_n,y_n)\}$ un échantillon. Le plande régression qui minimise la somme des carrés des erreurs est $y = \boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}$. La différence entre la valeur observée y_i et la valeur prédite $\widehat{y}_i = \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}$ s'appelle le résidu $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$.

Résidus

Soient $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$ les résidus estimés. Les résidus sont centrés et leur variance σ^2 est estimée par s^2 où

$$s^2 = \frac{1}{n-p}\sum_{i=1}^n \widehat{\varepsilon}_i^2 \text{ et } \widehat{\varepsilon}_i = 0.$$

Regression

```
> import numpy as np
> import statsmodels.api as sm
> x = np.array([[5,1], [15,4], [25,-5], [35,4],
    [45,-2], [55,2]])
> x = sm.add_constant(x)
> y = np.array([5, 20, 14, 32, 22, 38])
> model = sm.OLS(y, x)
> results = model.fit()
> print(results.summary())
=====================================================
            coef.  std err       t   P>|t|    [0.025   0.975]
-----------------------------------------------------
const     4.0581   3.370   1.204   0.315    -6.668   14.785
x1        0.5578   0.097   5.770   0.010     0.250    0.865
x2        1.5604   0.508   3.071   0.055    -0.057    3.178
-----------------------------------------------------
Dep. Variable:   y        R-squared:         0.931
Model:           OLS      Adj. R-squared:    0.886
                          F-statistic:       20.37
```

Regression

```
> df = data.frame(x1 = c(5, 15, 25, 35, 45, 55),
                  x2 = c(1, 4,-5, 4,-2, 2),
                  y  = c(5, 20, 14, 32, 22, 38))
> model = lm(y~x1+x2, data=df)
> summary(model)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.05810    3.37049   1.204   0.3149
x1           0.55783    0.09667   5.770   0.0103 *
x2           1.56037    0.50818   3.071   0.0545 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Residual standard error: 4.037 on 3 degrees of freedom
Multiple R-squared: 0.9314, Adjusted R-squared: 0.8857
F-statistic: 20.37 on 2 and 3 DF,  p-value: 0.01796
```

Propriétés de $\hat{\beta}$ ★★★

Propriétés de $\hat{\beta}$

Soit $(\boldsymbol{X}, \boldsymbol{y}) = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ un échantillon. Sous les hypothèses $\mathcal{H}_1$ et $\mathcal{H}_2$,

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta} \text{ et } \text{Var}[\widehat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$$

(admis)

On a vu (section 14) que si $y = \alpha + \beta x$ (droite de régression)

$$s^2_{\widehat{\beta}} = \frac{\frac{1}{n-2}\sum_{i=1}^n \widehat{\varepsilon}_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = s^2\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^{-1}$$

et ici $s^2_{\widehat{\beta}_j} = s^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{jj}$.

Propriétés de $\hat{\beta}$

```
> library(DALEX)
> reg = lm(m2.price~construction.year+surface+no.rooms
    , data=apartments)
> vcov(reg)
             (Intercept) const.year surface  no.rooms
(Intercept)  3550207.63   -1807.629 152.913 -3277.951
const.year     -1807.629      0.921  -0.080     1.171
surface          152.913     -0.080   2.562   -64.046
no.rooms       -3277.951      1.171 -64.046  1922.300
> summary(reg)

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) 6295.7095  1884.1995   3.341 0.000865 ***
const.year    -0.8829     0.9599  -0.920 0.357920
surface       -9.3827     1.6007  -5.862 6.22e-09 ***
no.rooms     -80.6139    43.8440  -1.839 0.066264 .

Residual standard error: 781.8, 996 degrees of freedom
Multiple R-squared: 0.2588, Adjusted R-squared: 0.2566
F-statistic: 115.9 on 3 and 996 DF, p-value: < 2.2e-16
```

Test (possiblement multiples) I

Test simple $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$

Soit $(\boldsymbol{X}, \boldsymbol{y}) = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ un échantillon. Si le plan de régression est $y = \boldsymbol{x}^\top \boldsymbol{\beta}$, pour tester $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$, la statistique de test est

$$T = \frac{\widehat{\beta}_j}{s_{\widehat{\beta}_j}} \text{ où } s_{\widehat{\beta}_j} = \sqrt{s^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{jj}}.$$

Si $H_0 : \beta = 0$ est vraie, et si $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $T \sim \mathcal{S}td(n-p)$. Et donc

- on rejette H_0 si $|t| > T^{-1}_{n-p}(1 - \alpha/2)$

où T_ν est la fonction de répartition de la loi de Student $\mathcal{S}td(\nu)$

Régression Linéaire

Sous les hypothèses $\mathcal{H}_1$ et $\mathcal{H}_2^{\mathcal{N}}$, on a, pour $j = 1, \ldots, p$

$$T_j = \frac{\widehat{\beta}_j - \beta_j}{s_{\widehat{\beta}_j}} \sim \mathcal{S}td_{n-p} \quad \text{où } s_{\widehat{\beta}_j} = s\sqrt{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{jj}}.$$

```
> library(DALEX)
> reg = lm(m2.price~construction.year+surface+no.rooms
    , data=apartments)
> summary(reg)

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) 6295.7095  1884.1995   3.341 0.000865 ***
const.year    -0.8829     0.9599  -0.920 0.357920
surface       -9.3827     1.6007  -5.862 6.22e-09 ***
no.rooms     -80.6139    43.8440  -1.839 0.066264 .

Residual standard error: 781.8, 996 degrees of freedom
Multiple R-squared: 0.2588, Adjusted R-squared: 0.2566
F-statistic: 115.9 on 3 and 996 DF, p-value: < 2.2e-16
```

Régression Linéaire et test multiple

- ξ un sous-ensemble d'indices $\xi \subseteq \{1, \ldots, p\}$ de cardinal $|\xi|$.
- $\bar{\xi}$ les indices du complémentaire de ξ dans $\{1, \ldots, p\}$,
  Rappel: $\xi \cap \bar{\xi} = \varnothing$ et $\xi \cup \bar{\xi} = \{1, \ldots, p\}$
- $\boldsymbol{X}_\xi$ sous-matrice des covariables $\boldsymbol{x}_j, j \in \xi$.
- $\boldsymbol{X}_{\bar{\xi}}$ sous-matrice des covariables $\boldsymbol{x}_j, j \in \bar{\xi}$ (ou $j \notin \xi$).
- $\boldsymbol{\beta}_\xi$ les paramètres dans le modèle (ξ) où seules les variables ξ sont conservées.
- $[\hat{\boldsymbol{\beta}}]_\xi$: coordonnées ξ du vecteur $\hat{\boldsymbol{\beta}}$,
  Note: $[\hat{\boldsymbol{\beta}}]_\xi \neq \hat{\boldsymbol{\beta}}_\xi$ en général (sauf si $\boldsymbol{X}_\xi \perp \boldsymbol{X}_{\bar{\xi}}$).

Note: $\boldsymbol{u} = (u_1, \cdots, u_k) = 0$ signifie $\forall j,\ u_j = 0$.
$\boldsymbol{u} = (u_1, \cdots, u_k) \neq 0$ signifie $\exists j$ tel que $u_j \neq 0$.

Régression Linéaire et test multiple

Test $H_0 : \boldsymbol{\beta}_\xi = 0$ contre $H_1 : \boldsymbol{\beta}_\xi \neq 0$

Soit $(\boldsymbol{X}, \boldsymbol{y}) = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ un échantillon. Pour tester $H_0 : \boldsymbol{\beta}_\xi = 0$ contre $H_1 : \boldsymbol{\beta}_\xi \neq 0$ on estime

$$\begin{cases} \boldsymbol{y} = \boldsymbol{X}_{\bar{\xi}}\boldsymbol{\beta}_{\bar{\xi}} + \boldsymbol{\varepsilon}_{\bar{\xi}} & (0) \text{ régression contrainte} \\ \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} & (1) \text{ régression non-contrainte} \end{cases}$$

La statistique de test est

$$F = \frac{\text{SCR}(\boldsymbol{\beta}_{\bar{\xi}}) - \text{SCR}(\boldsymbol{\beta})}{\text{SCR}(\boldsymbol{\beta})} \frac{n-p}{|\xi|} = \frac{n-p}{|\xi|} \frac{R^2 - R_0^2}{1 - R^2}$$

Si $H_0 : \beta = 0$ est vraie, et si $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $F \sim \mathcal{F}(n-p, |\xi|)$. Et donc

- on rejette H_0 si $f > F^{-1}_{n-p,|\xi|}(1-\alpha)$.

Régression Linéaire et test multiple

Sur nos données sur le prix des logements en Pologne

$$\begin{cases} y_i = \beta_0 \phantom{+ \beta_1 x_{1,i}} + \beta_2 x_{2,i} \phantom{+ \beta_3 x_{3,i}} + \eta_i & (0) \\ y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \varepsilon_i & (1) \end{cases}$$

```
> reg1 = lm(m2.price~construction.year+surface+no.
    rooms, data=apartments)
> reg0 = lm(m2.price~surface, data=apartments)
> anova(reg0, reg1)
Analysis of Variance Table

Model 1: m2.price ~ surface
Model 2: m2.price ~ construction.year + surface + no.
    rooms
  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1    998 611258600
2    996 608730962  2   2527638 2.0678  0.127
```

Régression Linéaire et test multiple

```
> linearHypothesis(reg1,
  c("construction.year = 0", "no.rooms = 0"))
Linear hypothesis test

Hypothesis:
construction.year = 0
no.rooms = 0

Model 1: restricted model
Model 2: m2.price ~ construction.year + surface + no.
    rooms

  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1    998 611258600
2    996 608730962  2   2527638 2.0678  0.127
```

Régression Linéaire et test multiple

Quand il y a beaucoup de variables, il est possible d'utiliser des méthodes de sélection de variables, les méthodes pas à pas, ou step-wise.
Les algorithmes les plus simples consistent à faire rentrer les variables une à une (méthode ascendante, forward), ou les à les faire sortir une à une (méthode descendante, backward).

Régression Linéaire et test multiple

Sélection de variable pas à pas

Soit $(\boldsymbol{X}, \boldsymbol{y}) = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ un échantillon.

- ξ, $\xi \subseteq \{1, \ldots, p\}$ de cardinal $|\xi|$.
- ξ_{+1}, $\xi_{+1} \subseteq \{1, \ldots, p\}$ de cardinal $|\xi| + 1$.
- $\xi \subset \xi_{+1}$, autrement dit $\xi = \xi_{+1} \cup \{j\}$, $j \in \{1, \ldots, p\}$.

$$\begin{cases} \boldsymbol{y} = \boldsymbol{X}_{\xi_{+1}} \boldsymbol{\beta}_{\xi_{+1}} + \boldsymbol{\varepsilon}_{\xi_{+1}} & (\xi) \\ \boldsymbol{y} = \boldsymbol{X}_{\xi} \boldsymbol{\beta}_{\xi} + \boldsymbol{\varepsilon}_{\xi} & (\xi_{+1}) \end{cases}$$

On préfère (ξ) à (ξ_{+1}) si $R_a^2(\xi) > R_a^2(\xi_{+1})$.

Note notion de parcimonie et rasoir d'Ockham.

Omission d'une variable explicative I

Oublier une variable importante peut avoir des conséquences importantes

- $y_i = \beta_0 + \boldsymbol{x}_1^\top \boldsymbol{\beta}_1 + \boldsymbol{x}_2^\top \boldsymbol{\beta}_2 + \varepsilon_i$: le vrai modèle
- $y_i = b_0 + \boldsymbol{x}_1^\top \boldsymbol{b}_1 + \eta_i$: le modèle que l'on considère

L'estimateur de $\boldsymbol{b}_1$ est

$$\begin{aligned}
\widehat{\boldsymbol{b}}_1 &= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{y} \\
&= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top [\boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}] \\
&= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_1 \boldsymbol{\beta}_1 + (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_2 \boldsymbol{\beta}_2 + (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{\varepsilon} \\
&= \boldsymbol{\beta}_1 + \underbrace{(\boldsymbol{X}_1' \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_2 \boldsymbol{\beta}_2}_{\boldsymbol{\beta}_{12}} + \underbrace{(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{\varepsilon}}_{\nu_i}
\end{aligned}$$

de tel sorte que $\mathbb{E}[\widehat{\boldsymbol{b}}_1] = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_{12} \neq \boldsymbol{\beta}_1$, en général.

Omission d'une variable explicative II

Comme le montrait Bickel, Hammel& O'Connell (1975) (avec un modèle plus complexe car les variables ne sont pas ici continues)

- y est l'admission aux études graduées
- x_1 est le genre (homme ou femme)
- x_2 est le programme où l'étudiant(e) a postulé

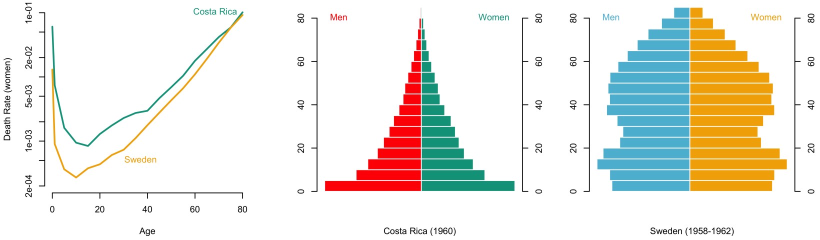
	Total	Men	Women	Proportions
Total	5233/12763 ~ 41%	3714/8442 ~ **44%**	1512/4321 ~ 35%	66%-34%
Top 6	1745/4526 ~ 39%	1198/2691 ~ **45%**	557/1835 ~ 30%	59%-41%
A	597/933 ~ 64%	512/825 ~ 62%	89/108 ~ **82%**	88%-12%
B	369/585 ~ 63%	353/560 ~ 63%	17/ 25 ~ **68%**	96%- 4%
C	321/918 ~ 35%	120/325 ~ **37%**	202/593 ~ 34%	35%-65%
D	269/792 ~ 34%	138/417 ~ 33%	131/375 ~ **35%**	53%-47%
E	146/584 ~ 25%	53/191 ~ **28%**	94/393 ~ 24%	33%-67%
F	43/714 ~ 6%	22/373 ~ 6%	24/341 ~ **7%**	52%-48%

Omission d'une variable explicative III

- y est la durée de vie résiduelle (en années)
- x_1 est le pays (Costa Rica ou Suède)
- x_2 l'âge de la personne

$$\mathbb{P}[Y \leq 1|\boldsymbol{X} = \text{Costa Rica}] < \mathbb{P}[Y \leq 1|\boldsymbol{X} = \text{Suède}]$$

$$\mathbb{P}[Y \leq 1|\boldsymbol{X} = (\text{Costa Rica}, x)] > \mathbb{P}[Y \leq 1|\boldsymbol{X} = (\text{Suède}, x)], \ \forall x$$



Confidence et prédiction

Intervalle de prédiction

Soit $(\boldsymbol{X},\boldsymbol{y}) = \{(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_n,y_n)\}$ un échantillon. Si le plan de régression est $y = \boldsymbol{x}^\top\beta$. On dispose d'une nouvelle observation $\boldsymbol{x}_{n+1}$. L'intervalle de confiance de la valeur moyenne prédite est:

$$\left[\boldsymbol{x}_{n+1}^\top\hat{\beta} \pm t_{1-\alpha/2,n-p}s\sqrt{\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}^\top}\right]$$

L'intervalle de confiance pour une valeur particulière est:

$$\left[\boldsymbol{x}_{n+1}^\top\hat{\beta} \pm t_{1-\alpha/2,n-p}s\sqrt{1+\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}^\top}\right]$$

Confidence et prédiction

$$\left[\boldsymbol{x}_{n+1}^{\top}\hat{\beta} \pm t_{1-\alpha/2,n-p}\, s\sqrt{1+\boldsymbol{x}_{n+1}^{\top}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{x}^{\top}n+1}\right]$$

```
> predict(reg, newdata = data.frame(construction.year
    =1992, surface=80, no.rooms=3), interval = "
    confidence")
       fit      lwr      upr
1 3544.494 3472.005 3616.983
> x=c(1,1992,80,3)
> t(x)%*%reg$coefficients
         [,1]
[1,] 3544.494
> residus = reg$residuals
> t(x)%*%reg$coefficients+qt(c(.025,.975),n-4)*sqrt(
    sum(residus^2)/(n-4))*sqrt(t(x)%*%solve(t(X)%*%X)
    %*%x)
[1] 3472.005 3616.983
```

Confidence et prédiction

$$\left[\boldsymbol{x}_{n+1}^\top\widehat{\beta} \pm t_{1-\alpha/2,n-p}s\sqrt{1+\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}^\top n+1}\right]$$

```
> predict(reg, newdata = data.frame(construction.year
    =1992, surface=80, no.rooms=3), interval = "
    prediction")
       fit      lwr      upr
1 3544.494 2008.663 5080.326
> x=c(1,1992,80,3)
> t(x)%*%reg$coefficients
         [,1]
[1,] 3544.494
> residus = reg$residuals
> t(x)%*%reg$coefficients+qt(c(.025,.975),n-4)*sqrt(
    sum(residus^2)/(n-4))*sqrt(1+t(x)%*%solve(t(X)%*%X
    )%*%x)
[1] 2008.663 5080.326
```

Sélection de variables

```
> library(olsrr)
> model = lm(mpg ~ disp + hp + wt + qsec, data =
    mtcars)
> ols_step_all_possible(model)

Index N       Predictors  R-Square Adj. R-Square
     1 1              wt 0.7528328     0.7445939
     2 1            disp 0.7183433     0.7089548
     3 1              hp 0.6024373     0.5891853
     4 1            qsec 0.1752963     0.1478062
     5 2           hp wt 0.8267855     0.8148396
     6 2         wt qsec 0.8264161     0.8144448
     7 2         disp wt 0.7809306     0.7658223
     8 2         disp hp 0.7482402     0.7308774
     9 2       disp qsec 0.7215598     0.7023571
    10 2         hp qsec 0.6368769     0.6118339
    11 3      hp wt qsec 0.8347678     0.8170643
    12 3      disp hp wt 0.8268361     0.8082829
    13 3    disp wt qsec 0.8264170     0.8078189
    14 3    disp hp qsec 0.7541953     0.7278591
    15 4 disp hp wt qsec 0.8351443     0.8107212
```

Sélection de variables

```
> model = lm(y ~ ., data = surgical)
> ols_step_forward_p(model)

                    Selection Summary
---------------------------------------------------------------
           Variable                     Adj.
 Step       Entered      R-Square     R-Square      RMSE
---------------------------------------------------------------
    1    liver_test       0.4545       0.4440     296.2992
    2    alc_heavy        0.5667       0.5498     266.6484
    3    enzyme_test      0.6590       0.6385     238.9145
    4    pindex           0.7501       0.7297     206.5835
    5    bcs              0.7809       0.7581     195.4544
---------------------------------------------------------------
```