Modèle linéaire

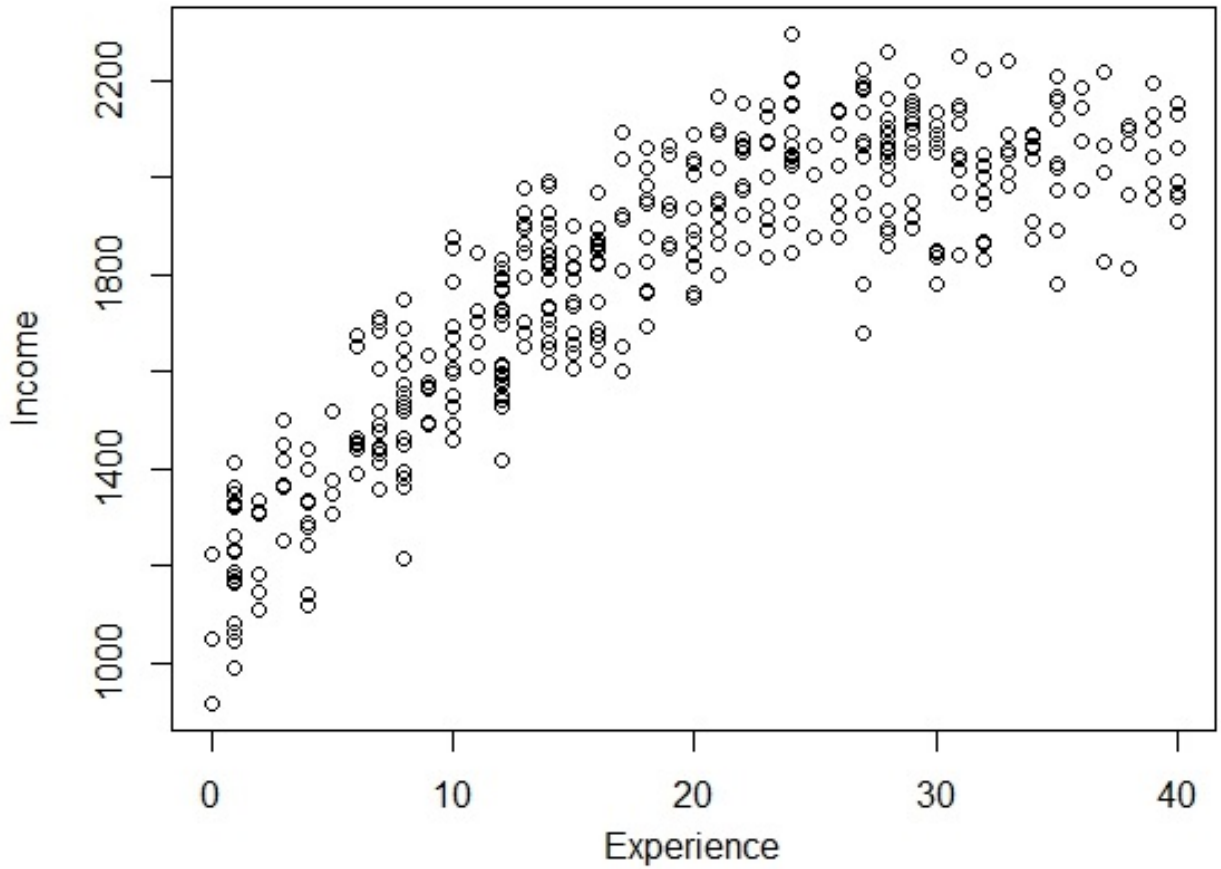
Hugo Harari-Kermadec

EOS - Econometrie

1 Modèle linéaire

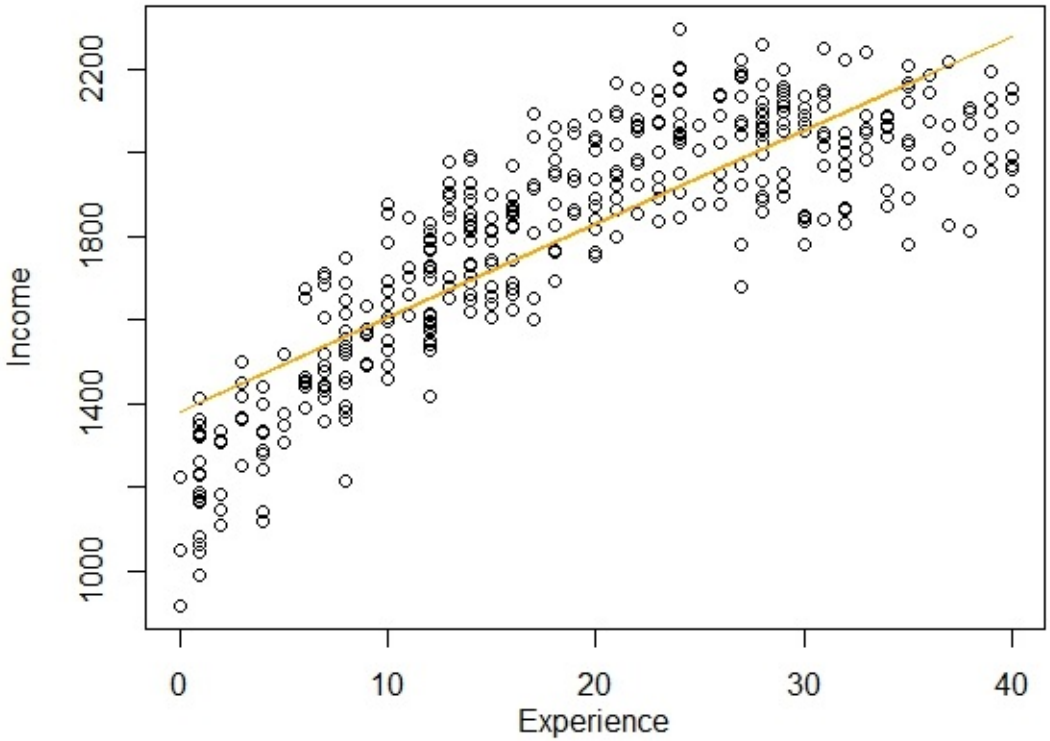
Modèle linéaire simple

Soient 2 variables, X, l'expérience et Y le revenu.



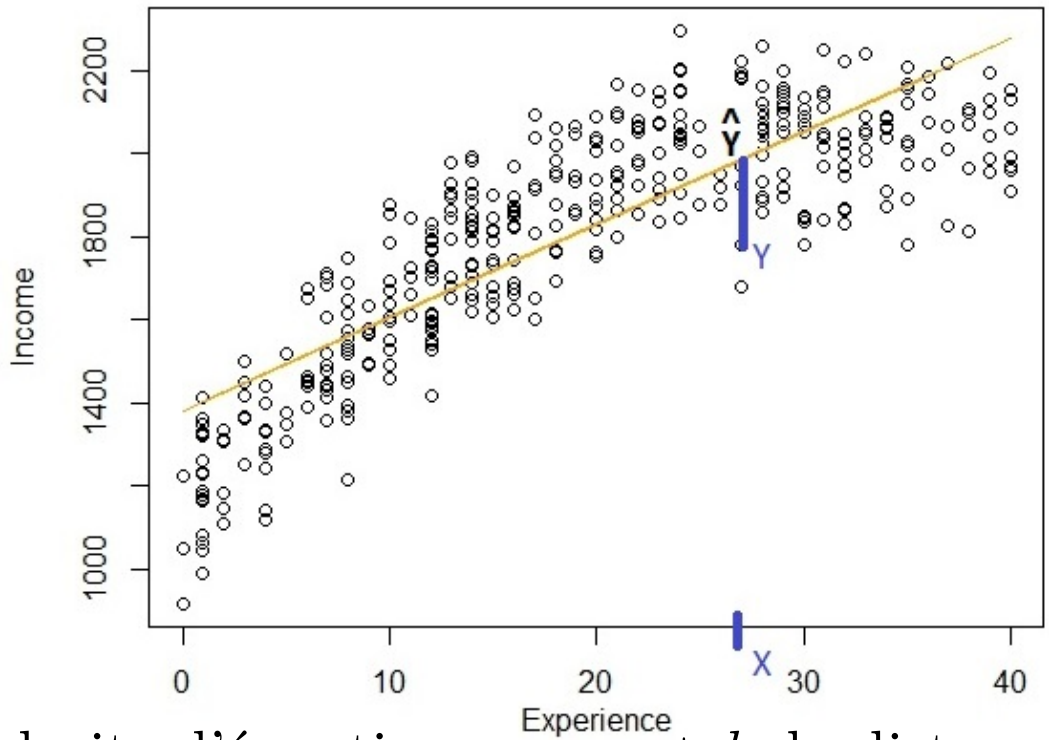
Moindres carrés ordinaires (OLS)

On cherche la droite qui minimise la distance **verticale** aux données.



Moindres carrés ordinaires (OLS)

On cherche la droite qui minimise la distance **verticale** aux données.



pour toute droite d'équation $y = ax + b$, la distance verticale à un point (X_i, Y_i) est :

$$|Y_i - (aX_i + b)|$$

Modèle linéaire

Le modèle formel s'écrit

$$Y_i = b + aX_i + \varepsilon_i$$

où ε_i est l'écart individuel inobservé au modèle.

Modèle linéaire

Le modèle formel s'écrit

$$Y_i = b + aX_i + \varepsilon_i$$

où ε_i est l'écart individuel inobservé au modèle.
Les estimateurs des moindres carrés (OLS) $\hat{a}$ et $\hat{b}$ minimisent :

$$\min_{a,b} \sum_i (Y_i - b - aX_i)^2 = \min_{a,b} \sum_i (\varepsilon_i)^2.$$

Modèle linéaire

Le modèle formel s'écrit

$$Y_i = b + aX_i + \varepsilon_i$$

où ε_i est l'écart individuel inobservé au modèle.
Les estimateurs des moindres carrés (OLS) $\hat{a}$ et $\hat{b}$ minimisent :

$$\min_{a,b} \sum_i (Y_i - b - aX_i)^2 = \min_{a,b} \sum_i (\varepsilon_i)^2.$$

On appelle valeurs modélisées (fitted values) $\hat{Y}_i = \hat{b} + \hat{a}X_i$
et résidus $\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{b} - \hat{a}X_i$.

Les estimateurs OLS $\hat{a}$ et $\hat{b}$ minimisent:

$$\min_{a,b} \sum_i (Y_i - b - aX_i)^2 = \min_{a,b} \sum_i (\varepsilon_i)^2.$$

L'estimateur efficace pour a est donc la corrélation:

$$\hat{a} = \frac{Cov(X, Y)}{Var(X)} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

et celui de b donne l'ordonnée à l'origine (intercept) :

$$\hat{b} = \overline{Y} - \hat{a}\overline{X}.$$

Les estimateurs OLS $\hat{a}$ et $\hat{b}$ minimisent:

$$\min_{a,b} \sum_i (Y_i - b - aX_i)^2 = \min_{a,b} \sum_i (\varepsilon_i)^2.$$

L'estimateur efficace pour a est donc la corrélation:

$$\hat{a} = \frac{Cov(X, Y)}{Var(X)} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

et celui de b donne l'ordonnée à l'origine (intercept) :

$$\hat{b} = \overline{Y} - \hat{a}\overline{X}.$$

La variance σ^2 de ε es estimée par

$$\hat{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Vecteurs

On écrit les équations individuelles les unes au dessus des autres :

$$\begin{aligned} Y_1 &= b + aX_1 + \varepsilon_1 \\ Y_2 &= b + aX_2 + \varepsilon_2 \\ \dots & \quad \dots \end{aligned}$$

Ca fait des vecteurs :

$$Y = (\mathbb{1}\ X) \times (b, a)' + \mathcal{E},$$

avec $Y = (Y_1, \dots, Y_n)'$, $\mathcal{E} = (\varepsilon_1, \dots, \varepsilon_n)'$, $\mathbb{1} = (1, \dots, 1)'$ et $X = (X_1, \dots, X_n)'$ appartenant à $\mathbb{R}^n$.

Avec ces notations, l'estimateur des moindres carrés s'écrit

$$(\hat{b}, \hat{a})' = \left((\mathbb{1}\ X)'(\mathbb{1}\ X)\right)^{-1} (\mathbb{1}\ X)' Y$$

Avec ces notations, l'estimateur des moindres carrés s'écrit

$$(\hat{b}, \hat{a})' = \left((\mathbb{1}\ X)'(\mathbb{1}\ X)\right)^{-1} (\mathbb{1}\ X)' Y$$

Sous les hypothèses

H_1 : les colonnes de X sont linéairement indépendantes,

H_2 : les ε_i sont d'espérance nulle et non corrélés aux X_i.

H_3 : ε_i sont indépendants entre eux, et de variance commune σ^2.

Alors l'estimateur OLS est sans biais et de variance minimale parmi les estimateurs linéaires.

Modèle multilinéaire

On ajoute simplement d'autres variables explicatives :

$$Y_i = b + a_1 X_{i1} + a_2 X_{i2} + \cdots + a_K X_{iK} + \varepsilon_i$$

Modèle multilinéaire

On ajoute simplement d'autres variables explicatives :

$$Y_i = b + a_1 X_{i1} + a_2 X_{i2} + \cdots + a_K X_{iK} + \varepsilon_i$$

on renomme b en a_0 pour simplifier, avec $X_{i0} = 1$ pour tout i

$$\begin{aligned} Y_i &= \sum_{k=0}^{K} a_k X_{ik} + \varepsilon_i \\ &= (X_{i0}, \ldots, X_{iK})(a_0, \ldots, a_K)' + \varepsilon_i \end{aligned}$$

Modèle multilinéaire

On ajoute simplement d'autres variables explicatives :

$$Y_i = b + a_1 X_{i1} + a_2 X_{i2} + \cdots + a_K X_{iK} + \varepsilon_i$$

on renomme b en a_0 pour simplifier, avec $X_{i0} = 1$ pour tout i

$$\begin{aligned} Y_i &= \sum_{k=0}^{K} a_k X_{ik} + \varepsilon_i \\ &= (X_{i0}, \ldots, X_{iK})(a_0, \ldots, a_K)' + \varepsilon_i \end{aligned}$$

Verticalement, pour n données

$$Y = X\theta + \mathcal{E},$$

avec $Y = (Y_1, \ldots, Y_n)' \in \mathbb{R}^n$, $\mathcal{E} = (\varepsilon_1, \ldots, \varepsilon_n)' \in \mathbb{R}^n$ et $X = (X_1, \ldots, X_n)' \in \mathcal{M}_{n,K+1}$.

Tests

Un test important est celui de significativité de l'effet de chaque variable :
Supposons les ε_i gaussiens : $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, alors

$$\hat{a} \sim \mathcal{N}(a, \sigma^2 (X'X)^{-1})$$

On estime σ^2 parh $\widehat{\sigma^2} = \frac{\widehat{\mathcal{E}}'\widehat{\mathcal{E}}}{n-K-1}$

$$\frac{\hat{a}_k - a_k}{\hat{\sigma}_k} \sim \mathcal{T}(n-K-1), \text{ where } \hat{\sigma}_k = \sqrt{\widehat{\sigma^2}(X'X)^{-1}_{kk}}.$$

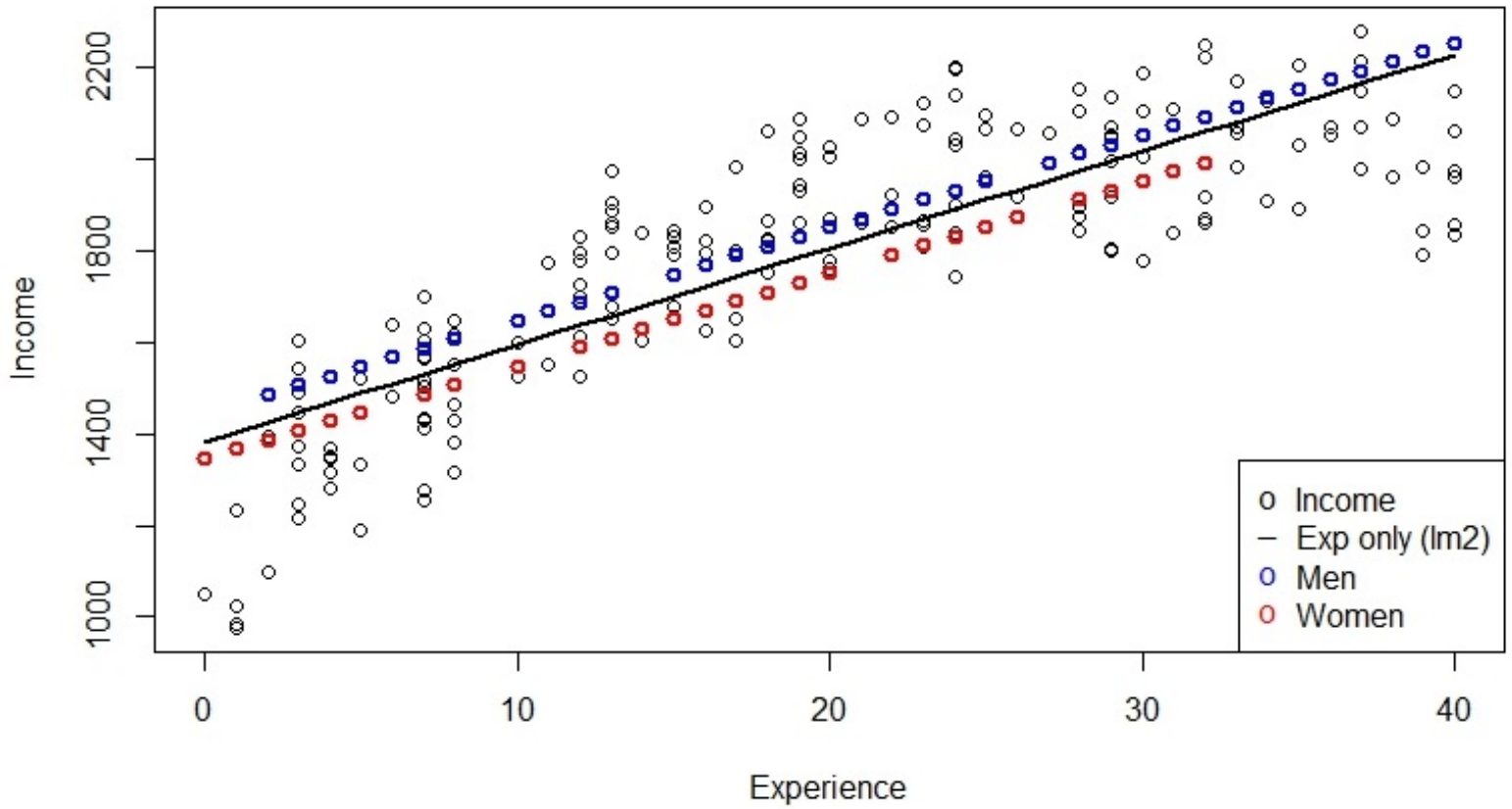
On peut alors utiliser un test de Student (t-test) pour tester $H_0 : \ a_k = 0$.

Expérience et genre

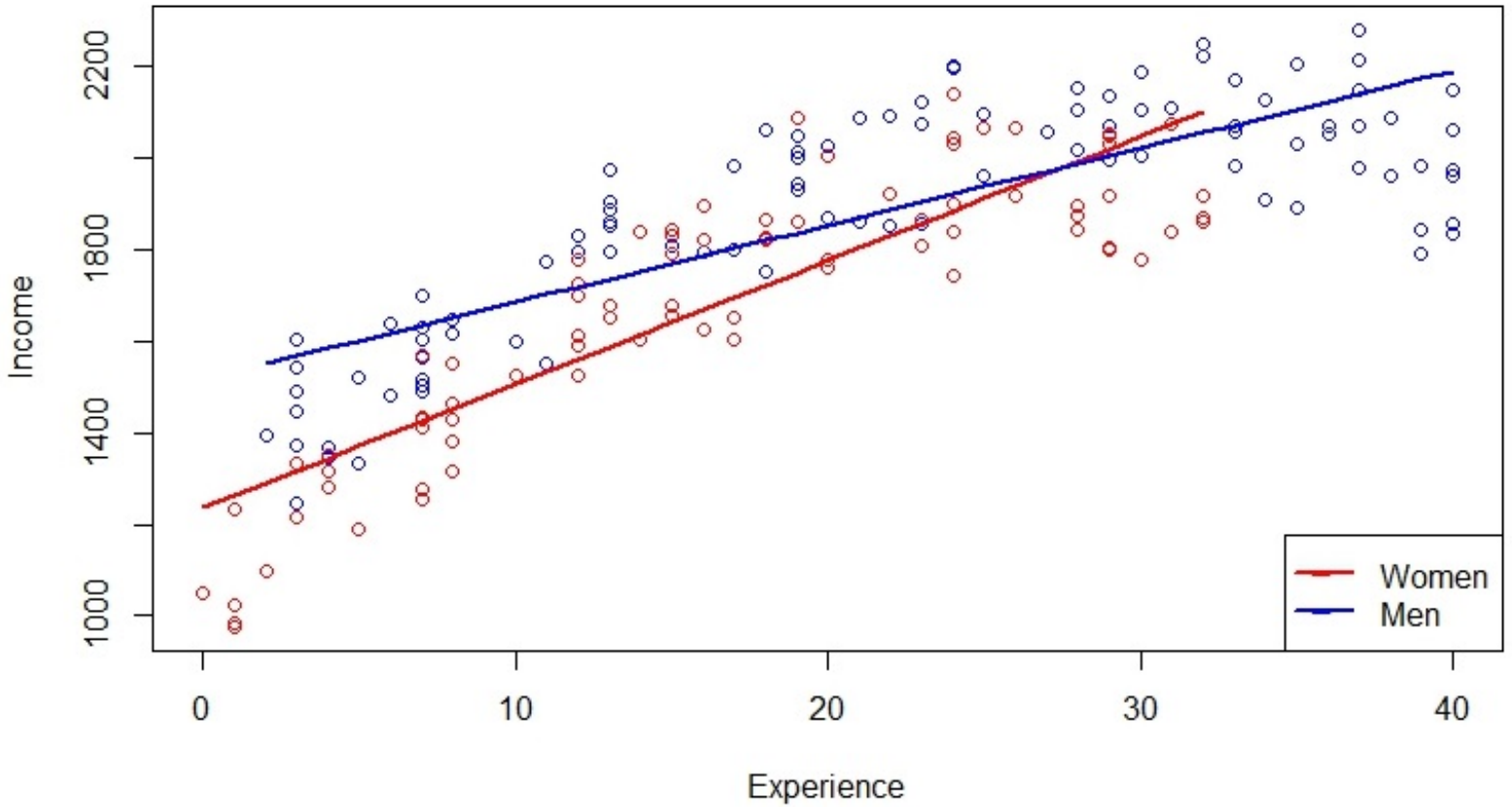
$$\textbf{Genre}= \begin{cases} 1 & \text{si i est une femme} \\ 0 & \text{sinon} \end{cases}$$ est une variable binaire (dummy).

```
lm(Income ~ Exp + Gender);summary(lm3)

Coefficients:
             Estimate  Std.Error t value Pr(>|t|)
(Intercept) 1446.202     26.530  54.511  < 2e-16 ***
Exp           20.247      1.032  19.612  < 2e-16 ***
Gender       -99.735     23.288  -4.283 2.88e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 160.6 on 197 degrees of freedom
Multiple R-squared:  0.6989,Adjusted R-squared:  0.6958
```

Expérience et genre



Expérience et genre



`lm(Income ~ Exp + Gender +Exp*Gender)`

Analyse de la Variance

$$Y_i = b + aX_i + \varepsilon_i$$

L'AnOVa teste si X permet d'expliquer l'essentiel de la variance de Y :

$$\sum_i (Y_i - \overline{Y})^2 = a^2 \sum_i (X_i - \overline{X})^2 + \sum_i \varepsilon_i^2$$

Si on suppose que tout est gaussien

$$\frac{\sum_i \varepsilon_i^2}{n-1-\dim(X)} \frac{\dim(X)}{a^2 \sum_i (X_i - \overline{X})^2} \sim F(n-1-\dim(X), \dim(X))$$

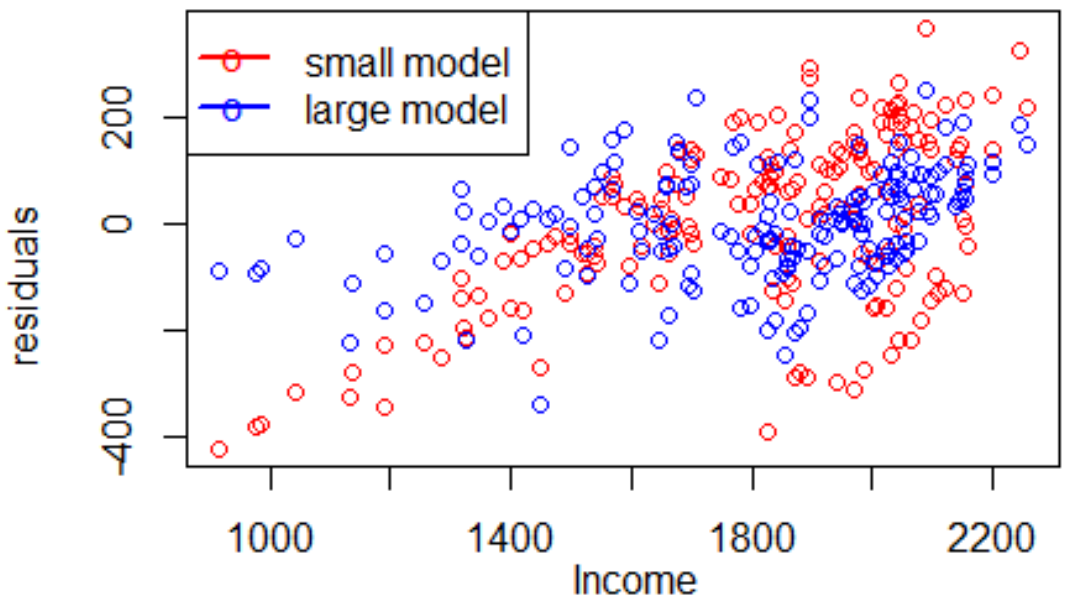
AnOVa et lm()

```
lm5<-lm(Income ~ Exp +Exp2 + Gender + Gender*Exp)
anova(lm5)

Response: Income
             Df   Sum Sq  Mean Sq   F value Pr(>F)
Exp           1 11319023 11319023 1068.2138 <2e-16 ***
Exp2          1  2386337  2386337  225.2065 <2e-16 ***
Gender        1  1100248  1100248  103.8340 <2e-16 ***
Exp:Gender    1     1380     1380    0.1302 0.7186
Residuals   195  2066262    10596
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

AnOVa pour les modèles emboités (nested)

$$Income = Exp + Gender \quad (1)$$

$$Income = Exp + Exp2 + Gender \quad (2)$$



AnOVa pour les modèles emboités (nested)

```
> anova(lmA,lmB)
Analysis of Variance Table

Model 1: Income ~ Experience + Gender
Model 2: Income ~ Experience + Experience2 + Gender

  Res.Df     RSS  Df Sum of Sq      F    Pr(>F)
1    197 5396845
2    196 2174223  1   3222622    290.51  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```