Modèle linéaire

Hugo Harari-Kermadec

M1 APE - Techniques quantitatives

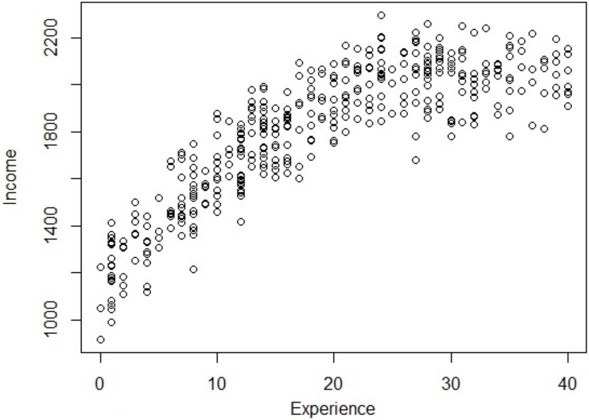
2020

1 Modèle linéaire

1. Modèle linéaire
2. Modèle multilinéaire et propriétés

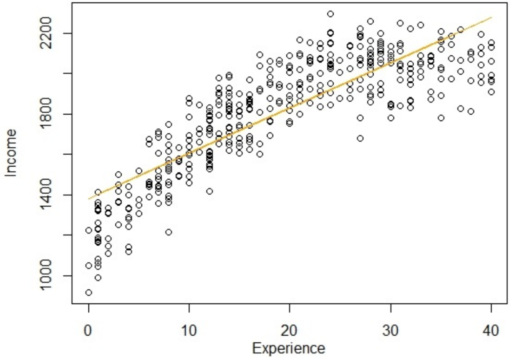
Modèle linéaire simple

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



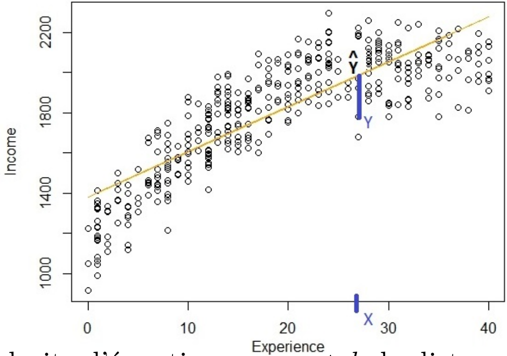
Moindres carrés ordinaires (OLS)

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 est estimée par

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

L'estimateur des moindres carrés est optimal

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.

Application

On veut expliquer le revenu Y par l'expérience X :

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

```
lm(Income ~ Exp)

Coefficients:
            Estimate  Std.Error t value Pr(>|t|)
(Intercept)  b=1446    26.5     54.511  < 2e-16 ***
Exp          â=20      1.03     19.612  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 170.6 on 198 degrees of freedom
Multiple R-squared:  0.65,Adjusted R-squared:  0.65
```

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

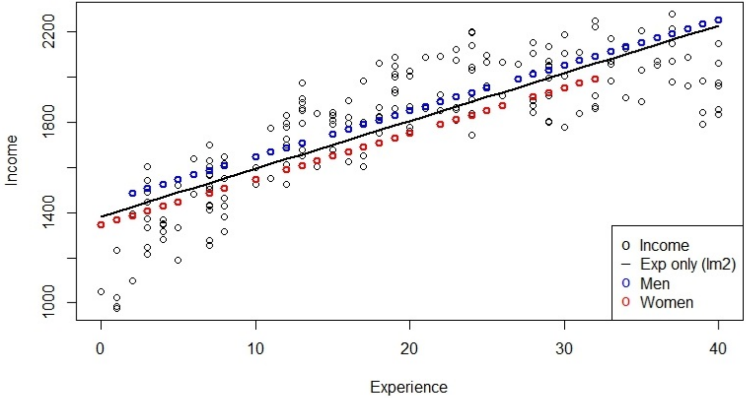
$$Y_i = \sum_{k=0}^{K} a_k X_{ik} + \varepsilon_i$$

Expérience et genre

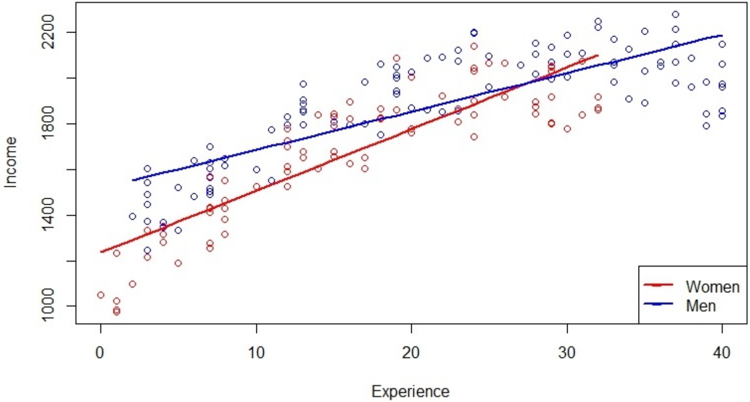
Genre$=\begin{cases} 1 & \text{si i est une femme} \\ 0 & \text{sinon} \end{cases}$ est une variable binaire (dummy).

```
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)
```