Linear model

Hugo Harari-Kermadec

EPOG - Econometrics

1. Simple linear model

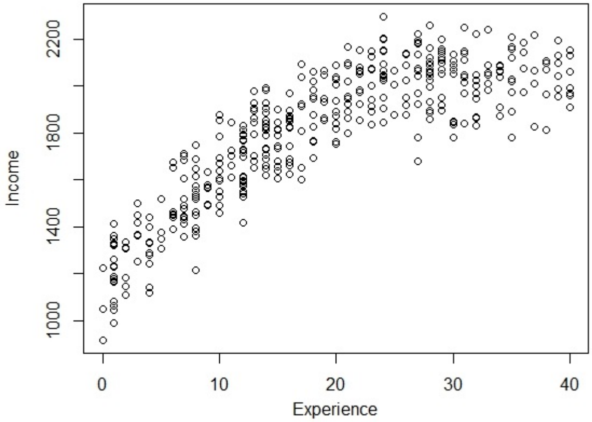
1. Simple linear model

2. Multilinear model and properties

1. Simple linear model
   - Method
   - Model
   - Properties
2. Multilinear model and properties
3. AnOVa

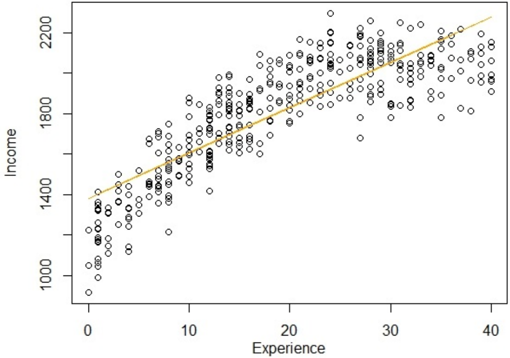
Simple linear model

Let's consider 2 variables, X, the experience, and Y the income.



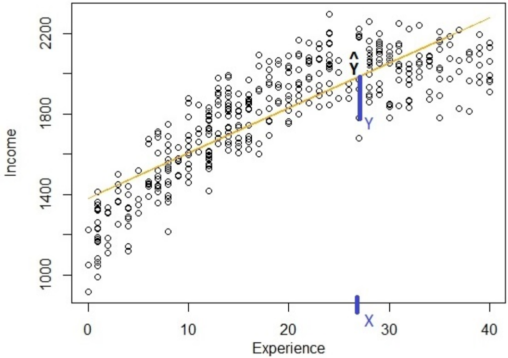
Least squares line

The line that minimizes the **vertical** distance with the data.



Least squares line

The line that minimizes the **vertical** distance with the data.



for any line with equation $y = ax + b$, the vertical distance with each individual data is: $Y_i - (aX_i + b)$.

Linear model

The formal model writes

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

where ε_i is an unobserved individual variation from the model.

Linear model

The formal model writes

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

where ε_i is an unobserved individual variation from the model. The OLS estimators $\hat{a}$ and $\hat{b}$ minimize:

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

Linear model

The formal model writes

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

where ε_i is an unobserved individual variation from the model.
The OLS estimators $\hat{a}$ and $\hat{b}$ minimize:

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

We call fitted values $\hat{Y}_i = \hat{b} + \hat{a}X_i$
and residuals $\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{b} - \hat{a}X_i$.

The OLS estimators $\hat{a}$ and $\hat{b}$ minimize:

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

The best estimator for a is the correlation:

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

and the best estimator for b sets the intercept:

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

The OLS estimators $\hat{a}$ and $\hat{b}$ minimize:

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

The best estimator for a is the correlation:

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

and the best estimator for b sets the intercept:

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

The variance σ^2 of ε is estimated by

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

Vectors

Write the individual equations one above the other:

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

It builds vectors:

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

with $Y = (Y_1, \dots, Y_n)'$, $\mathcal{E} = (\varepsilon_1, \dots, \varepsilon_n)'$, $\mathbb{1} = (1, \dots, 1)'$ and $X = (X_1, \dots, X_n)'$ all in $\mathbb{R}^n$.

With these notations, the least squares estimator writes

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

With these notations, the least squares estimator writes

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

Under the following hypothesis

H_1 : Columns of X are linearly independent,

H_2 : ε_i errors have 0 expectation and are uncorrelated with X_i.

H_3 : ε_i errors are uncorrelated with common variance σ^2.

The OLS estimator is unbiased and as minimum variance among linear estimators.

Multilinear model

It's the same, with more explanatory variables:

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

Multilinear model

It's the same, with more explanatory variables:

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

b is incorporated as a_0, with $X_{i0} = 1$

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

Multilinear model

It's the same, with more explanatory variables:

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

b is incorporated as a_0, with $X_{i0} = 1$

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

Vertically, for n data

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

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

Tests

One important test is the significance of the effect of each explanatory variable:
Assuming that the ε_i are normally distributed,

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

and with $\widehat{\sigma^2} = \frac{\widehat{\mathcal{E}}'\widehat{\mathcal{E}}}{n-K-1}$

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

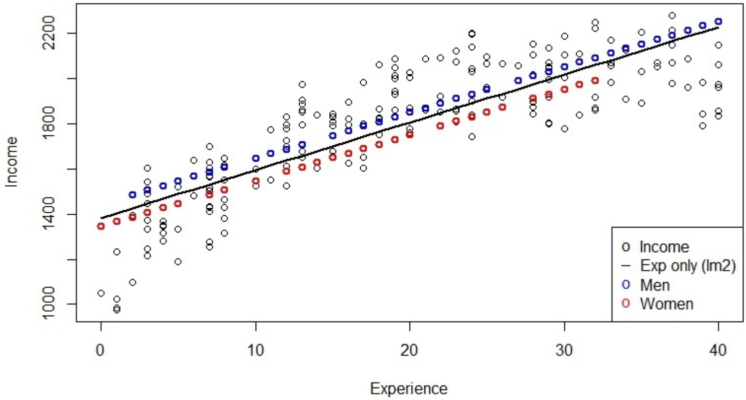
A t test can then test $H_0 : \ a_k = 0$.

Experience and Gender

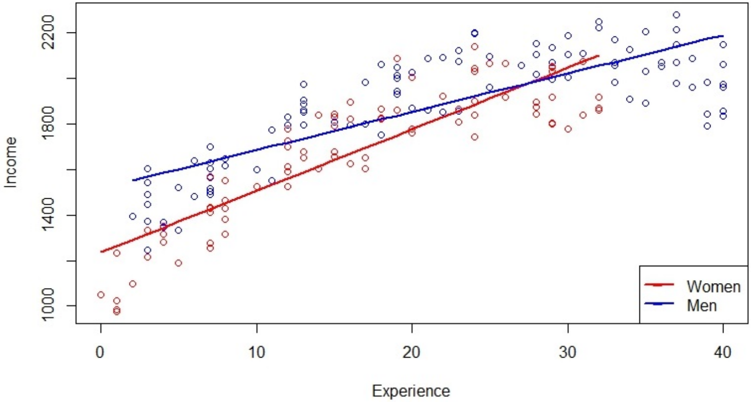
$$\textbf{Gender} = \begin{cases} 1 & \text{if the individual is a female} \\ 0 & \text{else} \end{cases} \text{ is a dummy.}$$

```
lm3<-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
```

Experience and Gender



Experience and Gender




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

Analysis of Variance

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

The AnOVa tests how much of the variance of Y is explained by X, and how much remains:

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

Assuming everything is gaussian

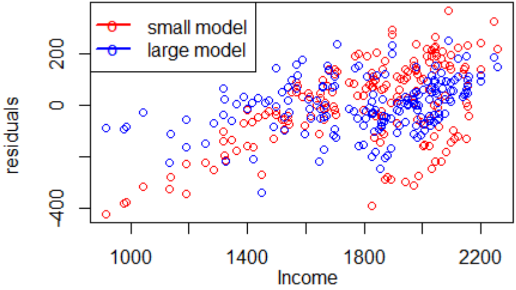
$$\frac{\sum_i \varepsilon_i^2}{n - 1 - \dim(X)} \frac{\dim(X)}{a^2 \sum_i (X_i - \overline{X})^2} \sim F(n - 1 - \dim(X), \dim(X))$$

AnOVa and 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 for nested models

$$Income = Exp + Gender \quad (1)$$
$$Income = Exp + Exp2 + Gender \quad (2)$$



AnOVa for nested models

```
> 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
```