Econometrics Part 1

The linear model: presentation

Jose Angel Garcia Sanchez

Université Paris 1 Panthéon-Sorbonne
jagarsanc@gmail.com

Sorbonne Data Analytics

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Outline

Topics addressed

- Linear regression
- Heteroskedasticity and autocorrelation
- Instrumental Variables
- Simultaneous Equations
- Panel Data
- Discrete choice
- Censored variables
- Time series

Econometrics

- Economic analysis relies on theoretical representations of behaviors and mechanisms (ex : macro-economic models, the Capital Asset Pricing Model, etc)
- The relevance of these models should be tested on real data (ex : does the CAPM "work"?)
- Parameters of interest should be assessed too (ex : what is the impact of a rise in oil prices or interest rates on car sales?)
- Econometrics uses statistical tools to provide an answer to these questions
- Here, theoretical and hands-on econometrics : How to manage real data? How to find the best model? Did I get what I expected? Can I trust this model?

Example

Consider an individual consumption function : for each individual i, we observe consumption C_i and income I_i. A very simple model explaining C_i would be :

$$\forall i, \quad C_i = a + bI_i$$

The goal of econometrics would be to test the relevance of this model (is it a *good approximation* of reality?), to provide *estimates* of parameters a and b and to assess the accuracy of these estimates (inference : tests). Plus, one may be concerned that the link between C and I is not that perfect.

Types of datasets

- *Cross-sectional data* : individual data collected at a particular point in time (ex : census data on individuals, households, firms)
- *Time series data* : variables collected over time (ex : yearly GDP, monthly unemployment rate)
- *Panel data* : individual data collected over time (ex : Medical Expenditure Panel Survey)

Cross-sectional data and panel data are collected not at the *population* level, but on a *sample*, representative of the population. We thus have to infer the characteristics of the population from the information we get from that particular sample.

Variables

- Quantitative variables are measured on a numeric or quantitative scale : income, age ...
- Qualitative variables are not : gender, region ..., but need to be coded with numbers anyway (*dummy variables*)

Consider model $\forall i,\ C_i = a + bI_i$

- C_i : the endogenous (*or* dependent *or* explained) variable which is the one explained by the model
- I_i and constant : exogenous (*or* independent *or* explanatory) variables that determine the endogenous variable

The simple linear model : example

Say we'd like to know about the health care habits of the French population. Assume that for each individual i, health care expenditure y_i could be explained by age x_i, in a linear way :

$$\forall i, \quad y_i = b_0 + b_1 x_i$$

For a newborn, expenditure would be b_0, and each additional year would incur b_1 extra cost. This makes sense, since we all know that health expenditure increase with age. This is a simple model because we have only one explanatory variable.

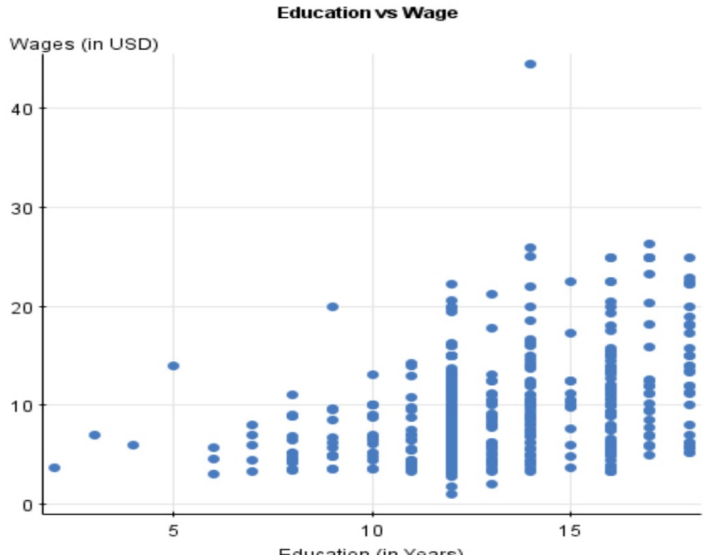
Error term - linearity

Of course, the link between age and expenditure cannot be that perfect, so we add an **error term** u_i that accounts for unobserved factors that make this relation not perfect (frailty, income, etc).

$$\forall i, \quad y_i = b_0 + b_1 x_i + u_i$$

This is a (simple) model of individual health care expenditure. This is how we believe the data was generated in the first place. Its relevance cannot be assessed on the whole population, but rather on a sample, say of size N. The model can be estimated using ordinary least squares (OLS), if some assumptions are verified. Notice that this is a linear model because parameters (and not necessarily variables) enter it in a linear way.

Example


Education vs Wage
Wages (in USD)
40
30
20
10
0
5
10
15
Education (in Years)

The simple linear model : assumptions

Let's call i an individual from the size N sample.

$$\forall i, \quad y_i = b_0 + b_1 x_i + u_i$$

Calling E the expected value and V the variance, the usual and so-called *Gauss-Markov assumptions* for such a model are :

1. $\forall i,\ E(u_i|X) = 0$
2. u and x are uncorrelated
3. $V(X) \neq 0$
4. $\forall i,\ V(u_i) = \sigma^2$
5. $\forall i \neq j,\ cov(u_i, u_j) = 0$

Everything is conditional on X, we don't write it every time for simplicity.

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

What they mean

1. We want the model to predict the right outcome, given any x :
   $\forall i,\ E(y_i|x) = E(b_0 + b_1 x_i + u_i|x_i) =$
   $E(b_0 + b_1 x_i|x_i) + E(u_i|x_i) = b_0 + b_1 E(x_i|x_i) = b_0 + b_1 x_i,$
2. u should be totally random, no way to guess u from X
3. The x_i can't be all the same,
4. The variance of the error term should be roughly the same over the individuals of the sample
5. Individuals should all be uncorrelated.

Reminder (1) : E and V

- $E(X)$ is the *expected value* of X : theoretical mean
- $\bar{X}$ is the *average* of observed X : empirical mean
- $V(X)$ is the *variance* of X : how we expect X to vary around its mean (theoretical)
- $V(X) = E((X - E(X))^2)$
- $\hat{V}(X)$ is the *empirical variance* of X : how X indeed varies around its mean
- $\hat{V}(X) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N}$
- When $\hat{V}(X)$ is used to estimate $V(X)$, there are minor adjustments : we divide by $(N-1)$ instead of N

Reminder (2) : Conditioning

- Earlier, we wrote : $E(y_i|x) = E(b_0 + b_1 x_i + u_i|x_i)$
- Say y is income and x is gender : 0 for males, 1 for females
- Expected income for a male, i.e. knowing $x = 0$:
  $E(y_i|x = 0) = b_0 + b_1 * 0 = b_0$
- Expected income for a female, i.e. knowing $x = 1$:
  $E(y_i|x = 1) = b_0 + b_1 * 1 = b_0 + b_1$
- So b_0 is the expected male income, $b_0 + b_1$ is the expected female income, and b_1 is the expected difference
- $\Rightarrow E(y_i|x)$ is used to predict individuals' outcomes, given their characteristics

Reminder (3) : Conditioning

- Without conditioning on x, this would be :
  $E(y_i) = E(b_0 + b_1 x_i + u_i) = b_0 + b_1 E(x_i)$
- $E(x_i)$ is the expected proportion of females in the population
- So $E(y_i)$ is the average income over the whole population, i.e. the weighted average of males' and females' incomes
- $\Rightarrow E(y_i)$ is used to predict the overall average outcome
- Usually, people discuss $E(y_i)$ while implicitly discussing $E(y_i|x)$ instead, because the latter is more informative

Interpretation of parameters

Say we have the following model :

$$\forall i, \quad y_i = a + bx_i + u_i$$

Then b can be interpreted as the marginal change in y when x increases by 1 unit, because since u is by definition unrelated to x, we have :

$$\frac{dy}{dx} = b$$

The linear model is in fact very general

Example : squared variables

The following model is linear, even if variable x is squared :

$$y_i = a + bx_i^2 + u_i$$

And a way to understand how y changes with x is to compute the following derivative (let's drop i's for convenience) :

$$\frac{dy}{dx} = 2xb$$

And we see it is not constant, it depends on the value of x. The relationship between x and y is not a line but a curve. It would be even better to put in the equation both x and x^2 to allow for more flexibility, that's what is done in practice :

$$y_i = a + bx_i + cx_i^2 + u_i$$

.

Example : logs (1)

- Let's have a simple model : $\forall i,\ log(y_i) = a + bx_i + u_i$
- How can we interpret b? (let's drop i's for convenience)
- Taking the exponential : $y = exp(a + bx + u)$

$$\frac{dy}{dx} = b.exp(a + bx + u) = b.y$$

$$\frac{dy/y}{dx} = b$$

b is thus interpreted as the **percentage** of variation of y when x increases by 1 unit. This kind of model has a *semi-log* form.

Example : logs (2)

- Let's have a simple model : $\forall i,\ log(y_i) = a + b.log(x_i) + u_i$
- How can we interpret b?
- Taking the exponential : $y = exp(a + b.log(x) + u)$

$$\frac{dy}{dx} = b.\frac{1}{x}exp(a + b.log(x) + u) = b.\frac{y}{x}$$

$$\frac{dy/y}{dx/x} = b$$

b is thus interpreted as an **elasticity** : the % of variation of y when x increases by 1%. This kind of model has a *log-log* form.

How to choose?

- Using economic theory, common sense, plots ...
- Example : the link between income and medical expenditures using the World Bank's 1997 Vietnam Living Standards Survey (source : Cameron & Trivedi, chapter 4)
- Model : $\textit{medical expenditures} = a + b.\textit{income} + u$
- Data : 5,999 households
- Total household income is usually not well captured in developing countries, so we use as a *proxy* total household expenditures

Our goal : estimate the parameters of the model

$$\forall i, \quad y_i = b_0 + b_1 x_i + u_i$$

- b_0 and b_1 are unknown.
- We thus want to find the best estimates for them : $\hat{b}_0$ and $\hat{b}_1$.
- One way of doing this is by using ordinary least squares : OLS.
- $\hat{b}_0 + \hat{b}_1 x_i$ should be as close as possible to y_i.
- We define $\hat{y}_i$ as the model prediction of y_i : $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$
- And we define $\hat{u}_i$ as $y_i - \hat{y}_i$: the residuals
- Warning : u and $\hat{u}$ are two different things

OLS estimators

- We want to minimize the global sum of residuals : this will give us the optimal $\hat{b}_0$ and $\hat{b}_1$
- But the residuals are known only when the regression line is drawn, and for that we need $\hat{b}_0$ and $\hat{b}_1$: what can we do?
- We need to express residuals as a function of $\hat{b}_0$ and $\hat{b}_1$, and then minimize their sum
- Since residuals can be either positive of negative, we minimize the sum of **squared** residuals with respect to $\hat{b}_0$ and $\hat{b}_1$:

$$Min \sum \hat{u}_i^2 = Min \sum (y_i - \hat{y}_i)^2 = Min \sum (y_i - \hat{b}_0 - \hat{b}_1 x_i)^2$$

And the solution is (proof below) :

$$\hat{b}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

and

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

Remark

$$\hat{b}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{s_{xy}}{s_{xx}}$$

- $s_{xx} = \frac{1}{N}\sum(x_i - \bar{x})^2$ is the empirical variance of x
- $s_{xy} = \frac{1}{N}\sum(x_i - \bar{x})(y_i - \bar{y})$ is the empirical covariance between x and y
- So that :

$$\hat{b}_1 = \frac{\widehat{cov}(X, y)}{\hat{V}(X)}$$

- Warning : correlation $\neq$ causality

Proof (1)

Calling $S(b_0, b_1) = \sum \hat{u}_i^2$, we need the two First Order Conditions :

$$\frac{\partial S(b_0, b_1)}{\partial b_0} = 0 \Rightarrow -2 \sum_{i=1}^{N} (y_i - \hat{b}_0 - \hat{b}_1 x_i) = 0 \quad (1)$$

$$\frac{\partial S(b_0, b_1)}{\partial b_1} = 0 \Rightarrow -2 \sum_{i=1}^{N} x_i (y_i - \hat{b}_0 - \hat{b}_1 x_i) = 0 \quad (2)$$

Since the function is globally convex and twice differentiable, we don't really need the Second Order Conditions : we know in advance that the minimum exists and is unique

Proof (2)

Equation 1 implies :

$$\sum y_i = N\hat{b}_0 + \hat{b}_1 \sum x_i \quad (3)$$

We divide everything by N, use the empirical means $\bar{y} = \frac{\sum y_i}{N}$ and $\bar{x} = \frac{\sum x_i}{N}$, and get :

$$\bar{y} = \hat{b}_0 + \hat{b}_1 \bar{x} \quad (4)$$

$\Rightarrow$ the regression line goes through the sample mean
(= the average person)
$\Rightarrow$ the average residual is always 0

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Proof (3)

Thanks to equation 4, we get $\hat{b}_0 = \bar{y} - \hat{b}_1\bar{x}$, and we plug-in $\bar{y} - \hat{b}_1\bar{x}$ instead of $\hat{b}_0$ in equation 2 :

$$\sum x_i(y_i - \bar{y} + \hat{b}_1\bar{x} - \hat{b}_1 x_i) = 0 \quad (5)$$

$$\sum x_i(y_i - \bar{y}) = \hat{b}_1 \sum x_i(x_i - \bar{x}) \quad (6)$$

$$\hat{b}_1 = \frac{\sum x_i(y_i - \bar{y})}{\sum x_i(x_i - \bar{x})} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad (7)$$

The OLS estimator is linear

We have :

$$\hat{b}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

So :

$$\hat{b}_1 = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} = \sum \frac{(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} y_i$$

Estimate $\hat{b}_1$ is thus a linear combination of elements y_i, and can also be written as a linear combination of elements u_i, since $y_i = b_0 + b_1 x_i + u_i$.

Are all observations given the same weight?

We have :

$$\hat{b}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

So :

$$\hat{b}_1 = \frac{\sum(x_i - \bar{x})^2 \frac{(y_i - \bar{y})}{(x_i - \bar{x})}}{\sum(x_i - \bar{x})^2} = \sum p_i \frac{(y_i - \bar{y})}{(x_i - \bar{x})}$$

Calling $p_i = \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$.
$\frac{(y_i - \bar{y})}{(x_i - \bar{x})}$ is the slope of the line drawn from the point corresponding to individual i to the sample average, and p_i is an increasing function of $(x_i - \bar{x})$.
Estimate $\hat{b}_1$ is thus highly influenced by extreme points (see Anscombe's quartet of identical regressions).

OLS estimators are "BLUE"

$$\forall i, \quad y_i = b_0 + b_1 x_i + u_i$$

- OLS estimators are the Best Linear Unbiased Estimators (BLUE) : this is the *Gauss-Markov theorem*, which is valid under the *Gauss-Markov assumptions* seen earlier
- They are linear in y_i, as seen earlier
- $(\hat{b}_0)$ and $(\hat{b}_1)$ are particular outcomes of random variables (their expression comprises y, which in turn comprises u), so we can compute their expected value and variance
- Unbiased : $E(\hat{b}_0) = b_0$, $E(\hat{b}_1) = b_1$
- "Best" means they have a minimal variance among unbiased linear estimators
- All this is true whatever the size of the sample
- Proof : in the general case with more than one X

The variance of estimators : summary

- Recall that $V(u) = \sigma^2$
- $V(\hat{b}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$
- $V(\hat{b}_0) = \frac{\sigma^2}{N} + \bar{x}^2 V(\hat{b}_1)$
- $cov(\hat{b}_0, \hat{b}_1) = -\bar{x} V(\hat{b}_1)$
- To get an estimate of these variances, we need an unbiased estimator for σ^2 : $\hat{\sigma}^2$
- $\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{N-2}$
- Indeed, $\hat{V}(\hat{u}) = \frac{\sum (\hat{u}_i - \bar{\hat{u}})^2}{N} = \frac{\sum \hat{u}_i^2}{N}$
- And we need a small correction to make it an unbiased estimator of $V(u)$: divide by $N-2$ instead of N
- This is because we lose 2 degrees of freedom, since there are 2 parameters to estimate in the model (proof in the general case)

OLS estimators : summary

In model $y_i = b_0 + b_1 x_i + u_i$:

$$\hat{b}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

and

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

- Unbiased : $E(\hat{b}_0) = b_0$ and $E(\hat{b}_1) = b_1$
- Smallest variance possible : $V(\hat{b}_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$ where $\sigma^2 = V(u_i)$
- Sensitive to extreme values

OLS estimators : distribution

- OLS estimators are a linear combination of u_i's
- If we had a large sample, we could rely on the Central Limit Theorem to learn about the distribution of estimators (see next section)
- But for any sample size, the only way to guess the distribution of estimators is to assume we know the distribution of u_i's
- A common assumption is that $u_i \hookrightarrow N(0, \sigma^2)$
- It implies that both OLS estimates follow a normal : $\hat{b}_0 \hookrightarrow N(b_0, \sigma^2_{b_0})$ and $\hat{b}_1 \hookrightarrow N(b_1, \sigma^2_{b_1})$
- Next, we'll check how these estimators behave with a large sample size

Asymptotics : a preview

- Do OLS estimates have good properties when sample size goes to infinity, i.e. "asymptotically"?
- Let's call $(\hat{b}_{0,N})$ and $(\hat{b}_{1,N})$ the series of estimators corresponding to a sample of size N
- $(\hat{b}_{0,N})$ and $(\hat{b}_{1,N})$ *converge in probability* towards b_0 and b_1
- Definition of convergence in probability : $(X_1, X_2, ..., X_t)$ converges in probability towards a if and only if :
  $P(X_t \neq a) \to 0$ when $t \to +\infty$
- It means that when sample size goes to infinity, the probability that estimates are exactly equal to the true parameters is close to 1

Additional assumption needed (1)

To prove this convergence, we need to assume that variable x follows this rule :

$$\lim_{N\to\infty} \frac{1}{N} \sum (x_i - \bar{x})^2 = \sigma_x^2 \neq 0$$

Which simply means that the empirical variance of x should have a given limit which is not zero. If it were zero, then it would mean that when we increase sample size, after a while variable x wouldn't vary anymore (it would stick to its average $\bar{x}$) so additional x's wouldn't provide any information.

Additional assumption needed (2)

- In fact, the only thing we need here is that the x keep some variance when sample size goes to infinity, which is very likely to happen.
- We also need the variance of X not to go to infinity, which implies *stationarity*[1] of the X's, so sometimes a weaker condition is used : $\lim_{N\to\infty} \sum(x_i - \bar{x})^2 = \infty$

[1] This part of the course focuses on micro data, stationarity is related to Time Series, covered later on

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Consistency

- Convergence in quadratic mean $\Rightarrow$ Convergence in probability
- Definition : a series of random variables $(X_1, X_2, ..., X_t)$ converges in quadratic mean towards variable X iff :
- $E[(X_t - X)^2] \to 0$ when $t \to \infty$
- Definition : $(X_1, X_2, ..., X_t)$ converges in probability towards a iff :
- $P(X_t \neq a) \to 0$ when $t \to \infty$

It means that we only need to prove that OLS estimators converge in *quadratic mean* towards their "true counterparts" to prove they are consistent in probability. So, we only need to prove that their variance goes to zero when sample size goes to infinity (easier).

Asymptotics

- Estimates are consistent, but can we guess what their distribution is?
- Since estimates depend on error terms, if we don't know the distribution of the error terms, we won't know the distribution of estimates
- But estimates are a linear combination of the error terms : thanks to the CLT[2], $\hat{b}_0$ and $\hat{b}_1$ are asymptotically normal
- So if sample size goes to infinity (=with a sample "large enough"), we know the distribution of estimates, so we may compute confidence intervals, run tests etc

[2] Central Limit Theorem

The Central Limit Theorem (CLT)

(Lindeberg-Feller version)

- Assume we have a series of random variables $(Z_1, Z_2, ..., Z_N)$ that are independent with finite expectancies and variances such that $\forall i$, $E(z_i) = \mu_i$ and $V(z_i) = \sigma_i^2$
- They can have different expectancies and variances, and don't need to share the same distribution
- Assume $\bar{z}_N = \frac{1}{N} \sum_{i=1}^{N} z_i$ and $\mu = \frac{1}{N} \sum \mu_i$
- Assume $\bar{\sigma}_N^2 = \frac{1}{N} \sum_{i=1}^{N} \sigma_i^2$ and $\lim_{N \to \infty} \bar{\sigma}_N^2 = \bar{\sigma}^2$
- Then : $\sqrt{N}(\bar{z}_N - \mu)$ converges in distribution towards $N(0, \bar{\sigma}^2)$, when $N \to \infty$

Asymptotic results

Using the CLT and focusing on parameter $\hat{b}_1$, we get :

- $\sqrt{N}(\hat{b}_1 - b_1) \rightarrow N(0, \frac{\sigma^2}{\sigma_x^2})$
- σ^2 is the variance of error terms u
- σ_x^2 is the asymptotic variance of x
- "Inflating" by $\sqrt{N}$ is needed, otherwise the variable would be degenerate (zero variance)

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Summary : distribution of parameters

- When sample size is large, OLS estimators become Normal thanks to the Central Limit Theorem (CLT) if error terms are *iid*, whatever their distribution, because estimators are a linear combination of error terms
- When sample size is not so large, parameters are Normal *only if* error terms are normal themselves, because these estimators are a linear combination of the error terms
- Knowing the distribution of estimates helps to draw inference (confidence intervals, tests, etc)

A measure of the "goodness of fit" of the regression

We call the R^2 or *coefficient of determination* :

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

- It is the ratio of the *variance of the outcome explained by the model* over the *total variance of the outcome*, and lies between 0 (very poor fit) and 1 (perfect fit).
- Formal derivation of the R^2 will be given in the more general case of the multiple linear model, as well as the conditions under which it can be used.
- Warning : if the x variable is a good predictor of y, but if the link is not linear, the R^2 will be low because of the lack of fit and not because of the poor choice of variables.

The multiple linear model

Say we get back to the simple model explaining health care expenditures by age, and we want to add to the model various other explanatory variables (income, household size, etc) :
$x_2, x_3, ..., x_{k-1}$

$$\forall i, \quad y_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + b_3 x_{i,3} + ... + b_{k-1} x_{i,k-1} + u_i \quad (8)$$

This is a (less simple) model of individual health care expenditure. It can be rewritten the following way, for individuals $i = 1$ to N :

$$\begin{aligned}
y_1 &= b_0 + b_1 x_{1,1} + b_2 x_{1,2} + b_3 x_{1,3} + ... + b_{k-1} x_{1,k-1} + u_1 \\
y_2 &= b_0 + b_1 x_{2,1} + b_2 x_{2,2} + b_3 x_{2,3} + ... + b_{k-1} x_{2,k-1} + u_2 \\
&... \\
y_N &= b_0 + b_1 x_{N,1} + b_2 x_{N,2} + b_3 x_{N,3} + ... + b_{k-1} x_{N,k-1} + u_N
\end{aligned}$$

Multiple linear model

This system of equations can be rewritten using simple vectors :

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = b_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + b_1 \begin{pmatrix} x_{1,1} \\ x_{2,1} \\ \vdots \\ x_{N,1} \end{pmatrix} + ... + b_{k-1} \begin{pmatrix} x_{1,k-1} \\ x_{2,k-1} \\ \vdots \\ x_{N,k-1} \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix} \quad (9)$$

And also in a more compact way, using matrices :

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,k-1} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,k-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \cdots & x_{N,k-1} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{k-1} \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix} \quad (10)$$

Rewritten : $y_{(N,1)} = X_{(N,k)} b_{(k,1)} + u_{(N,1)}$

Some notations

$$E(b) = \begin{pmatrix} E(b_0) \\ E(b_1) \\ \vdots \\ E(b_{k-1}) \end{pmatrix} \quad (11)$$

$$V(b) = E[(b - E(b))(b - E(b))'] \quad (12)$$

$$V(b) = \begin{pmatrix} V(b_0) & cov(b_0, b_1) & \cdots & cov(b_0, b_{k-1}) \\ cov(b_1, b_0) & V(b_1) & \cdots & cov(b_1, b_{k-1}) \\ \vdots & \vdots & \vdots & \vdots \\ cov(b_{k-1}, b_0) & cov(b_{k-1}, b_1) & \cdots & V(b_{k-1}) \end{pmatrix} \quad (13)$$

And notice that : $E(A.X) = A.E(X)$ but $V(A.X) = A.V(X).A'$

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Usual assumptions of the model

These are the same as the simple model, but generalized to any number of explanatory variables :

1. $E(u|X) = 0$
2. X and u are uncorrelated
3. $Rank(X) = k$ with $k < N$
4. $E(uu') = \sigma^2 I_N$
5. When $N \to \infty$, $\lim \frac{X'X}{N} = V_X$ where V_X is a finite non-singular matrix

(2) : consider for now that X is deterministic, we'll consider the more general case later. (3) : the rank of a matrix is the number of columns that are linearly independent. (5) : a non-singular matrix has an inverse ; the variance of X's is always non-zero.

Estimation process (1)

- OLS minimizes the sum of squared residuals, each one called $\hat{u}_i$
- Calling $\hat{u}$ the vector of residuals, this sum can be written as : $S = \sum \hat{u}_i^2 = \hat{u}'\hat{u}$
- The derivative of S with respect to the vectors of parameters $\hat{b}$ should be zero, for this vector of parameters to give an optimum for S
- Plus, the second derivative should be positive for this optimum to be a minimum
- We just need to compute the derivatives of S with respect to every single b_j, and set them to 0

The gradient

In fact, we take the derivative of a number with respect to a vector, and get the vector G of derivatives, called the *gradient* :

$$G = \partial S/\partial \hat{b} = \begin{pmatrix} \partial S/\partial \hat{b}_0 \\ \partial S/\partial \hat{b}_1 \\ \vdots \\ \partial S/\partial \hat{b}_{k-1} \end{pmatrix} \quad (14)$$

After some calculus, we get (proof below) :

$$\hat{b} = (X'X)^{-1}X'y \quad (15)$$

Optimization process (1)

We know that since $\hat{u} = y - \hat{y} = y - X\hat{b}$, then :

$$\begin{aligned} S &= \hat{u}'\hat{u} \\ &= (y - X\hat{b})'(y - X\hat{b}) \\ &= (y' - \hat{b}'X')(y - X\hat{b}) \\ &= y'y - y'X\hat{b} - \hat{b}'X'y + \hat{b}'X'X\hat{b} \\ &= y'y - 2y'X\hat{b} + \hat{b}'X'X\hat{b} \end{aligned}$$

Indeed, $y'X\hat{b}$ and $\hat{b}'X'y$ are each other's transpose, but they both are only scalars (matrix of size one by one), and a scalar and its transpose are the same thing.

Optimization process (2)

We can now compute the gradient G :

$$\begin{aligned} G &= \frac{\partial S}{\partial \hat{b}} \\ &= \frac{\partial (y'y)}{\partial \hat{b}} - \frac{\partial (2y'X\hat{b})}{\partial \hat{b}} + \frac{\partial (\hat{b}'X'X\hat{b})}{\partial \hat{b}} \\ &= 0 - 2X'y + 2X'X\hat{b} \end{aligned}$$

Since y is not a function of $\hat{b}$. We look for an optimum, so we want $G = 0$, which implies that $\hat{b} = (X'X)^{-1}X'y$ (given that $(X'X)^{-1}$ exists, that follows from the fact that $Rk(X) = k$).

Optimization process (3)

- We also need the second derivative of S to be positive, which in the matrix framework amounts to having the matrix of second derivatives *positive definite*.
- This matrix of second derivatives is equal to $\frac{\partial(2X'y+2X'X\hat{b})}{\partial\hat{b}} = 2X'X$ which is indeed positive definite.
- Reminder : a matrix A is positive definite if for any vector z of convenient size, $z'Az > 0$. So let's check this with our matrix $2X'X$.
- We only need to check that $X'X$ is positive definite. For any vector z of convenient size, $z'(X'X)z = (Xz)'(Xz)$. But matrix Xz is just a column vector, and a transposed vector times itself is a sum of squares, equal to its square norm (=length), which is always positive.

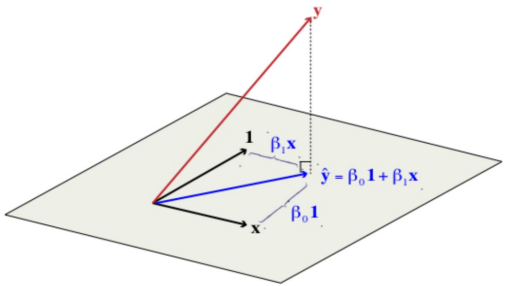
Geometrical interpretation of OLS (1)

See Wooldridge's IE, Appendix E

- In the minimization process, we got to a point where :
  $G = -2X'y + 2X'X\hat{b} = 0$
- Rewritten :
  $X'y - X'X\hat{b} = X'(y - X\hat{b}) = X'(y - \hat{y}) = X'\hat{u} = 0$
- Which means that $\hat{u}$ should be perpendicular to every column vector of matrix X, i.e. perpendicular to the vector space spanned by the column vectors of X
- Condition $X'\hat{u} = 0$ is called the *system of normal equations* (*normal* also means perpendicular)

Illustration

Assume variable y is explained by 2 variables, a constant and x :
$y = \beta_0 + \beta_1 x + u$
OLS will find parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the model prediction $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is as close as possible to the true y.



The vertical dashed line is vector $\hat{u}$ such that $y = \hat{y} + \hat{u}$. So we may find $\hat{\beta}$ starting directly from $X'\hat{u} = 0$ (normal equations)
Source : `https://sakai.unc.edu/`

Geometrical interpretation of OLS (2)

- Notice that $\hat{y} = X\hat{b} = X(X'X)^{-1}X'y = P_X y$.
- Where : $P_X = X(X'X)^{-1}X'$
- Matrix P_X is in fact the general form of a projector matrix, that projects orthogonally on $L(X)$ (vector space spanned by the columns of X)
- Prediction $\hat{y}$ is the orthogonal projection of y on $L(X)$
- Let's call $M_X = I - P_X$, with I the identity matrix of convenient size.
- M_X is a sort of complement to P_X
- M_X projects orthogonally on $L^{\perp}(X)$, vector space orthogonal to $L(X)$.
- We can see that residual $\hat{u} = M_X y$.
- Notice too that $y = \hat{y} + \hat{u} = P_X y + M_X y$: y can be split into 2 orthogonal parts

Projectors

For any orthogonal projection matrix P_X, we have :

- $P_X.P_X = P_X$: P_X is idempotent
- $P'_X = P_X$: P_X is symmetric

And some additional properties : $P_X + M_X = I$; $P_X.M_X = 0$; $P_X.X = X$; $M_X.X = 0$. The eigenvalues of a projection matrix are equal either to 1 (corresponding to the vector space they project onto) or 0 (corresponding to the supplementary vector space). For instance, P_X has k eigenvalues equal to 1 and $N - k$ eigenvalues equal to 0. (*From a geometrical point of view, OLS estimation is simply a projection.*)

Consequence : analysis of variance and R2

- Vector y can be split in two orthogonal parts : $y = \hat{y} + \hat{u}$
- Thanks to the Pythagorean theorem (proof below) :
- $V(y) = V(\hat{y}) + V(\hat{u})$
- The coefficient of determination is : $R^2 = V(\hat{y})/V(y)$
- R^2 represents the percentage of variance explained by the model
- But adding a variable, even if it is irrelevant, will artificially increase the R^2
- For models with a different number of explanatory variables, only the $\bar{R}^2$'s can be compared
- Adjusted R^2 : $\bar{R}^2 = 1 - \frac{N-1}{N-k}(1 - R^2)$ (warning : can be negative ...)

Proof

- Vector y can be split in two orthogonal parts : $y = \hat{y} + \hat{u}$
- Call e_N the size N vector only comprising 1's (the constant in the model), and $\bar{y}$ the average outcome : we can subtract $\bar{y}e_N$ on the right and on the left
- $y - \bar{y}e_N = (\hat{y} - \bar{y}e_N) + \hat{u}$
- $\hat{y} - \bar{y}e_N$ and $\hat{u}$ are orthogonal because $\hat{y} \perp \hat{u}$ (geometry of OLS) and $e_N \perp \hat{u}$ (since $e_N \in X$ and $X \perp \hat{u}$, again by the geometry of OLS)
- So we can apply the Pythagorean theorem
- $||y - \bar{y}e_N||^2 = ||\hat{y} - \bar{y}e_N||^2 + ||\hat{u}||^2$
- $\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(\hat{u}_i)^2$
- Since $\bar{y} = \hat{\bar{y}}$ because $\hat{u}$ averages to 0, this is equivalent to :
- $\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(\hat{u}_i)^2$
- So that : $V(y) = V(\hat{y}) + V(\hat{u})$

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

The R2

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{Sum\ of\ squares\ explained}{Sum\ of\ squares\ total} = \frac{V(\hat{y})}{V(y)}$$
$$= 1 - \frac{\sum \hat{u}_i^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{Sum\ of\ squared\ residuals}{Sum\ of\ squares\ total} = 1 - \frac{V(\hat{u})}{V(y)}$$

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Interpretations of the R squared

- R^2 represents the percentage of variance explained by the model
- It thus represents how well the model fits the data
- If there is a true relationship between the variables *but* this relationship is is fact non linear, the R^2 will be rather low
- The R^2 can be interpreted as the *multiple linear correlation coefficient* between y and the X's
- The R^2 is equal to the square of the linear correlation coefficient between y and its prediction $\hat{y}$
- If there is no constant in the model, it has no meaning because the way it is computed requires a constant term
- The R^2 is not enough to assess the relevance of a regression : we'll need statistical tests (see later)

Remarks (1)

- R^2 and adjusted R^2 are valid only when comparing models that have the same dependent variable
- So they are inappropriate to compare 2 models with y and $log(y)$ as the dependent variable
- Why : R^2 is a percentage of the variance of the outcome, so we can't compare percentages of different things
- See PE test (parametric encompassing) to compare models in log vs levels (see Verbeek)

Remarks (2)

What happens to the R^2 if we add an irrelevant explanatory variable to the model?

- $\hat{y}$ is the orthogonal projection of y on $L(X)$
- If X is "larger" because we have one more explanatory variable, then $\hat{y}$ cannot be worse than before
- Either there is no improvement to the projection, or there is a small improvement due to randomness
- So the R^2 cannot decrease, it can only increase
- So don't rely on the R^2 only to judge a model