A work-in-progress tentative econometrics' handout

Lucas Girard (CREST-ENSAE)

Current version: 17 September 2024

Contents

1 Introduction **5**

1.1 Syllabus – “Intermediate Econometrics”, Master of Economics (ENS Paris-Saclay) . . . 6

1.1.1 Objectives . . . 6

1.1.2 References . . . 6

1.1.3 Practical considerations . . . 6

1.1.4 Outline of the course . . . 7

1.2 Motivation . . . 8

1.2.1 Examples . . . 8

1.2.2 Tentative definition of econometrics . . . 12

1.2.3 Other key concepts . . . 12

1.3 Some probability and statistics reviews for econometrics . . . 17

2 Linear regressions – Ordinary Least Squares **19**

2.1 Setting and notation . . . 19

A Temporary work in Progress – elements to be put somewhere **23**

A.1 Notation . . . 23

Chapter 1

Introduction

1.1 Syllabus – "Intermediate Econometrics", Master of Economics (ENS Paris-Saclay)

1.1.1 Objectives

Main objectives of the course:

- Introduce (modern) econometrics: formalization of causal effects and how to identify it.
- Presenting a very popular and widely used method: the Ordinary Least Squares (OLS) and some of its extensions: Two Stage Least Squares (TSLS) and panel methods.

The course will remain a "theoretical" one: in a broad sense, we will do some maths. However, I will try to present practical examples (although, often, it will be rather illustrations than genuine, thorough empirical studies).

1.1.2 References

The course is mainly inspired by ENSAE's corresponding graduate course, "Econometrics 1" course conceived and taught by Xavier d'Haultfœuille. I taught that course as a Teaching Assistant (chargé de Travaux Dirigés) for many years, and it massively influenced my understanding and teaching of econometrics.

The material provided for the course should be self-contained.

However, if you want to go further or to take a look at other presentations, the main reference of the course is the book *Mostly Harmless Econometrics* (2009), Princeton University Press, by J. Angrist and S. Pischke [Henceforth MHE].

Another very interesting resource I know is the lecture notes from Gary Chamberlain's 2010 Econometrics class. They are available here thanks to Paul Goldsmith-Pinkham: `https://github.com/paulgp/GaryChamberlainLectureNotes/`.

Other classical references for such econometrics courses (but I admit I have worked with them only exceptionally) are:

- *Introductory Econometrics: A Modern Approach*, J. Wooldrige, 2008, South-Western College Publishing)
- *Econométrie : méthodes et applications*, B. Crépon and N. Jacquemet, 2018, Deboeck supérieur.
- *Econometric analysis of cross section and panel data*, J. Wooldrige, 2010, MIT Press ("the" Wooldridge, a very classical handbook in econometrics)

1.1.3 Practical considerations

Format of the course:

- 12 amphitheater classes of two hours and a half (including a 15-minute break) = 27 hours.
- 7 small classes (TD) of one hour and a half = 10,5 hours.

Grading:

- 40%: 1h30 written midterm exam (to be held on November 4th);
- 60%: 2h written final exam (to be held in the first week of January).

Both are closed-book exams, that is, without access to any documents. A priori, both will have a similar structure:

- For about one third: a multiple choice quiz
- For about one half: an exercise mixing theoretical and more practical questions (comments and interpretations of regression outputs, etc.)
- For the remaining part: a more theoretical exercise.

1.1.4 Outline of the course

Key ideas The logical progression of the course can be summed up in this way (Figure 1.1):

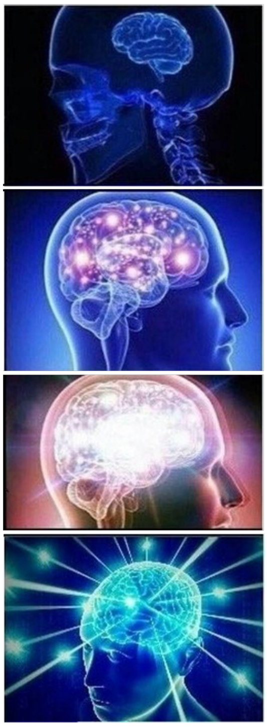


Figure 1.1: A "sum-meme-ry" of the course:

Under mild moment conditions, the theoretical linear regression is *always* well defined: a so-called "linear model" meaning (P) $Y = X'\beta_0 + \varepsilon$ with $\mathbb{E}[X\varepsilon] = 0$ is, in that sense, tautological.

However, linear representation (P), "simple projection", *does not generally coincide* with the causal representation involving potential outcomes $Y(d)$ and causal parameters, which are *jointly defined: causal effect* :=: *differences of* $Y(d)$.

For that and to estimate an average (weighted or on a sub-population) causal effect, assumed to be linear, by a simple linear regression (OLS), *there needs to be no selection bias.* Example and conceptual benchmark: experiments where the treatment D is randomly drawn.

In most applications, the absence of selection bias is not very credible. Yet, we can nonetheless identify causal effects: (*i*) by adding adequate control variables G such that the absence of conditional selection is plausible, or (*ii*) by using valid instrumental variables Z (TSLS).

Structure That progression yields the following structure of the course:

1. First part: tools from probability and statistics OLS

   (a) Estimation: fundamental of linear regressions

   (b) Inference (parametric tests and confidence regions): statistical uncertainty in linear regressions

   (c) Use of linear regressions for non-causal predictions in stable environments[1]

2. Second part: definitions of causal effects and links with (simple) linear regressions

   (a) Formalization of the notion of causality through potential outcome variables (Neyman-Rubin causal model)

   (b) Non-causal linear representation (that is, theoretical linear projection) and causal representation: do they coincide? In other words, is there a selection bias / an endogeneity issue?

3. Third part: several identification strategies to recover average causal effects despite (unconditional) selection bias / endogenous treatments

   (a) Add adequate control variables

   (b) Rely on valid instrumental variables (IV, or instruments)

   (c) Take advantage of richer longitudinal data: difference-in-differences, panel data methods (within and first difference estimators)

[1] This part is likely to be additional material: not studied during lectures, not at the exam program, with material support to go further if interested.

1.2 Motivation

1.2.1 Examples

A first example During some web navigation or at the library, you encounter the (title of the) following article published in *The Harvard Business Review* in May-June 2021:[2]

Banks with More Women on Their Boards Commit Less Fraud

by Scott Berinato

From the Magazine (May–June 2021)

In such situations where causal effects are evoked, what questions may/should you ask yourself?

. . .

. . .

. . .

Three (types of) questions We will study that example with more details later. As of now, we can consider three main questions (that should become reflexes you have in front of any causal claims). Importantly, those three questions are distinct ones. Remark that, in particular, knowing the answer of one is not informative of the answer of the two others.

- **Identification of causal effects (causality or correlation)?** Is the asserted causal effects really one? Very often, we can think of other reasons explaining the correlation without any causal link (confounder; chicken or egg issue, reverse causality). How much should I trust the causality claimed?

  Problem: we can, at best, observe correlations, but we are interested in causal effects.

  Important warning to keep in mind: *no causation without manipulation.*

- **Uncertainty quantification (statistical significance)?** Be it causal or not, is the correlation "real" or just an artefact from a particular finite sample (as opposed to a census entirely covering a given population of interest)?

  Related concepts and key distinctions to keep in mind: estimand vs. estimator vs. estimate. Always be aware of the NATURE OF THE OBJECTS, in particular, whether they are (modelled as) stochastic or not.

  Tools: probabilistic modelling and inference (tests, confidence regions). If not explicitly stated otherwise, we place ourselves in a frequentist setting.

- **Is the correlation "important" (practical significance)?** Again, irrespective of the interrogation causality/correlation, is the link important in practice?

[2] This particular example is chosen and discussed here; feel free to find other examples, namely any situation where a causal effect is asserted.

Some comments about the example and those three questions

Causality of correlation.

Here, we can think of a simplistic/toy explanation as this one: imagine that the direction of each bank is influenced by another variable, namely the morality/social awareness of its direction. Note that this variable is quite hard to observe and to quantify. However, if so, that variable would be a confounder: it causes both the fact that the bank tends to recruit more top-management women and tend to commit less fraud. Thus, it is not because there are more women on their boards that such banks commit less fraud.

A clue for such issues is to reverse the statement (reverse causality): "banks that commit less fraud have more women on their boards."

If the converse [réciproque] statement seems, a priori, as likely as the initial implication, it does not say at all that the study is false, but it is a warning that the causal links and its sense are rather difficult to disentangle and identify. It suggests that we should be particularly careful in assessing the identification of a causal effects.

Statistical significance.

Is the study conducted in a random sample of banks, some specially selected banks? What is the population of interest? In other words, when and where the asserted statement applies? Northern Italian banks during Renaissance? French banks in the 2020s? American banks in the XXth century? Banks with more than one trillion of assets?

Question: is the sample representative of the population of interest?

These are very important questions in practice, somewhat upstream of any econometric or statistical analysis. It will not be the core of the course, but we will discuss it.

Practical significance.

A toy example. (Newtonian) Gravity. No doubt there is a real causal gravity effect of the Moon on me. Yet, in the current setting of me in Earth, it is totally negligible in practice, it is not a first order (and not even second or higher order) factor to explain my physical movement. Remark that it is not because the gravity effect of the Moon is "small" in an absolute sense, but because it is relatively small compared to the first-order effect due to the Earth's gravity effect.

Similarly, we could imagine a well conducted randomized experiments performed in a large and representative sample of banks so that we do find a significant and causal effect of the number of women in boards on the amount of fraud. Yet, perhaps the magnitude of the effect is totally negligible in practice (no practical significance), say: each additional woman in a board decrease the total amount of fraud by 42 euros.

Another example The explained variable Y is the final grade obtained in an ENSAE subject. The treatment D is the number of hours worked on the subject, averaged over the semester. We wonder whether there is a causal effect of D on Y.

Figure 1.2: The more you work on a subject, the worse your grade on average… Why doing reviews?! Simple linear regression of Y on D (and a constant):

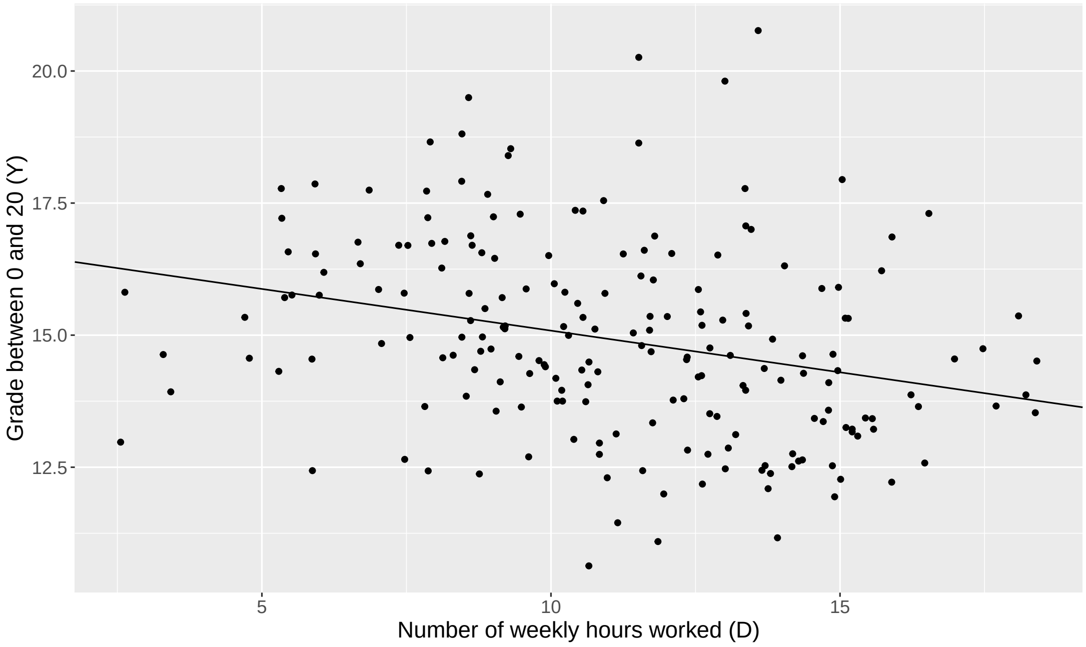


Figure 1.3: But maybe students do not randomly choose how much time they spend studying a subject! There is no randomized experiment on the amount of time spent per subject; in other words, the agents (in this case, the students) choose the "treatment" themselves. We can probably even think (a bold hypothesis) that they work harder in the more difficult subjects, where grades tend to be lower anyway. *Simpson's paradox*: paradoxical because it's obvious once you say it but very easily missed or forgotten before you see it, and, moreover, in most applications, we do not observe these relevant controls in the data.
Here, however, we observe as controls G the subject, shown in color below:

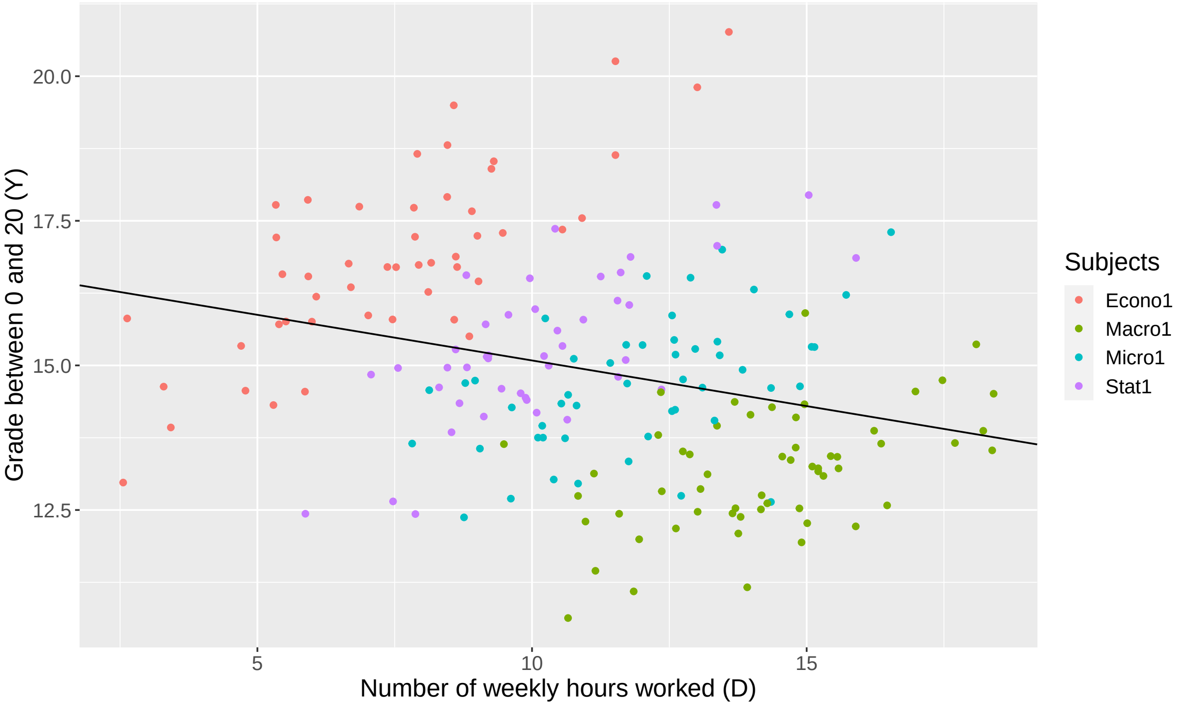
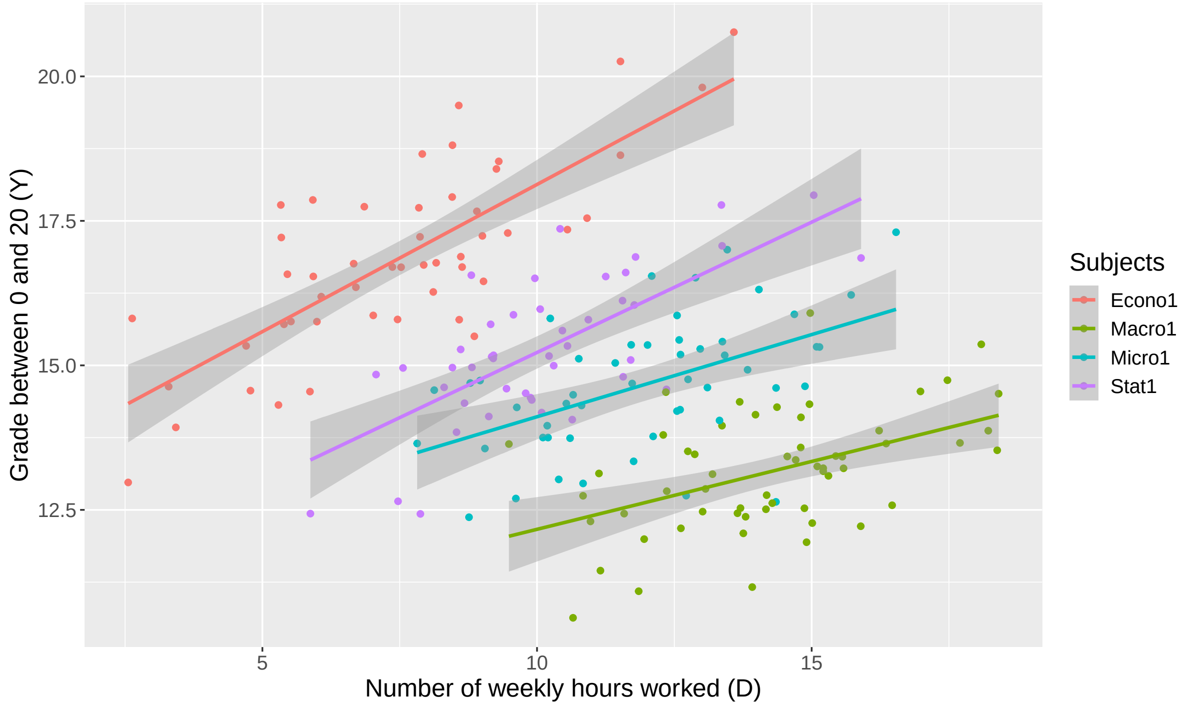


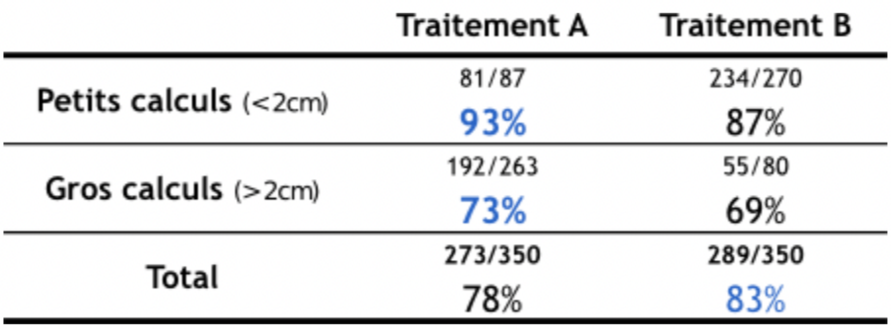
Figure 1.4: The previous simple linear regression suffers from a selection bias: it does not identify an average causal effect of D on Y. Indeed, we have an omitted variable bias if we do not control for the difficulty of each subject, which is correlated *both* with the treatment variable (the number of hours worked D) and with the explained variable (the grade Y). Phew, with this control, "work pays off": good reviews! Note: actually, the graph below does not present a multiple linear regression with G as a control, but separate regressions according to the modalities of G (discrete here):



Simpson's paradox The previous example is an instance of the Simpson's paradox. Reading/Watching: some additional references:

- (in French) un billet de blog (*lien*) et une vidéo (*lien*) de David Louapre de "Sciences Etonnantes" sur le paradoxe de Simpson, ou cette vidéo (*lien*) de Lê Nguyên Hoang de "Science4All" sur les facteurs de confusion.
- Wikipedia page on Simpson's paradox (*link*)

From the blog post, an example of Simpson's paradox:



	Traitement A	Traitement B
Petits calculs (<2cm)	81/87 93%	234/270 87%
Gros calculs (>2cm)	192/263 73%	55/80 69%
Total	273/350 78%	289/350 83%

Source: Charig, C. R., et al. « Comparison of treatment of renal calculi by open surgery... » British medical journal (Clinical research ed.) 292.6524 (1986): 879.

Main problem to keep in mind: the treatment are not allocated (as-if) randomly.
[In class: illustration and introduction, for a first view, of potential outcomes]

1.2.2 Tentative definition of econometrics

Definition We may define (a part of) econometrics as the field of study that answers the previous question: "In such situations where causal effects are evoked, what questions may/should you ask yourself?", that is, that tries to

(i) Determine what are the important questions to ask ourselves in settings where causal effects are evoked;

(ii) Formalize those questions by making links between real-life situations (as expressed in our ordinary language) and the mathematical properties (stated in a formal language) of random variables introduced to model the situation;

(iii) Answer them: provide conditions (at least sufficient, and even better if they are also necessary) that guarantee that the relations measured by proper statistical tools do recover causal relations;

(iv) Those issues are addressed by theoretical econometrics. Then applied econometrics use that methodology to measure causal effects of interest in some specific applications, notably to evaluate public policies and thus inform decisions made by policy-makers. By doing so, applied econometricians especially need to assess the plausibility of the conditions outlined by theoretical econometricians, that is, again, they need to make links between real-life situations and properties of random variables.

This tentative definition calls for various remarks.

Remark 1 The precision "(a part of)" comes from the fact that a typical dictionary-like definition of econometrics is broader[3]: the application of mathematical/statistical methods to the field of economics.

Contrary to the previous definition, that broader sense does not restrict to causal effects. Here, econometrics aims at using data to make predictions about economic objects (see details below) or to test economic theories. That objective of testing economic theories is notably sought by so-called *structural econometrics*, often opposed to *reduced-form econometrics*.[4]

Remark 2 In the previous definition, why is this field of study called *econo*metrics although there is no reference to economics or the economy? Why not sociometrics, biometrics, psychometrics, geometrics, etc.? The question is relevant, and there is no real reason apart that the course is part of a curriculum in economics. For that reason, we will mainly consider economic applications and examples, but the tools are more general, and can be used to study and quantify causal effects in any kind of applications: sociology, biology, psychology, geology, etc. In other words, another name for the course could be: causal inference.

1.2.3 Other key concepts

1.2.3.1 Use of data to make predictions and recent explosion of econometrics

Econometrics use data to make predictions (in a broad sense, see details below) and test (economic) theories.

Econometrics has massively developed since its beginning in the middle of the last century. This is due to "science" with the development of statistical and econometric methods, but also due to concrete, technological changes:

- the considerable development of data storage and computing capacities
- the increasing availability of (economic) data: survey data, administrative data, firm data.

[3] For instance, *Oxford Advanced Learner Dictionary* reads "the branch of economics concerned with the use of mathematical methods (especially statistics) in describing economic systems".

[4] I introduce that terminology here for your general culture in economics/econometrics, but there are various fuzzy definitions for that opposition, and I am uncertain the opposition is interesting by itself. If someone asked you to locate the course within that distinction, you could answer reduced-form econometrics.

Overall, the development of econometrics has participated to a deep change in the practice of the economics. Basically, at least for a non-negligible fraction of its sub-fields, economics has been moving from a theoretical to an empirical science with more and more attention devoted to data:

- See the very broad and fuzzily defined field of "applied econometrics". Here, essentially, researchers tries to answer empirical questions using data and theory, but the use of data is preponderant (as opposed to the sub-field of "theory" in economics, where mathematical models are preponderant).
- Example: 13% of academic articles in labor economics published between 1965 and 1969 were at least partly empirical (meaning: they use data), compared to 80% between 1994 and 1997 (and probably even an higher percentage nowadays).

1.2.3.2 Different types of predictions (causal or non-causal)

Econometrics allows to make predictions of different kinds.

Causal/counterfactual predictions Some predictions will be called "causal" in the sense that they require to build counterfactual: what would happen (ex ante) or what would have happened (ex post)?

The use of the word "causal" comes from the fact that, later in the course, following the current dominant paradigm in econometrics, causality will be jointly defined with "potential" or "counterfactual" variables; a causal effect := a difference between potential outcome.

Examples:

- Evaluation of existing policies (ex post analysis): reduction in class sizes. Prediction question: what would have happened in the absence of the policy? In this case, what would a student's performance has been, had the student been in a larger class?

- Evaluation of future policies (ex ante analysis): impact of a merger. Prediction question: what would happen in the presence of the policy? In this case, what would happen on prices if a merger between two companies were authorized by a competition regulator?

Non-causal predictions Another type of predictions is deemed "non-causal" in so far as these predictions are not concerned with counterfactuals: they are done in a "stable environment", without policy/structural change.

Examples:

- Using data on health, socioeconomic background and dates of death, how and what do we best predict the life expectancy of a 40-year-old woman holding a college degree? Same question for a 40-year-old woman holding a, high school degree?

- Using socioeconomic data, how and what do we best predict the wage of a 40-year-old woman holding a college degree? Same question for a 40-year-old woman holding a, high school degree?

If you were to bet, without any additional information, you will very likely predict an higher wage for the college-degree owner woman than for the high-school-degree owner woman? *Crucially, it does not matter whether the wage gap is genuinely caused by the difference in education or not.*

Wage difference between individuals with different schooling can be due to other factors than education (typical evoked confounders in this example: motivation, cognitive skills, etc.).[5]

Non-causal predictions do not care: they use education to predict wages because, be it by itself (there is a causal effect of education on wages) or not (it is a confounder, correlated with other factors explaining wages), education is a good predictor of wages = it has a good explanatory power of wages = there exist a significant (both statistically and practically) correlation between education and wages.

[5] You can also think of M. Spence's screening theory of the education system.

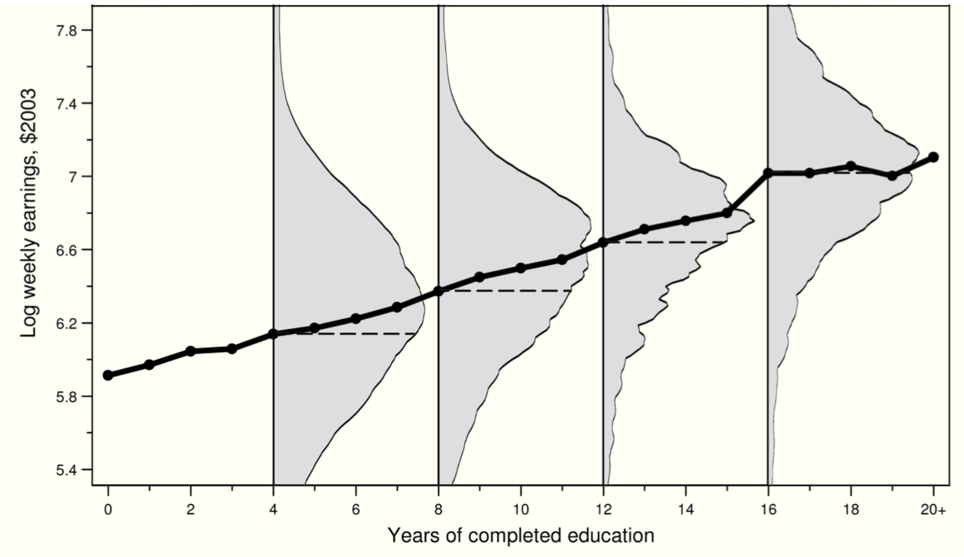


Figure 1.5: Reproduction of Figure 3.1.1 of MHE – Raw data and the conditional expectation function of average log weekly wages given schooling. The sample includes white men aged 40-49 in the 1980 IPUMS 5 percent file.

On the contrary, in causal predictions/analyses, we need to isolate the effect of education per se on wages. This is linked to the notion of "all other things equal" (*ceteris paribus*): all other things equal, what would, on average, the wage gap between high-school-degree owners and college-degree owners? Potential outcome variables aim to formalize this question.

1.2.3.3 Overfitting

We could be tempted to use as much explanatory variables as possible to improve predictions.

In causal predictions, we will see later why this can be a bad idea (included variable bias).

In non-causal predictions too, it can deteriorate predictions: in a model with too many variables, the predictions become unstable in the sense that they massively depends on the particular finite sample used, and often have difficulties to extrapolate to new data (out-of-sample prediction, which is the final aim – as opposed to in-sample prediction or fitted values).

That issue will not be a central part of the course, but it remains important.

overfitting, in French "sur-apprentissage" or rather "sur-interprétation".

The `R` script available on Pamplemousse presents examples of over- and under-fitting.

In particular, the following graph shows an example of "extreme" overfitting in the sense of the setting of Question 6: $n = k$ and $\sum_{i=1}^{n} X_i X_i'$ invertible, namely, we have as many degrees of freedom, free parameters, coefficients in the model as data points.

Therefore, in the same way we can always find a line connecting two points, we can always find a model that perfectly fits the data: for any observation i, $\widehat{Y_i} = Y_i$. In the graph, this corresponds to the fact that the green curve that represents the predicted curve of the model goes exactly through each black dot (the data points): the in-sample prediction is perfect ($R^2 = 1$)

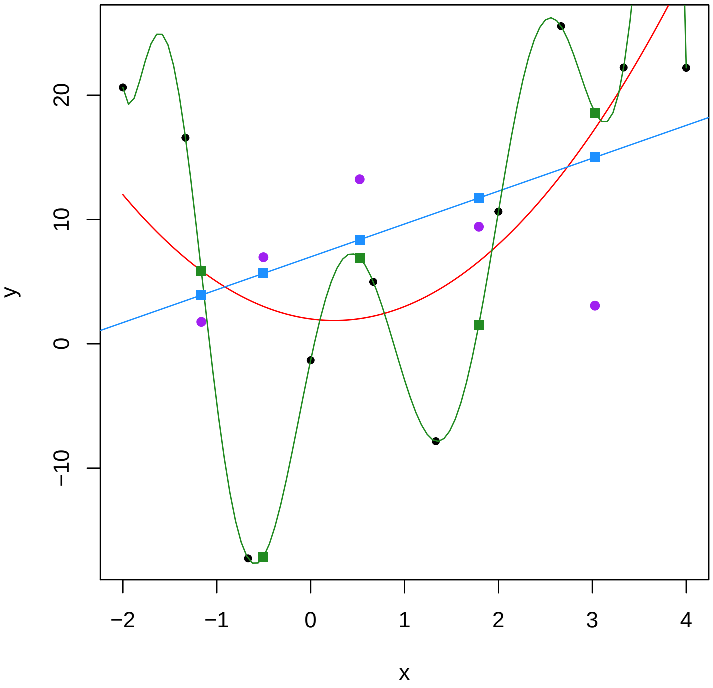
However, the out-of-sample prediction – predicting the outcome value Y_{new} from a *new* (not used to estimate $\widehat{\beta}$, to train/learn the model) observation X_{new} of covariates – can be quite bad. In that sense, the R^2 can be a misleading (over-optimistic) measure of the accuracy of a prediction.

Chapter 3 of the Econometrics 1 course will discuss this issue of over-fitting (see also the Machine Learning course in Semester 2). You can also look at the videos on Artificial Intelligence and Machine Learning by Lê Nguyên Hoang (Youtube channel "Science4All") (`link to the full playlist`), in particular

- episode 11 (`link to episode 11: ‘‘La sur-interprétation (overfitting)’’`)
- and episode 12 (`link to episode 12: ‘‘Fat Tony et Dr. John (biais-variance)’’`)

For a more general epistemological debate (related to that trade-off between over- and under-fitting) about theory versus empiricism, you can also look at this video (and his other videos on epistemology if you are interested) by Thibaut Giraud (Youtube channel "Monsieur Phi"):

- "Merci Captain Ad Hoc ! | Grain de philo #23" (`link`)
- Monsieur Phi's playlist on logic, epistemology, and language (`link`)



Explanations (See the related `R` script for details):

- Red curve: oracle, true model $\mathbb{E}[Y \mid X]$ here assumed to be quadratic in X; $Y = \mathbb{E}[Y \mid X] + \varepsilon = \beta_{00} \times 1 + \beta_{01} \times X + \beta_{02} \times X^2 + \varepsilon$, with $\mathbb{E}[\varepsilon \mid X] = 0$.
- Black dots: training sample $(Y_i, X_i)_{i=1,\ldots,n}$ used to learn/estimate two models (in the graph, $n = 10$).
- Purple dots: testing sample $(Y_{\text{new}}, X_{\text{new}})$ for out-of-sample prediction.
- Blue curve (first model that under-fits): simple linear regression of Y on X (polynomial of order 1 on X.

  The blue squares are the out-of-sample predictions made by this model 1.
- Green curve (second model that over-fits): multiple linear regression of Y on X, X^2, X^3, ... (polynomial of order $n-1$) to have as many parameters as data points in the training sample and thus be in the setting of Question 6 with perfect in-sample prediction.

  The green squares are the out-of-sample predictions made by this model 2.

The following image is another (artistic) representation of overfitting:

1.2.3.4 Use for decision-making and quantifying uncertainty (through probabilistic modeling)

We will simply allude to this point, but it is important. Often (and especially in economics for economic topics) but not always, we are not interested in causal effects only to describe the world, but to inform decisions by policy-makers.

The resulting decisions may differ depending on the amount of uncertainty around them (decision-makers being, in general, not neutral with respect to risk).

Example: whether to reduce or not by five students average class size is likely to differ if the causal effect of such a reduction is an increase of students' reading ability (through some normalized PISA – Programme for International Student Assessment – score, say) by[6]

- 10% ± 2%;
- 10% ± 10%;

in the latter case, the positive causal effect could simply due to chance, that is, comes from the particular random realization of the sample used in the regression analysis.

The quantification of uncertainty is also important in predictions without causality (which, can also inform decisions).

That question is tackle with a probabilistic modelling and the use of statistical tools to conduct inference (confidence regions and parametric tests).

[6] We will give later more precise signification of this ±, notably through the notion of confidence regions/intervals.

1.3 Some probability and statistics reviews for econometrics

To be completed.

To begin with, see the two documents distributed in class.

Chapter 2

Linear regressions – Ordinary Least Squares

2.1 Setting and notation

In this chapter, we present/review the fundamental of linear regressions. From now on, the objective is to predict/model/explain a variable, generically denoted Y, by a *linear* combination of other variables, denoted X.

Various names are used to call those variables.
The variable Y is referred to as

- the outcome variable,
- the explained variable,
- the dependent variable.

The variables X are referred to as

- the covariates,
- the explained variables,
- the regressors,
- the independent variables.

I will use those expressions as synonyms, except for the terminology of "dependent" and "independent" variables that is too dangerous by confusing with the mathematical properties of independence or non-independence between random variables.

Key general principle (nature of objects - NO). Always be aware of the nature and dimension of objects. In particular, is it stochastic or not?

By default, the vectors (be they stochastic or not) are column matrices, and $\cdot'$ denotes transposition.

For any pair of sets E and F, F^E is the set of applications/functions from E into F. We take advantage of that standard notation to underline the distinction between, on the one hand, random variables or random vectors that are stochastic and, on the other, non-stochastic parameters.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ the underlying probability space over which all random elements that we will consider are defined. Do not panic, we will not really need it. Besides, most often in applied mathematics as statistics using probability tools, the underlying probability space is left implicit.

In fact, we use it mostly for that notation: $\mathbb{R}^\Omega$, the set of functions defined on Ω with values in $\mathbb{R}$. A real random variable over Ω is an element of $\mathbb{R}^\Omega$ that is measurable with respect to the sigma-algebra $\mathcal{A}$. Here, we will not care of that matter of measure theory. We only use the notation $Y \in \mathbb{R}^\Omega$ to say that Y is a real random variable, and thus modelled as stochastic. In contrast, soon, we will define β_D the theoretical coefficient of the slope in a simple linear regression. Crucially, in our frequentist (as opposed to bayesian) setting, $\beta_D \in \mathbb{R}$, that is, β_D is a real number non-stochastic. Granted it is unknown, but it is not stochastic.

Here and henceforth, $Y \in \mathbb{R}^\Omega$ is a real random variable. We will briefly discuss the case where the support of Y, $\text{Support}(Y) = \{0, 1\}$, that is the outcome variable is binary (being employed or unemployed among active people in the job market, for instance). However, most of the time, we have in mind continuous real random variable (meaning continuous with respect to Lebesgue measure), quantitative, they have a quantitative meaning (as opposed to qualitative).

Here and henceforth, $X \in (\mathbb{R}^d)^\Omega$ is a real random *column*[1] vector of dimension $d := \dim(X) \in \mathbb{N}^*$. For $j \in \{1, \dots, d\}$, we denote by X_j or X^j (depending on context) the j-th component of that vector; in other words

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix}, \quad \text{with transpose} \quad X' = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix}' = (X_1, X_2, \dots, X_d).$$

Remark that, for $j \in \{1, \dots, d\}$, $X_j \in \mathbb{R}^\Omega$ is a real random variable.

The regressors can be continuous or discrete random variables with

- quantitative (example to come: an average temperature expressed in Celsius degrees), or
- qualitative (example: in the wage equation, gender or the type of contract, such fixed-term – CDD – or without fixed-term – CDI))

meaning. We will discuss soon how to include such quantitative regressor (an indicator variable for each modality[2] except one, treated as the reference modality) and how to interpret their coefficients in linear regressions.

The result we will see in this chapter are unaffected.

Finally, very often, we include a so-called constant or intercept among the regressors, which, formally, is simply a degenerate constant random variable, for simplicity, equal to 1. In this case, we have $X_1 = 1$ and

$$X = \begin{pmatrix} 1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix}$$

Note that choosing the first component to put the constant has no importance: the order of the regressors has no impact.

We will see soon why we do that as a default.

That being said, if not stated otherwise, the general results we will see for the linear regression of Y on X hold whether there is a constant or not.

We will distinguish

- *multiple* linear regressions (MLR) where $\dim(X) > 2$;
- *simple* linear regressions (SLR) where $X = (1, D)'$ with $D \in \mathbb{R}^\Omega$, that is, the regressors are made of only a constant and a scalar/real explanatory variable – $\dim(X) = 2$ with a constant.
  Within the particular case of simple linear regressions, a particular case is when D is a binary variable: $\text{Support}(D) = \{0, 1\}$.

Remark about randomness/stochastic modelling: it comes from the act of random sampling: before an individual is drawn from the population, we do not know what its associated variables Y and X will turn out to be; yet, we can assign a joint probability $\mathrm{P}_{(Y,X)}$ to the couple (Y, X).

Example: wage equation. Again, remember that the tools we present can be applied to any applications. Nonetheless, as one of the leading examples, we will consider a traditional econometrics study where the outcome is wage and the regressors include: age, experience, the number of years of education, etc.

Key remark: linear means linear in parameters.

Considering only *linear* combination may seem rather restrictive. It is restrictive, but much less than at first sight. Indeed, a key point to keep in mind is that linear regressions are linear *in parameters.*

[1] Following the standard convention that identifies vectors of dimension d with column matrix of dimension $d \times 1$.

[2] I tend to use the word "values" for the values taken by a quantitative variable as opposed to "modalities" for a qualitative one.

The meaning of that is the following: from initial variables of interest, you can specify transformations of those variables so that the model is a linear regressions of the transformed variables, yet entirely able to capture non-linear relations between your initial variables of interest.

Example [wage equations] Consider a simplified model where Y is the wage (in practice, it should, of course, be further specified, say the net monthly income before income tax) and among X there is the number of year of education. Imagine we consider France in the 2020s.

A strictly speaking linear model implies that, for each additional year of education, we predict the same change (a priori, an increase) in wage. In particular, it says, for instance, that, as final study, going from "première" (complete only the penultimate year of high school) to "terminale" (that is, complete last year of high school and get the high school diploma) has the same effect as going from L2 to L3 (obtaining an undergraduate diploma) or to M1 to M2 (obtaining a graduate diploma).

Given the concrete conditions of job market in the population of interest, a more realistic modeling is that each additional year of education increases the wage by a given percentage, the same for any year (relative increase as opposed to an absolute increase in the previous linear model between wage and the number of year of education).

There is no issue in embracing that modeling in a linear regression. It suffices to consider for the outcome variable the logarithm of the wage.

Formally, write with tilde the initial, fundamental variables of interest:

- $\widetilde{Y}$ is the wage
- $\widetilde{X_1}$ is the number of year of education.

Define

- $Y := \log(\widetilde{Y})$ as the outcome varialbe (logarithm of wage)
- $X_1 = \widetilde{X_1}$ as one of the covariates, the number of year of education.

In this "log-level model", a linear regression specifies a linear relation between Y and X_1, which entails a non-linear relation between the genuine variables of interest $\widetilde{Y}$ and $\widetilde{X_1}$. We will present those models and their interpretation more in-depth later.

Another example $\widetilde{Y}$ = consumption of electricity for heating or cooling home $\widetilde{X_1}$ = average temperature. Consider $X := (1, \widetilde{X_1}, \widetilde{X_1}^2)'$ to allow for quadratic effect of average temperature on the consumption of electricity.

Appendix A

Temporary work in Progress – elements to be put somewhere

A.1 Notation

We try to stick to the following conventions for consistency and easier understanding. Stand-alone Roman uppercase letters such as D, Y, Z, or X are used to denote random variables *observed* by the econometrician as opposed to a function-type notation such as $Y(d)$ or $D(z)$ for *potential* variables, which are random variables, one of which are observed by the econometrician while the others are counterfactual. Lowercase Roman letters are typically used as "free variables"; beware of this terminology: they are not stochastic but denote arbitrary values, for instance, any possible value $d \in \text{Support}(D)$ taken by the treatment. For examples, if the treatment is binary, $d \in \{0, 1\}$; if the treatment is a (non-negative) price, $d \in [0, +\infty)$. Lowercase Greek letters generally denote non-stochastic unknown parameters or estimands. Following standard notation, the exception is the letters ε, ν, or sisters that denote disturbances or error terms: random variables not observed by the econometrician. Also, λ, μ, ω, and α are used to denote non-stochastic weights. The uppercase Greek letter Δ denotes the individual causal effect; it is a random variable (hence the uppercase). Lowercase δ, possibly with sub- and superscripts, is used to refer to causal average effects, which are the primarily parameters of interest here (and γ for quantile effects). In contrast, the letter β is typically used to denote estimands, automatically identified from the data, as opposed to the δ for which this requires identifying assumptions.