Structural summary of Econometrics 1 course

(Lucas Girard) – This version: 15 January 2024

English version (French version available on Pamplemousse)

Preamble *The following document is not an exhaustive account of the course and, in particular, only encompasses some of the elements that can be asked at the exam. Rather, it aims at emphasizing the main structure and outlines of the course to ease its overall understanding for your reviews of the exam and, first and foremost, for your long-term comprehension and memorization of the important notions and good reflexes, studied in this Econometrics 1 course, to keep in mind in front of any causal data analysis.*

Notation The notation used below follows that of the course (see also the reminders about notation at the beginning of the quizzes if needed)
As we always assume to have independent and identically distributed (i.i.d) samples, remember that we often omit the observation index i to denote a generic instance with the same distribution as the data. For example, $(Y, D) \sim \mathrm{P}_{(Y,D)}$ denotes a generic instance of the couple outcome × treatment that has the same distribution, denoted $\mathrm{P}_{(Y,D)}$, as the distribution of any observation (Y_i, D_i), $i \in \{1, \dots, n\}$, where n is the sample size.
It is crucial to be well aware of the *nature* and *dimension* of the objects.
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 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. For example, the individual causal effect $\Delta \in \mathbb{R}^\Omega$ is a real random variable, that is, formally, a measurable map from Ω into $\mathbb{R}$, where $(\Omega, \mathcal{A}, \mathbb{P})$ is the underlying probability space on which are defined all the random variables that we consider; in contrast, $\delta := \mathbb{E}[\Delta] \in \mathbb{R}$, the average treatment (or causal) effect (ATE), is a non-stochastic real number (frequentist setting). Remember that an estimator (as a function of the observations, which are modeled as stochastic) is a random variable. Example: the OLS estimator, $\widehat{\beta} \in (\mathbb{R}^{\dim(X)})^\Omega$.

Contents

1 First part – Tools from probability and statistics: OLS **2**
1.1 Estimation – Chapter 1: the fundamentals of linear regressions 2
1.2 Inférence – Chapter 2: statistical uncertainty in linear regressions 4
1.3 Use for prediction (in stable environments) – Chapter 3: linear regressions and non-causal predictions 7

2 Second part – definition of causal effects and links with linear regressions – Chapter 4: linear regressions and causality, Sections 1 and 2 **8**
2.1 Formalization of the notion of causality through potential outcome variables 8
2.2 Non-causal linear representation (theoretical linear projection) and causal representation: do they coincide? That is, is there a selection bias? 9

3 Third part – two identification strategies to recover an average causal effect despite (unconditional) selection bias **13**
3.1 Add the adequate control variables – Chapter 4, Sections 3 and 4 13
3.2 Rely on a valid instrument – Chapter 5 : instrumental variables 13
3.3 (Anticipated) Development: they exist other identification strategies – Econometrics 2 in Spring semester and third year econometric courses 15

4 Complements and conclusion **16**
4.1 To prepare specifically for the theoretical exercises 16
4.2 An example 16
4.3 To sum up the summary (in 4 images and about 160 words) 18

1 First part – Tools from probability and statistics: OLS

1.1 Estimation – Chapter 1: the fundamentals of linear regressions

Theoretical linear regression (or projection) Under mild moment conditions (that, hereafter and in the slides, problem sets, and exams, are always assumed if the contrary is not explicitly stated):

- a finite second-order moment for the real explained / outcome random variable Y: $\mathbb{E}[Y^2] < +\infty$;
- a finite second-order moment for each component of the random vector of covariates / explanatory variables / regressors X : $\mathbb{E}[\|X\|^2] = \mathbb{E}\left[\sum_{j=1}^{\dim(X)} X_j^2\right] < +\infty$ (where X_j) denotes the j-th component of X):
- no perfect colinearity among regressors: $\mathbb{E}[XX']$ invertible;

we can *always* define the *theoretical* linear regression of $Y \in \mathbb{R}^\Omega$ on $X \in (\mathbb{R}^{\dim(X)})^\Omega$:

$$Y = X'\beta_0 + \varepsilon \text{ with } \mathbb{E}[X\varepsilon] = 0,$$

where $\beta_0 := \mathbb{E}[XX']^{-1}\,\mathbb{E}[XY] \in \mathbb{R}^{\dim(X)}$ is a (column) vector that is non-stochastic (it is a parameter) and function of the joint distribution $\mathrm{P}_{(Y,X)}$ of the couple (Y, X), and $\varepsilon := Y - X'\beta_0 \in \mathbb{R}^\Omega$ is a real random variable.

The theoretical linear regression can be seen as an orthogonal projection, within the Hilbert space L^2 of random variables admitting a finite variance, of the variable Y on the space of *linear* functions of X. $\mathbb{E}[X\varepsilon] = 0$, which is equivalent to X and ε orthogonal, is the associated orthogonality condition.

Estimation by OLS β_0 can be estimated by the corresponding (empirical – as opposed to theoretical) linear regression:

$$\widehat{\beta} := \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) = \widehat{\mathbb{E}}[XX']^{-1}\,\widehat{\mathbb{E}}[XY],$$

the Ordinary Least Squares (OLS) estimator of Y on X (remember that $\widehat{\beta} \in (\mathbb{R}^{\dim(X)})^\Omega$ is a stochastic vector since it is an estimator, thus a statistic, namely a measurable function of the observations, which are modeled as stochastic; in other words, we can compute a realization of the estimator $\widehat{\beta}$, an estimate, from a given sample), under the above-mentioned moment conditions and with independent and identically distributed samples $(Y_i, X_i)_{i=1,\dots,n} \overset{\text{i.i.d.}}{\sim} \mathrm{P}_{(Y,X)}$ (i.i.d sampling is also always assumed if not state otherwise in the course slides, problem sets, exams, quizzes, and in this document),

- is well defined with proability approaching 1 as n goes to infinity (w.p.a.o), and
- converges in probability to the theoretical coefficient β_0, that is, $\widehat{\beta}$ is a consistant estimator of the parameter β_0.

Besides, the empirical equivalent of the orthogonality condition of the theoretical regression holds: $\widehat{\mathbb{E}}[X\widehat{\varepsilon}] = \overline{X\widehat{\varepsilon}} = n^{-1}\sum_{i=1}^n X_i\widehat{\varepsilon}_i = 0$ where $\widehat{\varepsilon}_i := Y_i - X_i'\widehat{\beta}$ is the (estimated)[1] residual of observation $i \in \{1, \dots, n\}$.

Key reference for that first part of the course: **Chapter 1, Proposition 5** (definition of theoretical linear projection / regression and consistent estimation by OLS).

There is no (not yet) causality At this stage, we do *not* consider causality questions; we only have a prediction problem of Y using linear functions of X.

In that perspective, the R^2 of the regression (Chapter 1, slides 12 and 21) quantifies the quality of the prediction of Y through a linear function of X: $\widehat{Y} := X'\widehat{\beta} \xrightarrow[n\to+\infty]{P} X'\beta_0$.

[1] In English, by default, *residual* means estimated residuals (“résidus estimé” in French) while *error term* corresponds to the “theoretical residual” (“résidus” sans adjectif) in French.

Conditional expectation and theoretical linear regression In general, there is absolutely no reason that $\mathbb{E}[Y \mid X] = X'\beta_0$, namely, that the conditional expectation of Y given X be linear.

However, the theoretical linear regression of Y on X, that yields the prediction $X'\beta_0$, is the best (with respect to mean square error / L^2 norm) *linear* approximation of $\mathbb{E}[Y \mid X]$ (Chapter 1, Proposition 5, second equality of point 2): $\beta_0 = \arg\min_{b \in \mathbb{R}^{\dim(X)}} \mathbb{E}\left[(\mathbb{E}[Y \mid X] - X'b)^2\right]$

An important particular case: simple linear regressions when X is simply made of a constant / intercept and a univariate / scalar regressor, that is, a real random variable $D \in \mathbb{R}^{\Omega}$, that is, formally, when $X = (1, D)'$, the OLS estimator of the slope, denoted $\widehat{\beta}_D$, that is the estimator of the theoretical coefficient associated with D (writing $\widehat{\beta} = (\widehat{\alpha}, \widehat{\beta}_D)'$) in the linear regression of Y on D (and on a constant, which is always implicitly present if there is no contrary indication) is equal to

$$\widehat{\beta}_D := \frac{\widehat{\mathbb{C}\text{ov}}(Y,D)}{\widehat{\mathbb{V}}[D]} = \frac{(n-1)^{-1}\sum_{i=1}^n (Y_i - \widehat{\mathbb{E}}[Y])(D_i - \widehat{\mathbb{E}}[D])}{(n-1)^{-1}\sum_{i=1}^n (D_i - \widehat{\mathbb{E}}[D])^2},$$

and, under the classical moment conditions, it converges in probability to the empirical counterparts:

$$\widehat{\beta}_D \xrightarrow[n \to +\infty]{P} \beta_D := \frac{\mathbb{C}\text{ov}(Y,D)}{\mathbb{V}[D]},$$

denoting, in this case of a simple linear regression, $\beta_0 = (\alpha_0, \beta_D)'$ (Chapter 1, slide 34).

Particular case of the particular case : if, in addition, D is binary (Support$(D) = \{0,1\}$), then

$$\beta_D = \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0],$$

and the corresponding empirical expression with $\widehat{\mathbb{E}}[\,\cdot\,]$ also holds for $\widehat{\beta}_D$ (Chapter 1, slides 9 et 34).

Frisch-Waugh theorem and links between simple and multiple linear regressions That particular case provides a simple and quite understandable, intuitive

- (*stochatic estimators*) of the OLS estimator, $\widehat{\beta}_D \in \mathbb{R}^{\Omega}$, a real random variable,
- (*non-stochastic parameters*) and of its probability limit, the theoretical coefficient $\beta_D \in \mathbb{R}$, a non-stochastic real number,

of the slope in a simple linear regression (SLR): it is more or less (modulo a positive multiplicative factor) the (linear or Pearson's) correlation between the explained variable Y and the single explanatory variable or regressor D, since

$$\beta_D := \frac{\mathbb{C}\text{ov}(Y,D)}{\mathbb{V}[D]} = \underbrace{\frac{\mathbb{C}\text{ov}(Y,D)}{(\mathbb{V}[D]\mathbb{V}[Y])^{1/2}}}_{=\mathbb{C}\text{orr}(Y,D)} \times \underbrace{\frac{\mathbb{V}[Y]^{1/2}}{\mathbb{V}[D]^{1/2}}}_{\text{constant } \geq 0},$$

and the same equality holds with the empirical counterparts (symbolically, adding the hats) for $\widehat{\beta}_D$.

In comparison, for multiple linear regressions (MLR), the expressions of the theoretical parameter, β_0, and of its OLS estimator, $\widehat{\beta}$, provide an explicit expression, yet less intuitive. Frisch-Waugh Theorem (F.W) can be seen as a way to address that issue, and its result is sometimes referred to as the formula of "the anatomy of a (multiple) regression". Indeed, that theorem enables us to express, for a given univariate regressor of interest, its theoretical coefficient (or its empirical OLS estimator) as the coefficient (implicitly, the slope coefficient) of a specific *simple* linear regression.

Statement of the theoretical version of the theorem (Chapter 1, Proposition 6) – idem with empirical counterparts in the "OLS estimators" version (Chapter 1, Proposition 3). If we are interested in a particular regressor $X_j \in \mathbb{R}^{\Omega}$ (a real random variable), with $j \in \{1, \ldots, \dim(X)\}$, then the coefficient (a non-stochastic real number) associated with X_j in the theoretical linear regression of Y on X, denoted $\beta_{0j} \in \mathbb{R}$, the j-th component of β_0, is equal to

$$(\beta_0)_{j\text{-ème comp.}} \overset{\text{noté}}{=} \beta_{0j} \overset{\text{F.W}}{=} \frac{\mathbb{C}\text{ov}(Y, \xi)}{\mathbb{V}[\xi]} = \text{theoretical (slope) coefficient in the SLR of } Y \text{ on } \xi,$$

where ξ is the error term in the theoretical linear regression of X_j on all the other components of X, that is, on the random vector of dimension $\dim(X) - 1$ often denoted X_{-j}.

ξ can be interpreted as the "residual" or "net" variation remaining in X_j once subtracted the variation that can be explained by linear functions of X_{-j}. β_{0j} is thus interpreted as, modulo a positive multiplicative factor, the correlation between the explained variable Y and the regressor X_j net of its variations explained by linear functions of the other regressors X_{-j}.

Hence, also, the usual formulations of marginal effects "all else being equal", "keeping fixed the values of the other regressors" when interpreting multiple linear regressions.

Another important result linking SLR and MLR is the so-called "omitted variable bias" formula although, at this stage, it is noteworthy that there is no notion of bias: the formula is only an algebraic relationship between a "short" regression and a "long" regression:

- Chapter 1, Proposition 4 for the empirical version with OLS estimators,
- Chapter 1, Proposition 7 for the theoretical version with the coefficients of the theoretical linear regression.

Beyond particular notations, it is interesting to remember "with words" to memorize it better (and this is a good habit more generally for any result). In a nutshell (see Quiz 1, Question 5 for further details), it says:

$$\textit{short} = \textit{long} + \textit{omitted} \times \textit{coefficients of omitted on included}.$$

1.2 Inférence – Chapter 2: statistical uncertainty in linear regressions

Having a consistent estimator $\widehat{\beta}$ of β_0 is one thing. Yet, it remains an estimator, and thus, for a given sample, a single realization of that estimator ("estimateur"): an estimate ("une estimée" in French). What can be deduced and learned about the unknown parameter β_0? This is the inference question studied in Chapter 2 of the course.

Asymptotic normality of the OLS estimator The fundamental result of that chapter is, under moment conditions that remain mild and i.i.d sampling, the asymptotic normality of the OLS estimator $\widehat{\beta}$ around the theoretical coefficient β_0 (**Chapter 2, Theorem 1**):

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow[n \to +\infty]{d} \mathcal{N}(0, \underbrace{\mathbb{E}[XX']^{-1}\,\mathbb{E}[\varepsilon^2 XX']\,\mathbb{E}[XX']^{-1}}_{=: V_a(\widehat{\beta}) \text{ the asymptotic variance of } \widehat{\beta}}).$$

That result is obtained *without assuming* homoscedasticity of the error term ε (with respect to regressors X): $\mathbb{E}[\varepsilon^2 XX'] = \mathbb{E}[\varepsilon^2]\,\mathbb{E}[XX']$ (Chapter 2, Equation (Hom)). This is nice because the homoscedasticity condition is often restrictive in practical applications by imposing that the magnitude of the prediction error is not correlated with the regressors, that is, heuristically, that there is as much variation, dispersion in the Y whatever the value of X. A classical counter-example: Y = wage, X = number of years of education, for instance, in France in the 2020s.

As a default (and contrary to Stata's default, be careful), we thus use the option `robust` in the Stata command `regress` in order to compute standard errors (that are used in tests and confidence intervals) that are robust to heteroscedasticity.

Precision of the OLS estimator As corollaries of Theorem 1, the first section of Chapter 2 presents two results that provide a more intuitive expression of the asymptotic variance of the OLS estimator in a simplified setting that assumes homoscedasticity.

(SLR) Chapter 2, Equation (2): in a simple linear regression, $X = (1, D)'$, the asymptotic variance of the OLS estimator $\widehat{\beta}_D$ of the slope is

$$V_a(\widehat{\beta}_D) = \frac{\mathbb{E}[\varepsilon^2]}{\mathbb{V}[D]}.$$

(MLR) Chapter 2, Proposition 1: in a multiple linear regression, the asymptotic variance of the OLS estimator $\widehat{\beta}_j$ of a given coefficicent β_{0j} associated with the univariate regressor X_j is

$$V_a(\widehat{\beta}_j) = \frac{\mathbb{E}[\varepsilon^2]}{(1 - R^2_\infty)\mathbb{V}[X_j]},$$

where R^2_∞ is the probability limit of the R^2 in the regression of X_j on the other regressors X_{-j}.

Those two expressions are valid under the homoscedasticity hypothesis. However, more generally, the two results can be understood as a way to give insight, make more intuitive and understandable[2] the expression of the asymptotic variance(-covariance matrix) $V_a(\widehat{\beta})$.

Qualitatively, regarding the main drivers, those results remain valid more generally, without assuming homoscedasticity. Thus, they show that the asymptotic variance of the OLS estimator

- increases (that is, the precision of the OLS estimator decreases) when the variance of error terms increases;
- increases (the précision decreases) when the regressor of interest X_j is more correlated with the other regressors X_{-j};
- decreases (the précision increases) when the variance of the regressor of interest increases.

Tests and confidence intervals Provided consistent estimation of the asymptotic variance $V_a(\widehat{\beta})$, and we can do so (at the cost of an additional moment condition) by replacing the unknown theoretical expectations $\mathbb{E}[\cdot]$ by the corresponding sample means $\widehat{\mathbb{E}}[\cdot]$ and the unobserved error term ε by the residuals $\widehat{\varepsilon}$ (Chapter 2, Proposition 2), the asymptotic normality of $\widehat{\beta}$ enables to build tests and confidence intervals for β_0 (or components thereof) that are robust to heteroscedasticity and have the proper asymptotic guarantees:

- Chapter 2, Proposition 3: bilateral simple tests
- Chapter 2, Proposition 4: unilateral simple tests
- Chapter 2, Proposition 5: “multiple” or “joint” tests
- Chapter 2, Equations (8) et (9): confidence intervals (for scalar parameters) and confidence regions (for multivariate parameters), they have connexions with tests (Chapter 2, slide 33).

[2] Similarly to Frisch-Waugh's theorem (in an estimation perspective as opposed to the inference question studied here) that, for a given component, makes more understandable the expression of the random vector $\widehat{\beta}$ (empirical version of F.W) or of the non-stochastic vector β_0 (theoretical version of F.W).

Interpretation of a régression In practice, it is expected (and often asked!) to know how to read and comment the Stata output of a regression (see the different Problem Sets and, for instance, Quiz 2, Question 11) notably by discussing, for the coefficient associated with some specific given regressor X_j, the following three elements:

1. The sign of the coefficient (qualitative interprétation): is it expected? surprising? What is the direction of the effect in terms of prediction? Later, once causality has been introduced, given the sign, is it sensible to think that it is the estimate of an (average) causal effect?

2. Its statistical significance: these are the answers, or the p-value to summarize, to the question of the simple bilateral test of nullity of the associated theoretical coefficient: $H_0 : \beta_{0j} = 0$ against $H_1 : \beta_{0j} \neq 0$. Reminder: the *p-value* of a test is the lowest level for which the null hypothesis H_0 is rejected against the alternative hypothesis H_1 (Chapter 2, slides 21 and 25).

   Exemple. If the p-value of that test is equal to $0.03 = 3\%$, then:

   - $3\% > 1\%$: X_j is said to be *not* statistically significant at 1%;
   - $3\% < 5\%$: X_j is said to be statistically significant at 5% (and, a fortiori, at 10% or at any higher level).

   Typically, if the p-value is lower than 1%, X_j is said to be statistically significant at any usual level (most people think of 1%, 5%, 10%). Conversely, if the p-values is larger than 10%, X_j is said to be not statistically significant at any usual level.[3].

3. The value of the coefficient, its practical significance (quantitative interprétation, and not only a qualitative interpretation of the sign of the coefficient). To do so:

   - Be careful about the model:
     (i) level-level, log-level, level-log, log-log?
     (see Quiz 4, Question 5 for the different interpretations)
     (ii) are there regressors that are present in the regression with powers or interactions
     (see Quiz 1, Question 12 about *marginal effects*)
   - Be careful also about the unit in which are expressed the variables Y and X_j.
     Reminder: percentage points are used to state the *absolute* variation of a percentage, one p.p. $= 0.01$. Example: going from 20% to 28% is an *absolute* increase of 8 percentage points and a *relative* increase of 40% since $0.40 \times 0.20 = 0.08 = 28\% - 20\%$.
     We can always (by the construction of OLS and theoretical linear projection) interpret the results in terms of predictions, *but, in contrast, not necessarily in terms of causal effects!* (see Part 2 below). Hence, that kind of helpful formulation for quantitative interpretation of the value of an estimated coefficient: "All else being equal, for an increase of . . . (specifying in detail the units), we predict an increase/decrease of . . . (again with specification of the units)".

Inference under stronger assumptions The third section of Chapter 2 ("Particuliar cases*") presents two cases in which we can expect an improved inference in the sense of more precise: confidence intervals with the same (asymptotic) level, but with narrower length. However, the cost is stronger hypotheses:

- homoscedasticity (often restrictive in practical applications);
- Gaussian / normal error terms independent of regressors (often very restrictive).

The second case also permits to build tests and confidence intervals that are *exact*, namely valid *non-asymptotically*, for any sample size n as opposed to guarantees that are only asymptotic when $n \to +\infty$. For different (arguable, but not discussed here) reasons, the dominant paradigm in the course for inference is *asymptotic* and robust to heteroscedasticity.

[3] Statistical significance and p-values are important and useful notions. Yet, they also have drawbacks and are criticized (or, at least, their bad use) – outside the course, see nonetheless Quiz 2, Question 15.

1.3 Use for prediction (in stable environments) – Chapter 3: linear regressions and non-causal predictions

Non-causal predictions as opposed to causal (counterfactual) predictions Chapter 3 studies the use of OLS for non-causal prediction *in a stable environment*, which formally means:

$$(Y_{n+1}, X_{n+1}) \overset{d}{=} (Y_i, X_i),\ \forall i \in \{1, \ldots, n\} \overset{d}{=} (Y, X) \sim \mathrm{P}_{(Y,X)} \text{ (a generic instance w. the same distrib.)},$$

that is: the new variables, indexed by $n+1$, for which we seek to predict the component Y using X, are *not* (this is *out-of-sample prediction*) in the initial sample whose observations are indexed by $i = 1, \ldots, n$, but they have the same distribution.

This setting of non-causal predictions differs from the setting of so-called "counterfactual" predictions (Chapter 0, Section 2). Those counterfactual predictions enable to quantify causal relationships for *causality will be defined in terms of counterfactual variables called potential outcome variables.*

In contrast, regarding non-causal predictions in stable environments, there are no questions about causality. In other words (see Quiz 3, Question 10 for further details), we are not trying to discover and explain the underlying, genuine reasons that explain why X can be useful to explain Y; *we are only trying to predict Y as precisely as possible (in terms of mean square error typically) using functions (here, linear functions using OLS) of X.*

Main issues You will study more in-depth that question in your statistical learning (or machine learning) courses. Chapter 3 can be seen as an introduction (with already some contents!) to that question in the specific case of predictions based on linear models. Despite that restriction, several important ideas or notions are presented and remain relevant and crucial in a more general setting (not restricted to linear predictions):

- Trade-off in the out-of-sample prediction error (Chapter 3, Theorem 1) between

  (i) A "bias" term, decreasing in the complexity / richness of the model.
  Underlying interrogation: neglecting statistical uncertainty, is the model rich enough, realist enough so that, on average, over possible samples, the obtained predictions coincide with the optimal / oracle prediction (in the case of MSE, with the conditional expectation)?

  (ii) A "variance" term, increasing in the complexity / richness of the model.
  Underlying interrogation: to what extent do the estimated model and the resulting predictions vary or, on the contrary, are rather stable depending on the particular sample drawn?

  Another formulation of that trade-off is the opposition between

  (i) *Under-fitting*: the model is not rich or complex enough; the bias term prevails.

  (ii) *Over-fitting*: the model is too complex; it learns random fluctuations by "sticking to" the particular data of the training sample used that are irrelevant and, consequently, fails to generalize its prediction in out-of-sample setting properly; the variance term prevails.

  Theorem 2 of Chapter 3 provides an expression of those bias and variance terms under further assumptions (*objective of simplification to get more understandable, intuitive expressions*).

- The principle of cross-validation to select a model with the split of the initial sample into a training / estimation set and a set for validation / test.

- The idea of penalized regressions and, more generally, of penalization:
  - "norm 0" penalization: selection among regressors and information criteria; it entails computational issues since the resulting optimization program is not convex;
  - to address that issue, idea of convex relaxation: consider an optimization program close to the initial problem, but convex, hence easily solved ⟶ norm 1 (Lasso) penalization.
  - norm 2 (Ridge regression) penalization.

2 Second part – definition of causal effects and links with linear regressions – Chapter 4: linear regressions and causality, Sections 1 and 2

Genuine start of econometrics That second part corresponds to the first two sections:

- the case of a single binary variable
- the case of a single non-binary variable

of Chapter 4: linear regressions and causality.

In a way, the course of econometrics really starts here. Following the first three chapters (Part 1) that are essentially about statistics, we study econometrics per se, which is at the crossroads between applied mathematics (statistics) and also social sciences (economics, etc.) $\longrightarrow$ THE KEY (AND THE DIFFICULTY, BUT ALSO THE INTEREST): *make links between a real-life situation (economic or in other fields – ordinary language) and the (mathematical – formal language) properties of random variables* (Chapter 0, slide 4 of the French version).

2.1 Formalization of the notion of causality through potential outcome variables

The beginning of Chapter 4 formalizes and defines the notion of causality, more precisely of *the causal effect* of a treatment D on an outcome or explained variable Y.

That definition of causal effects is done jointly with the introduction of *potential outcome variables (or simply, potential outcomes)*: the $Y(d)$ with d a free variable, a possible value, that is a real number (in the case, for now, of a univariate treatment), of the real random variable D: $d \in \text{Support}(D) \subset \mathbb{R}$.

Those potential outcomes are *counterfactual*. They represent, for each individual (the $Y(d)$ are generic instances of the $Y(d)_i$ for each individual / unit / observation i), what would have been the value of the outcome variable for an individual if the latter would have got realization / value d for the treatment variable D (in a way, in a parallel universe if it eases your understanding)[4].

$Y(d)_i$ is thus what *would have got* unit i for its outcome variable had it got $D_i = d$ in terms treatment variable.

The individual causal effect specific to an individual and his or her potential outcomes are *jointly defined. We need to accept that it is the definition of causality in this course (and in current econometrics).*

The case of a binary treatment When D is binary, meaning that D is a real random variable with a Bernoulli distribution, taking value 0 (untreated) or 1 (treated), the individual causal effect of D on Y is defined as

$$\Delta := Y(1) - Y(0) = \text{the difference between potential outcomes.}$$

We only observe *the realized or observed outcome variable* defined as

$$Y := Y(D) = DY(1) + (1 - D)Y(0) = Y(0) + D[Y(1) - Y(0)] = Y(0) + D\Delta,$$

so that $Y = Y(1)$ if $D = 1$ and $Y = Y(0)$ otherwise if $D = 0$.

For a given individual, it is crucial to understand that we can thus observe only one of its potential outcomes; the other remains counterfactual. Precision about the meaning of *counterfactual*: it is a more radical problem than “missing data” or “unobserved” variables; the point is not that the data collection

[4] A justification or rather a philosophical explanation behind the construction (mainly to improve your understanding of potential variables): “1.2 The world divides into facts. 1.21 Any one [of those facts] can either be the case or not be the case, and everything else remain the same.” *Tractatus logico-philosophicus*, Wittgenstein, propositions 1.2 and 1.21.

has been subject to some mistakes or failures and that we could have observed the counterfactual variable (but missed it in some way), no, rather that variable will never be available.

In other words, once the random variable D is realized (an ω is drawn) for a unit, one and exactly one of its potential outcomes is realized and is observed: $Y := Y(D)$, the other (when D is binary) or all the others (in the more general case of a non-binary D) are counterfactual.

Behind an interest in causality, there is often an interest in public policy (or treatment) evaluation: should we implement or not a given treatment? If Y represents the utility of an individual (net of the cost, expressed in terms of utility, of a treatment), an economist would like to treat that individual, assign him the treatment (that is, formally, having $D = 1$) if and only if Δ is positive. Ideally, we would thus like to know the individual causal effects Δ_i for each individual i.

Without further assumption, the previous paragraph says it is impossible: there is no hope of retrieving the individual Δ. That is why we restrict to *aggregated* causal effects, which formally are functions of the probability distribution P_Δ of Δ (among other dependencies according to the parameter we are interested in). Indeed, as Δ is specific to each individual and unobserved, it is modeled as random, stochastic: mathematically, Δ is a real random variable (*heterogeneous* causal effects).

In this course, we focus on *average* causal effects (possibly with weights, weighted averages, and also possibly an average over a sub-population only).

In the case of a binary treatment D, the two main parameters of interest are:

$\delta := \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\Delta]$, the average causal effect (over the entire population of interest),

$\delta^{\mathrm{T}} := \mathbb{E}[Y(1) - Y(0) \,|\, D = 1] = \mathbb{E}[\Delta \,|\, D = 1]$, the average causal effect over treated units.

The case of a univariate treatment, not necessarily binary When D is not necessarily binary but keeps a quantitative meaning and an order, we assume a *linear* causal effect of D on Y (or linear in known transforms of D and of Y, see the log-log, log-level, etc. models; besides, it is also possible to include interactions or powers to relax that linearity assumption): it exists $d_0 \in \mathrm{Support}(D)$ and a unique real random variable $\Delta \in \mathbb{R}^\Omega$ such that, for all $d \in \mathrm{Support}(D)$,

$$Y(d) - Y(d_0) = \Delta(d - d_0) \qquad \text{(Chapter 4, Equation (Lin. effects))}. \tag{L}$$

Remark: if this holds for some d_0, it is also the case for any $d_0 \in \mathrm{Support}(D)$; here, d_0 only plays the role of a reference value to express that the individual causal effect of D on Y is linear.

That linearity assumption was not required in the case of a binary D because, then, D taking only two values, linearity was somehow mechanically satisfied: for a given individual, there is a unique causal effect Δ, the effect of going from untreated ($D = 0$) to treated ($D = 1$).

It is not the case anymore when D is not binary. The assumption (L) ("L" standing for Linear) of a linear causal effect authorizes to get back to a *unique causal effect of D on Y* per individual: what is the cause on the outcome variable Y of an increase by 1 of the treatment D.

When D is a continuous variable[5], Δ can be interpreted as a *marginal causal effect*[6] insofar as, for all $d \in \mathrm{Support}(D)$, $\dfrac{\partial Y(d)}{\partial d} = \Delta$.

Remark: in (L), Δ remains random or stochastic, specific to each individual.

2.2 Non-causal linear representation (theoretical linear projection) and causal representation: do they coincide? That is, is there a selection bias?

The result to keep in mind if there was only one Morally (in a nutshell, as the morale of a short story or mathematical theorems and propositions here!), the **fundamental results of the beginning of Chapter 4**,

[5] That is when the real random variable D admits a density with respect to Lebesgue's measure.

[6] Remark: not to be confused with the notion of (non-causal) marginal effects, which can be defined without any reference to causality (see Chapter 1, slides 17, 18, and 36, as well as Quiz 1, Question 12).

- **Chapter 4, Proposition 1** for the case of binary D,
- **Chapter 4, Proposition 2** for the case of non-binary D,

assert that *we can identify and consistently estimate through OLS (the probability limit of the slope OLS estimator is equal to the average causal effect of interest) a causal average parameter (with possibly a weighted average or over a sub-population only) by a simple linear regression of Y on D* IF *there is no selection bias, namely if the potential outcomes $Y(d)$ are uncorrelated with the treatment variable D.* In such a case, D is also said to be *exogenous* with respect to the causal model written with potential outcomes.

Therefore, it is NOT MECHANICAL: "correlation is not causality" as often said, "No causation without manipulation" as sometimes said.

AN ASSUMPTION IS REQUIRED with a concrete meaning (not only a mathematical meaning; this is where the bridge between econometrics and social sciences is) for the linear regression of Y on D to consistently estimate some average causal effect of D on Y: the realized, actual treatment D ought to be generated, realized so that it is uncorrelated with the potential outcomes $Y(d)$, $d \in \text{Support}(D)$.

To get that point and understand why it makes sense to wonder whether there is no correlation – or, more generally, independence, but here, heuristically, for *linear* regressions, the absence of (*linear*) correlation is enough – between D and $Y(d)$, for $d \in \text{Support}(D)$, we need to realize that conceptually, there is no link between the realized treatment D and the potential outcome variables $Y(d)$: they are two distinct objects. The observed outcome will be $Y := Y(D)$, but, at the level of potential outcomes, D and $Y(d)$ are two different things and we wonder whether the treatment D is determined in a way that depends or not on the potential outcomes $Y(d)$.

As a general rule, the determination of D depends on the potential outcomes: there is a selection bias, and the probability limit of the OLS estimator, the theoretical coefficient β_D, mixes a causal effect and a selection effect: the treated individuals with $D = 1$ and the untreated individuals with $D = 0$ are not the same (in terms of potential outcome variables).

See also Simpson's paradox (Quiz 4, Question 2).

Some details about the mathematical formalization of that idea

The case of a binary treatment D.

Chapter 4, Proposition 1:

$$\underbrace{\beta_D}_{\widehat{\beta}_D \xrightarrow[n\to+\infty]{P} \beta_D} = \delta^{\mathrm{T}} \iff \mathbb{C}\mathrm{ov}(D, Y(0)) = 0 \iff \underbrace{\mathbb{E}[Y(0) \mid D = 1] - \mathbb{E}[Y(0) \mid D = 0]}_{=: B, \text{ the selection bias}} = 0.$$

Chapter 4, slides 12 et 22:

$$\underbrace{D \perp\!\!\!\perp (Y(0), Y(1))}_{D \text{ randomly drawn (experience)}} \implies \forall\, d \in \underbrace{\{0,1\}}_{\text{Support}(D)}, \mathbb{C}\mathrm{ov}(D, Y(d)) = 0 \implies \underbrace{\beta_D}_{\widehat{\beta}_D \xrightarrow[n\to+\infty]{P} \beta_D} = \delta = \delta^{\mathrm{T}} = \delta^W.$$

The case of a non-binary treatment D. Chapter 4, Proposition 2:

$$\overbrace{\text{(L)}}^{\text{linear causal effects}} \quad \text{et} \quad \overbrace{\forall d \in \text{Support}(D),\ \mathbb{C}\mathrm{ov}(D, Y(d)) = 0}^{\text{absence of selection bias}}$$

$$\implies \underbrace{\beta_D}_{\widehat{\beta}_D \xrightarrow[n\to+\infty]{P} \beta_D} = \delta^W := \mathbb{E}[W\Delta], \text{ où } W := \frac{(D - \mathbb{E}[D])^2}{\mathbb{V}[D]}.$$

The latest result can be seen positively: under those two conditions, a simple linear regression of Y on D does estimate a causal parameter, δ^W, a *weighted* average causal effects with weights larger as an individual's treatment D is further (the square of the Euclidean distance) from the average treatment $\mathbb{E}[D]$. Yet, it can also be seen more negatively: even in the absence of selection bias and with linear causal effects, without any restriction on the heterogeneity of individual causal effects (Δ is stochastic and we do not impose any restriction on the probability distribution P_Δ), then at best, a linear regression can identify an average causal effect, but *weighted*, with weights W that have the advantage of being non-negative but, however, do not have a clear interpretation as regards the decision to implement or not the treatment under study.

Average causal effects and decisions of implementing a treatment Some preliminary remarks about the previous point.[7] To simplify the exposition, we neglect here *statistical uncertainty*: in fact, at best, we have the realization of a consistent estimator of a causal parameter, and we can perform inference (tests, confidence intervals) on that causal parameter, but we never know it with certainty.. We also neglect the cost of the treatment or, instead, we include it in Y, which represents the utility of an individual net of the treatment cost.

Imagine an egalitarian (she attributes the same weight to each individual of a given population of interest), benevolent (those weights are non-negative), and utilitarian (she cares only about the total, or, equivalently, the average of the utilities Y) social planner.

She wonders whether to implement on the population of interest (thereof she gets and observes a representative sample) and, if so, how, some public policy, the treatment D.

The “treatment” can be binary (for example, a training program such as JTPA studied in Problem Set 9); in this case, the social planner wonders whether to treat or not individuals, namely, whether they follow the treatment or program. More generally, the treatment may be non-binary (for example, the number of years of education); in this case, the social planner wonders whether the treatment should be increased by one (or several) units (to resume the example: implement an additional year of education by increasing by one year the legal mandatory age of studies) for the individuals of the population under study.

- $\delta > 0$ (on average in the population, the individual causal effect Δ is positive) is a necessary and sufficient condition (NSC) for the social planner to decide a strict implementation of the treatment, meaning to really force all individuals to follow the treatment in the binary case: $D = 1$, or, more generally, $D \to D + 1$ (increase by one unit, for instance, the intensity of the non-binary treatment) since, on average, and thus also for the total in the population, that increase has a positive effect.

- $\Delta > 0$ almost surely (a.s) (the individual causal effect Δ is positive for each individual) also implies that strict implementation of the treatment; even more, if so, it is the case whatever the (non-negative as the social planner is benevolent, she does not want to harm anyone) weights that are used in the planner's criterion that aggregates individual utilities.

- $\delta^{\mathrm{T}} > 0$ (on average among treated individuals, the individual causal effect Δ is positive) is an NSC for the social planner to decide free access to the treatment, meaning to allow anyone who wants to do so to follow it (in the case of a binary treatment since we consider δ^{T}), but without imposing the treatment, since this will have a positive effect on average among treated individuals[8].

- In contrast, $\delta^W > 0$ is *not* a piece of sufficient information by itself for the social planner to decide whether to implement or not the treatment. It is possible to have $\delta^W > 0$, but $\delta < 0$ and

[7] More generally, see the field of statistical decision theory, taking into account statistical uncertainty, as well as the issues regarding the “scalability” of a treatment, a public policy; many questions arise that we neglect here but are crucial in practice: as a start, you can think of the notions of *iinternal validity* as opposed to *external validity* and *general equilibrium effects* or *externalities* (that you already know by your economics training).

[8] In perfect complete information (no uncertainty, that would require introducing risk aversion, etc.), if each agent knows his or her Δ without doubt, and provided they behave to maximize their utility, the individuals who decide to follow the treatment are precisely those with $\Delta > 0$ – (see Problem Set 6, model 2 for such a model where individuals auto-select in the treatment; on the contrary, model 1 of Problem Set 6 is the case of a randomized controlled trial).

$\delta^{\mathrm{T}} < 0$; $\Delta > 0$ a.s implies $\delta^W > 0$, but the converse is obviously false. The condition $\delta^W > 0$ would be an NSC for a strict implementation of the treatment if and only if the utilitarian social planner used the weights W (involved in the weighted average causal effect δ^W) as her weights in the aggregation of individual utilities. However, there is no reason a priori to rationalize such a weighting scheme.[9]

Heterogeneous or homogeneous causal effects (Individual) causal effects are said to be *heterogeneous* when Δ is "random", that is, formally, is not a constant (degenerate) real random variable: $\mathbb{V}[\Delta] > 0$.

Otherwise, if there exists a non-stochastic real number δ_0 such that $\Delta = \delta_0$ a.s; equivalently, if P_Δ is a Dirac mass on δ_0, causal effects are said to be *homogeneous*.

In most practical applications, it is a strong assumption, often unrealistic. Its key interest is returning to a *unique* causal parameter. Consequently, in the binary treatment case (in which δ^{T} is naturally relevant) or in the general non-binary case,

$$\exists\, \delta_0 \in \mathbb{R} : \Delta = \delta_0 \,\text{a.s} \implies \delta = \delta^{\mathrm{T}} = \delta^W = \text{ any weighted or over a sub-population average of } \Delta$$

because the individual causal effects are all equal to δ_0.

The assumption $\mathbb{E}[\Delta \mid D, G] = \delta_0$ in models (Lin. mod. 1) and (Lin. mod. 1) of Chapter 4 authorizes heterogeneous effects, but restricts the heterogeneity: causal effects cannot depend on D nor G. That form of causal effect heterogeneity can be called idiosyncratic since causal effects can be random, specific to each individual, but without any systematic differences due to the realization of their treatment D or their control variables G (see Quiz 4, Questions 10 et 11 for further details about that hypothesis and, more generally, about linear models 1 and 2 of Chapter 4).

In the perspective mentioned above of the social planner, that restriction is crucial: essentially, it allows to boil down the analysis to a *single average* causal effect: δ_0; in particular, $\mathbb{E}[\Delta] =: \delta = \delta_0$.

Including interactions in the model can partially relax the assumption (see Chapter 4, slide 29).

The particular case (and a conceptual benchmark) of randomized experiments There exists a particular case where there is indeed absence of selection / exogeneity of D: when the effective realized treatment D is "random" *in the ordinary language meaning of the term "random"* (note that D is always "random" in the mathematical formal language sense of being stochastic since D is a random variable), namely drawn, determined randomly. This occurs in Randomized Controlled Trials (RCT) in which the effective treatment D is properly "randomized".

On the contrary, in "natural experiments", administrative data, surveys, or observations obtained in real life without all the logistics and manipulation involved in randomized experiments, most of the time, and in particular as soon as individuals have some leeway to determine D, they can, in a way or another (even if only partially), choose their realized value for D, it is likely that D and $\{Y(d)\}_{d \in \text{Support}(D)}$ are correlated, namely, that D is *endogenous*. In this situation, there is a selection bias: the simple linear regression of Y on D does *not* consistently estimate a causal parameter (which, here, is defined as some average of individual causal effects).

In practice, YOU CAN AND SHOULD ASK YOU THE FOLLOWING QUESTION: *to what extent the way the effective treatment variable D is determined in the observed data is close or far from the situation of a randomized controlled trial in which the treatment D would be randomly allocated, drawn at random, haphazardly?*

Again (but this is crucial!), in general, as soon as individuals have some leeway over D, we get further from that situation of randomized controlled trials, in which D is imposed on individuals on the contrary. Outside of random experiments with perfect compliance, that is, when the effective

[9] As explained before, an egalitarian social planner considers uniform weights across individuals. Other political philosophies might prefer other weights; yet, a priori, those weights have no reason to coincide with the weights W, which only depend on the distance with the average treatment intensity.

treatment D is equal to the random assignment Z (using Chapter 5's notation), *the assumption of the absence of selection is often not plausible*.

From there, **the third part of the course** can be understood as a presentation of two ways to try to overcome that problem and estimate a causal parameter despite selection bias.

3 Third part – two identification strategies to recover an average causal effect despite (unconditional) selection bias

3.1 Add the adequate control variables – Chapter 4, Sections 3 and 4

Identification and consistent estimation of an average causal effect by a multiple regression Although $\mathbb{C}\text{ov}(D, Y(d)) = 0$ for all $d \in \text{Support}(D)$ is often unrealistic, provided adequate controls variables G, *the absence of conditional selection*: $\forall\, d \in \text{Support}(D), \mathbb{C}\text{ov}(D, Y(d) \,|\, G) = 0$ can be more credible.

If so, under the absence of conditional selection, still with the assumption of *linear* causal effects and provided *restricted heterogeneity* of individual causal effects Δ by assuming $\mathbb{E}[\Delta \,|\, D, G] = \delta_0$, a non-stochastic real number, it is possible to consistently estimate the causal parameter of interest which is the average causal effect $\delta := \mathbb{E}[\Delta] = \delta_0$ by a *multiple* linear regression of Y on D and G (**Chapter 4, (Lin. mod. 1) and (Lin. mod. 2), Proposition 4**).

Omitted variable bias On the contrary, if we omit among the control variables a variable

(a) that affects, influences, mathematically that is correlated with the explained variable Y,

(b) AND (*both conditions are required*) that is correlated with the treatment D,

then there is an omitted variable bias (Chapter 4, Proposition 5).

The intuition is thus to put as controls G any variable that can be correlated simultaneously with Y and with D to prevent such an omitted variable bias and obtain the absence of conditional selection. In other words, consider variables G such that *when we condition by those variables, the realization of D can be considered "as-if" it were drawn randomly, picked at random* (behind, heuristically, you can and should always think of the conceptual reference setting of a randomized controlled experiment).

Remark: be careful that reasoning does not mean we should put as many control variables as possible, any variable in G (see Chapter 4, slides 37 and 38: included variable bias).

To obtain such adequate control variables G, first conceive and find them, have the idea of the relevant controls, and, second, be able to measure them in the data, is often tricky. Hence, another identification approach is through instrumental variables.

3.2 Rely on a valid instrument – Chapter 5 : instrumental variables

Intuition Morally, an instrumental variable Z is a variable:

- that affects (formally, that is correlated with) the treatment variable D, even net of possible controls G (*relevance condition*), and
- that is exogenous (*exogeneity condition*), in the sense of uncorrelated with all unobserved determinants, summed up in the error term η of the course (that intervenes in the causal representations with potential outcomes), which affect the potential outcomes $Y(d)$ (and thus also the observed outcome Y):

  (i) it is credible to consider that the instrument Z is determined "as-if" it were drawn randomly (possibly conditional on controls G), and

  (ii) Z affects Y only through D; there is no effect, influence by itself of Z on Y (exclusion restriction).

Intuition of the identification results of Chapter 5. A variation of Z induces an exogenous variation of D, which, in turn, can affect and make Y vary. Thus, we obtain an exogenous variation in the sense that we can conceive that variation as if drawn randomly (again and always the conceptual reference of a randomized controlled trial). Intuitively, such a variation allows to recover an average causal effect of D on Y *over the sub-population of the individuals that are affected by the instrument*, those whose treatment D exogenously varies as a response to a variation of the instrument Z.

Section 1: randomized experiments with imperfect compliance (Z et D binary, heterogeneous causal effects, no controls G) In this setting, the exogeneity of instruments (more precisely, part (i) of that condition) is guaranteed by their random draw, and

- Z = the indicator variable of being initially allocated to the treatment

  a priori *differs from*

- D = the indicator to actually follow the treatment (that can depend on individual choices: register, effectively follow a training, for instance, etc.).

Because D is (at least partially) determined by agents, D is a priori endogenous, and there is a selection bias. However, the initial assignment Z is exogenous, for drawn randomly (independence assumption, Equation (Indep.) of Chapter 5). Furthermore, in this setting, Z is indeed (positively) correlated with D (assumption $\mathbb{E}[D \mid Z = 1] > \mathbb{E}[D \mid Z = 0]$, Chapter 5, Theorem 1).

Under those hypotheses, assuming in addition monotonicity (Equation (Monoton.) of Chapter 5), then we can identify by a two-stage regression, Two-Stage Least Squares (TSLS), the causal parameter δ^{C}: the average of individual causal effects Δ *over the sub-population of compliers*, the individuals who comply to the instrument (**Chapter 5, Theorem 1**).

For now, there is no restriction on the heterogeneity of the individual causal effects Δ. As a consequence, we can identify an average causal effect *only* over the sub-population of the population of interest; hence the term of LATE, "Local Average Treatment Effect", for the causal parameter δ^{C}.

Remark: in the setting of Chapter 4 with a binary D, δ^{T} could also be called a *local* effect. Indeed, it is an average over a specific sub-population: the treated individuals (who, a priori, are different from the untreated individuals regarding potential outcomes; this is precisely the selection bias). However, that terminology is generally reserved to δ^{C}.

Sections 2 and more: generalization and "natural" experiments (Z and D non-binary, possibly multivariate, possible controls G, homogeneous causal effects) However, starting from Section 2 of Chapter 5, we assume *homogeneous* (and *linear* as usual) causal effects:

$$Y(d) = \zeta_0 + \delta_0 d + \eta, \qquad \underbrace{\mathbb{E}[\eta] = 0}_{\text{without loss of generality since there is a constant } \zeta_0} \qquad \text{(Chapter 5, Lin. model 1)}$$

$$Y(d) = \zeta_0 + G'\gamma_0 + \delta_0 d + \eta, \quad \mathbb{E}[\eta] \;=\; \underbrace{\mathbb{E}[G\eta] \;=\; 0}_{\text{that is, } G \textit{ exogenous}} \qquad \text{(Chapter 5, Lin. model 2)}$$

but we do *not* assume $\mathbb{E}[D\eta] = 0$:the treatment D can be *endogenous*.

The assumption of homogeneous causal effects implies: $\delta = \delta^{\mathrm{T}} = \delta^{W} = \delta^{\mathrm{C}} = \delta_0$.

In Model 1 (unconditionally), or in Model 2 (conditionally on controls G), it is worth noting that the only stochastic part in $Y(d)$, with $d \in \text{Support}(D)$, is the error term η, which aggregates the *individual unobserved heterogeneity* that affects the potential outcomes $Y(d)$ (and thus also affects the observed outcome Y).

Consequently, for any random variable A, the (conditional on G) covariance between A and $Y(d)$ is equal to the covariance between A and η, also equal to $\mathbb{E}[A\eta]$ (resp. $\mathbb{E}[A\eta \mid G]$) since η is centered.

In particular, for instance, unconditionally, under the assumption of Lin. model 1, we obtain:

$$\mathbb{C}\mathrm{ov}(D, Y(d)) = \mathbb{C}\mathrm{ov}(D, \eta) = \mathbb{E}[D\eta] \qquad \text{and} \qquad \mathbb{C}\mathrm{ov}(Z, Y(d)) = \mathbb{C}\mathrm{ov}(Z, \eta) = \mathbb{E}[Z\eta].$$

In the generalized setup of Sections 2 and following, if the causal effects are *linear* and *homogeneous* and if the *instrument* Z of D is *valid*, then we can identify and consistently estimate the homogeneous causal effect δ_0 by two-stage least squares:

- **Chapter 5, Theorem 4** for univariate D (a single endogenous regressor) – identification,
- **Chapter 5, Theorem 5** for multivariate D (several endogenous regressors; in that case, at least as many instruments as endogenous variables are required) – identification,
- **Chapter 5, Theorem 6** – estimation and inference: consistency and asymptotic normality of the TSLS estimator $\widehat{\beta}_{2\text{SLS}}$.

A *valid* instrument means that the instrument satisfies the two conditions: exogeneity and relevance. The exogeneity of the instrument cannot be tested, whereas we can and should test the relevance condition:

- Chapter 5, Proposition 1 for univariate D,
- Chapter 5, Proposition 2 for multivariate D,
- Chapter 5, slide 39 for the practical implementation of the test.

3.3 (Anticipated) Development: they exist other identification strategies – Econometrics 2 in Spring semester and third year econometric courses

If we are not in the setting of a randomized experiment (where the initial allocation Z into the treatment is randomly drawn), similarly as at tend of subsection 3.1 of this document (adding adequate control variables), the exogeneity of Z (even conditional, controlling by G) is often doubtful: *finding a valid instrument can be tough*.

There exist other ways to proceed, other identification strategies of causal effects. We will see some of them together in Econometrics 2 next semester.

In particular, a strategy is to use richer data in the sense of data that enables to follow the same individual over time (*panel data*) or sample from the same population of interest at several dates (*repeated cross sections*). See Chapter 0: Introduction, Section 3: “Different types of data”, for the definitions.

By-the-way remark: in the Econometrics 1 course, we are always in the setting of *cross-sectional data*. Furthermore, we always assume i.i.d sampling, and to lighten notation, we often omit the i-th indices to denote a generic variable or vector with the same probability distribution as the observations.

There are two main interrogations regarding cross-sectional data:

1. Is the sample “representative” of the population of interest we consider?
2. Is the assumption of independence between the vectors of random variables associated with the different units credible?

The Econometrics 2 course will address the second point (with the notion of “clustering”) and also, partially, point 1 (with selection models). Point 1 is crucial in practice, even if the Econometrics 1 course focuses on other aspects that are also important (see Quiz 1, Question 13 and Problem Set 7, Questions 7 to 10, about that issue of the representativeness of a sample).

4 Complements and conclusion

4.1 To prepare specifically for the theoretical exercises

It is useful to know and know how to use in computations (non-exhaustive list, only some fundamental points):

- the definition of theoretical linear regressions,
- the law of iterated expectations and its generalization (the composition of projections),
- the definition of the observed variables $Y := Y(D)$ (observed outcome variable) or $D := D(Z)$ (observed treatment variable; see Chapter 5, Section 1),
- the notion of conditional expectation and, more generally, the use of conditioning.

Example Below is an example for the last two points (the previous formulation is likely too abstract).

In the setting of a binary treatment D (Chapter 4, Section 1), the selection bias is defined as

$$B := \mathbb{E}[Y(0) \,|\, D = 1] - \mathbb{E}[Y(0) \,|\, D = 0].$$

Question. In the previous subtraction, one of the two expectations is counterfactual (it cannot be identified), while the other is identified in the data; which one?

Answer. We have

$$\begin{aligned}\mathbb{E}[Y \,|\, D = 0] &= \mathbb{E}[Y(D) \,|\, D = 0] \qquad \text{(by definition of } Y := Y(D)\text{)}\\ &= \mathbb{E}[Y(0) \,|\, D = 0] \qquad \text{(using the conditioning).}\end{aligned}$$

Thus $\mathbb{E}[Y(0) \,|\, D = 0] = \mathbb{E}[Y \,|\, D = 0]$ is a function of the probability distribution $\mathrm{P}_{(Y,D)}$ of the observations, and is therefore a quantity identified in the data.

Reminder: when discussing identification, we ask ourselves what can be learned *reasoning as if* we knew the distribution of the data, that is as if we had an “infinite sample”.

What is more (and better), in addition to being identified, we can consistently estimate that expectation by the empirical mean of the Y_i on the sub-sample of observations such that $D_i = 0$, provided i.i.d data (the required moment conditions to apply the law of large numbers here are implicitly assumed to hold, as presented in the course slides sometimes and as often presented in Problem Sets or exams):

$$\frac{\sum_{i=1}^n Y_i \mathbb{1}\{D_i = 0\}}{\sum_{i=1}^n \mathbb{1}\{D_i = 0\}} = \frac{\sum_{i \in \{1,\dots,n\} : D_i = 0} Y_i}{\mathrm{Card}(\{i \in \{1, \dots, n\} : D_i = 0\})} \xrightarrow[n \to +\infty]{P} \mathbb{E}[Y \,|\, D = 0] = \mathbb{E}[Y(0) \,|\, D = 0].$$

It is only an instance of the type of reasoning often used in the proofs of the main theorems of Chapters 4 and 5. Knowing how to do this kind of reasoning is helpful, especially for theoretical exercises in mid-term and final exams.

4.2 An example

Since a good example is often better than long speeches... below is an example (based on simulated data)[10] to illustrate the crux of the Econometrics 1 course (see also the examples in the course slides and Problem Sets, as well as the examples of Quiz 4, Question 3).

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.

[10] If you are interested, please contact me to obtain the texttR script used to generate the example.

Figure 1: 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):

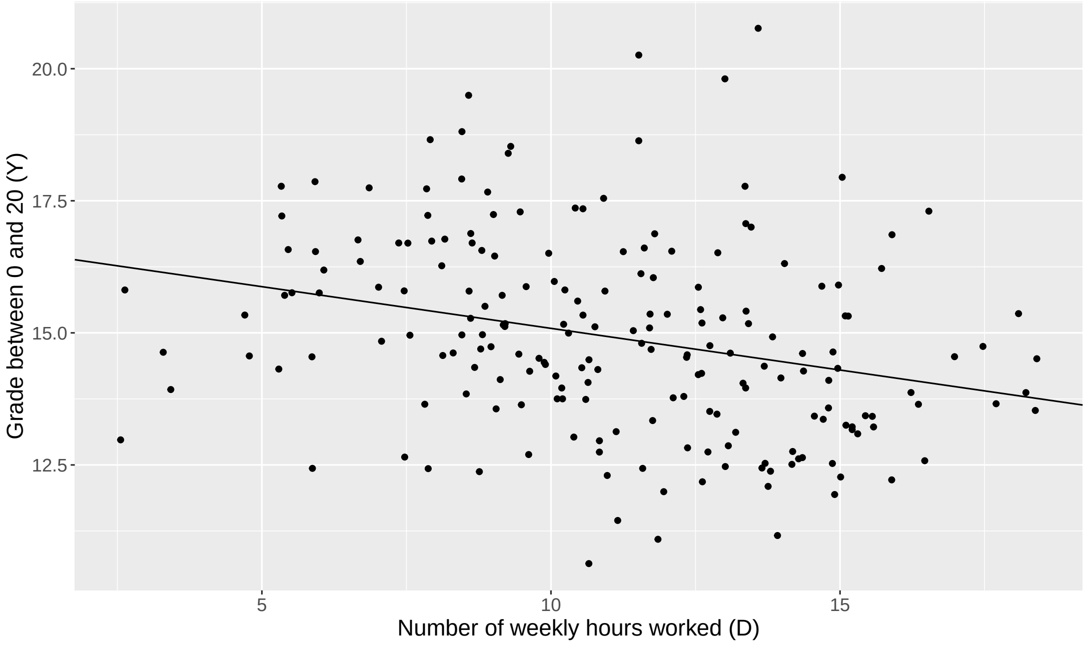


Figure 2: But maybe students do not randomly choose how much time they spend studying a subject! There is no randomized experiment on the amount of time spent per subject; in other words, the agents (in this case, the students) choose the "treatment" themselves. We can probably even think (a bold hypothesis) 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:

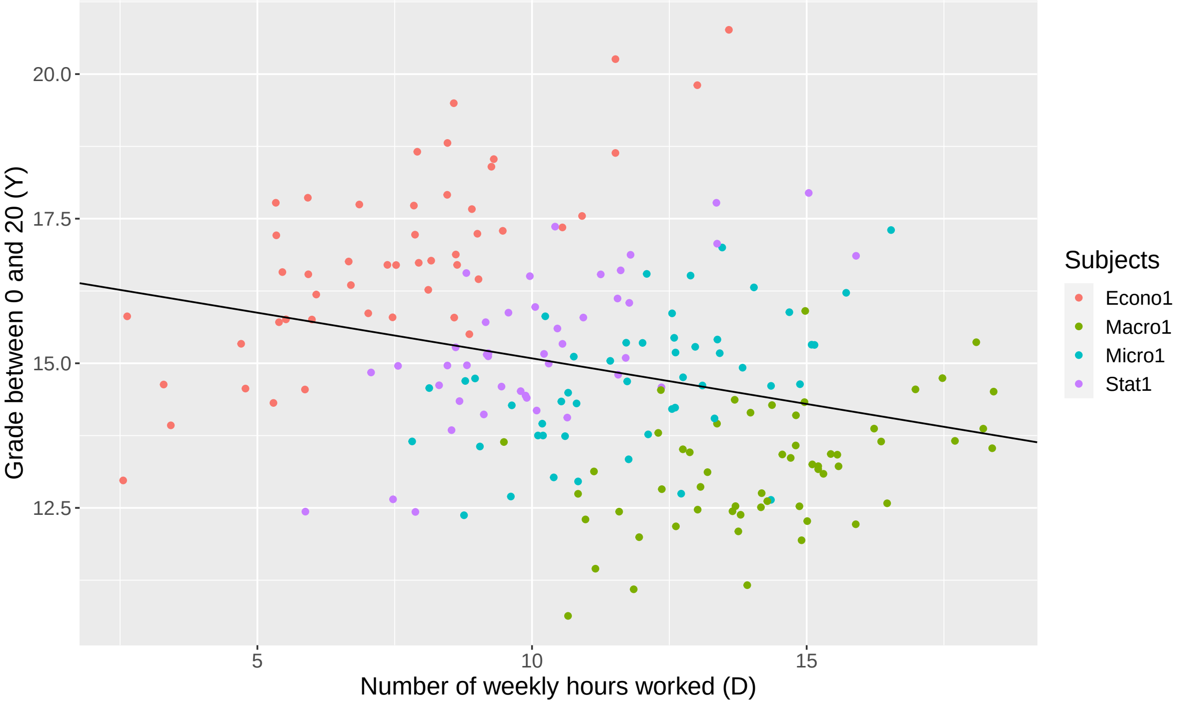
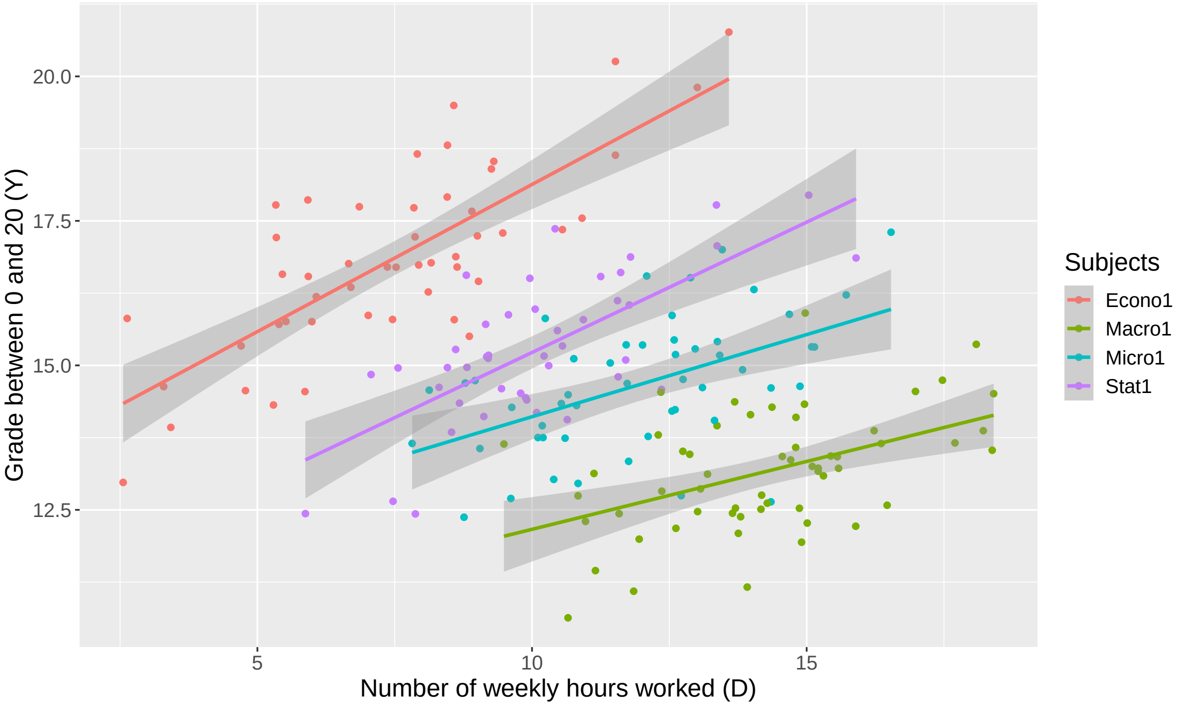


Figure 3: The previous simple linear regression suffers from a selection bias: it does not identify an average causal effect of D on Y. Indeed, we have an omitted variable bias if we do not control for the difficulty of each subject, which is correlated *both* with the treatment variable (the number 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), which corresponds more closely to Proposition 3 of Chapter 4 (slide 27):



4.3 To sum up the summary (in 4 images and about 160 words)

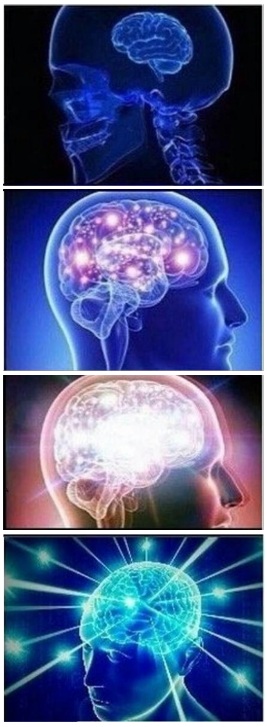


Figure 4: A "sum-meme-ry" of the course[11]:

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).

[11] In my current understanding, I have the impression that there is a fifth stage, partially mentioned in that summary, by the way, but outside the course's material, and there may exist a sixth and even higher stages: good thinking!