Quiz 4 – Chapter 4: Linear Regressions and Causality

(Lucas Girard) – This version: 11 January 2025.
Compared to 5 December 2024: references to new summary documents that supersede Question 11

Solutions

The quizzes are provided as training to help you check your knowledge and understanding of the course; the course and the TD remain the only reference. The quizzes are not necessary, all the less so sufficient, to study Econometrics 1 but might nonetheless be helpful in your learning[1].

Some words about the quiz. Notation. *Beyond notations, try to be constantly aware of the nature of the objects they denote*: is it a non-stochastic parameter like β_0? Or an estimator, thus a random variable (since it is a function of the stochastic observations), like $\widehat{\beta}$? Likewise, be careful about the dimension of the objects (vectors, matrices, numbers) in computations. Absent contrary indication, the notation used follows that of the course's slides.
$D \in \mathbb{R}^{\Omega}$, a real random variable, is the treatment variable, either binary (Support$(D) = \{0, 1\}$) or a non-binary *ordered quantitative* variable (Support$(D) = \mathbb{R}$). (Sometimes, we may also consider multivariate treatment: $D \in (\mathbb{R})^{\dim(D)}$, with $\dim(D) \geq 2$.) We are interested in the causal effect of D (the "treatment") on an outcome real variable Y. For any $d \in$ Support(D) (note that d is thus not a random variable, it is just a "free variable" – *variable muette* in French), we denote by $Y(d)$ the *potential* outcome associated with the value d of the treatment D: $Y(d) \in \mathbb{R}^{\Omega}$ is a real random variable. Remember that, in addition to D (and G if present), we only observe $Y := Y(D) \in \mathbb{R}^{\Omega}$, the *observed* outcome. In particular, when the treatment D is binary, we observe $Y := DY(1) + (1 - D)Y(0)$, but not the couple of potential outcomes $(Y(0), Y(1))$. That distinction between *observed* and *potential* outcomes is crucial.
We thus assume to observe an i.i.d. sample $(Y_i, D_i)_{i=1,\dots,n}$, with $n \in \mathbb{N}^*$, that have the same distribution, denoted $\mathrm{P}_{(Y,D)}$, as a generic couple (Y, D), written without the index i. Sometimes, we also have access to a column vector $G \in (\mathbb{R}^{\dim(G)})^{\Omega}$ of control variables. In this case, we observe $(Y_i, D_i, G_i')_{i=1,\dots,n}$ i.i.d with distribution $\mathrm{P}_{(Y,D,G)}$.
We consider simple linear regressions of Y on D, or multiple linear regressions of Y on D and G. Absent contrary indications, all regressions include an intercept/constant as usual. As in problem sets and exams, if not stated otherwise, we implicitly assume i.i.d. sampling of the observations and that the three standard moment conditions of Chapter 1, Proposition 5 hold: Y and X admit finite second-order moments, with $X := (1, D)'$ (without controls) or $X := (1, D, G')'$ (with controls), and $\mathbb{E}[XX']$ is invertible.
Questions. As in the course slides, questions marked with an asterisk are more advanced.
Questions 1 to 4 deal with the first section of Chapter 4: binary treatments. Question 1 is to check your knowledge of the fundamental definitions used to formalize causal effects. Question 2 is about Proposition 1. Question 3 presents two concrete examples to apply the notions of Chapter 4; they are interesting exercises. Part (a) of Question 4 is a must-know about Proposition 1, again. Part (b) is more advanced and deals with the issue of testing the absence of selection.
Questions 5 and 6 are concerned with the interpretation of linear regressions and how, despite being linear (*in parameters!*), linear regression can account for non-linear effects, at the cost of using known transformations of the initial outcome and/or the treatment variables.
Question 7 is about Section 2: the case of a single non-binary (but ordered and with a quantitative – as opposed to qualitative – meaning) treatment, notably Proposition 2.
Questions 8 to 9 are concerned with the second part of Chapter 4, sections 3 and 4 with control variables. More precisely, Question 8 is also linked to sections 1 and 2 and compares the assumptions of absence of selection and of absence of *conditional* selection, which are critical conditions for the purpose of Chapter 4, namely, identify causal parameters through linear regressions. For that goal, Question 9 discusses, on an example, whether one should add control variables.
Questions 10 and 11 are proposed to explore several aspects of the causal linear models (Lin. mod. 1) and (Lin. mod. 2) studied in Chapter 4. Question 10 is quite important to understand the distinction between heterogeneous or homogeneous causal effects. Question 11 (marked with an asterisk) is a more open question, but it might be interesting to check your understanding of the course deeper.
About the solution to Question 11. It happens to be much longer (pages 28 to 43) than initially thought. It is a sort of general commentary on the linear models 1 and 2 of Chapter 4, with also some links with the model of Chapter 5. Note that this material is more advanced compared to the course and also not very well structured as of now. I would recommend reading first the structural summary ("`RStructuralSummary`", available on Pamplemousse in French or in English); then, if you are interested, you can read and think about Question 11.
Update (January 2025): new summary documents that present the main results of Chapter 4: causal effects and linear regressions are available on Pamplemousse – see `Quiz4_Chapter4_complements_..`. Question 11 and its solution are superseded by these documents (better structured and more synthetic); I would suggest the general structural summary, then the two syntheses for Chapter 4 (without and with control variables); then look at Question 11 if interested.

Bonne lecture ! Do not hesitate if you have any questions.

[1] See "auto-test", one of the pillars of efficient learning – reference: David Louapre (Science Étonnante)'s video on learning how to learn (*link*). If you have not seen this video yet, I advise you to stop this quiz immediately and first watch it: the returns you can get from this 29-minute video likely eclipse any specific quiz, lecture note, or review.

1 Fundamental objects for causality (binary treatment)

As in the first section of Chapter 4, we consider a single binary variable $D \in \{0,1\}^\Omega$.

(a) Chapter 4 introduces the following objects and notations: Δ, δ, δ^{T}, B, and β_D. Give their respective name and definition.

For a binary treatment D, Δ (in *upper case* $\neq \delta$) is defined as the difference between the two potential outcomes: $\Delta := Y(1) - Y(0)$ and is *the individual causal effect of the treatment D on the outcome variable Y* (slide 5, Chapter 4). In this econometrics course, for a given individual, the individual causal effect of a binary treatment variable D on an outcome variable Y is formalized as the difference between the potential outcome $Y(1)$ in the presence of the treatment minus the potential outcome absent the treatment $Y(0)$.

Remember that we assume i.i.d. variables in the Econometrics 1 course: the variables without subscript i are simply generic instances of the variables (with the same distribution since they are i.i.d.). Thus, the previous definition is equivalent to saying: for any individual i, $\Delta_i := Y(1)_i - Y(0)_i$ is the *individual* causal effect of D on Y *for individual* i.

A priori, there is no reason why the causal effects should be the same across individuals. In other words, the Δ_i vary across individuals i; for the econometrician, as an individual i is drawn randomly from a population of interest, it amounts *to considering Δ_i as a random variable*; in practice, it allows for the causal effects of a treatment to be different for distinct individuals.

For a given individual i, we only observe $Y_i := Y(D_i)$, never the couple $(Y(0)_i, Y(1)_i)$ of potential outcomes. Consequently, Δ_i is unobserved and, without additional assumptions, cannot be retrieved from the data; formally, it is not identified. This is why we will restrict to the identification and estimation of some features of the distribution of Δ (see second bullet-point of slide 5, Chapter 4).

A first feature, often considered, is δ (in *lower case* $\neq \Delta$). δ is defined as the expectation of Δ: $\delta := \mathbb{E}[\Delta]$: the expected individual causal effect, or *the average treatment effect* (note: *treatment effect* is used as a synonym of *causal effect* in those expressions) (slide 5, Chapter 4).

The key point to understand is the following: **absent specific data-collection schemes such as randomized controlled trials, there is no reason in general that D is determined independently of the potential outcomes $Y(0)$ and $Y(1)$.** In most settings, a priori, absent specific arguments in favor of the opposite, the individuals with $D = 1$ (treated) and the individuals with $D = 0$ (non-treated) are likely to be different in terms of potential outcomes $Y(0)$ and $Y(1)$.

In particular, the average individual causal effect among the treated likely differs from the average individual causal effect among the non-treated. A second feature often considered is thus the average individual causal effect among the treated, which is denoted δ^{T} (with the superscript "T" for "Treated"): $\delta^{\mathrm{T}} := \mathbb{E}[\Delta \,|\, D = 1]$ is *the average treatment/causal effect on the treated*.

A way to express that the treated and non-treated individuals differ in terms of potential outcomes is to consider the expectations of $Y(0)$, the potential outcome absent the treatment, in each sub-group (treated and non-treated) and see whether they differ across treated and non-treated.

Remark that, when Support$(D) = \{0,1\}$,

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

Consequently, looking at $Y(0)$ is like looking at a baseline value for the outcome absent treatment. This is somewhat the idea behind the definition (slide 8, Chapter 4) of *the selection bias B*

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

$B = 0$ means the expected potential outcome absent the treatment is the same for treated and non-treated individuals:

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

that is, since $D \in \{0, 1\}$, the conditional expectation of $Y(0)$ given D does not depend on D and is equal to the unconditional expectation $\mathbb{E}[Y(0)]$: $Y(0)$ *is said to be mean-independent of* D.

In particular, it implies that $Y(0)$ and D are uncorrelated. Remark that the converse is generally false but holds if D is binary (to be checked as an exercise).

β_D is the limit in probability (under i.i.d. sampling and adequate moment conditions as in Chapter 1, Proposition 5) of the OLS estimator of the slope in the simple linear regression of Y on D (and a constant as always) (Proposition 5, Chapter 1)

$$\widehat{\beta}_D := \frac{\widehat{\mathbb{C}\mathrm{ov}}(Y, D)}{\widehat{\mathbb{V}}[D]} \xrightarrow[n \to +\infty]{P} \beta_D := \frac{\mathbb{C}\mathrm{ov}(Y, D)}{\mathbb{V}[D]} \overset{\text{because } D \text{ binary}}{=} \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0].$$

Note that B is the difference of the conditional expectations (knowing D, $|D = 1$ minus $|D = 0$) of the potential outcome $Y(0)$ while β_D is the difference of the conditional expectations of the observed outcome Y.

Now, the main question is: under mild moment conditions and with i.i.d. observations, **we can identify/estimate β_D consistently, fine; but,*under which conditions is β_D equal to some average causal effects of interest*, such as δ or δ^{T}?** The answer to this fundamental question of the course is given, in the case of a simple linear regression with a binary treatment D, by **Proposition 1 of Chapter 4**.

(b) In general, what can we say about δ and δ^{T}?

In general, that is, without any other information as regards the context, that is, the way the data is collected (notably: whether it is a randomized controlled trial or not – "natural experiment"), we can only say $\delta \neq \delta^{\mathrm{T}}$ (see the last bullet-point of slide 5, Chapter 4).

1. $\delta > \delta^{\mathrm{T}}$ – **False**.
2. $\delta < \delta^{\mathrm{T}}$ – **False**.
3. $\delta \neq \delta^{\mathrm{T}}$ – **True**.
4. $\delta = \delta^{\mathrm{T}}$ – **False**.

(c) In this sub-question, we assume *homogeneous causal effects*, namely $\exists\, \delta_0 \in \mathbb{R} : \Delta = \delta_0$, that is, Δ is a degenerate/constant random variable ($\mathbb{V}[\Delta] = 0$). In this case, what can we say about δ and δ^{T}?

As explained in the solution to (a) above, *a priori, the individual causal effects vary across individuals:* ***Δ is stochastic****; this case is called the case of* ***heterogeneous causal effects***.

However, sometimes, we make the **assumption** of *homogeneous causal effects*: Δ is a degenerate constant random variable: all the individual causal effects are equal to a non-stochastic real number (here denoted δ_0, τ is another frequent notation). It means that the treatment has the same causal effect for all individuals. Remark that, in most situations, it will be a very *strong assumption.*

If $\Delta = \delta_0 \in \mathbb{R}$ non-stochastic, we obtain by linearity of unconditional and conditional expectations

$$\begin{aligned} \delta &:= \mathbb{E}[\Delta] = \mathbb{E}[\delta_0] = \delta_0 \times \mathbb{E}[1] = \delta_0, \\ \delta^{\mathrm{T}} &:= \mathbb{E}[\Delta \mid D = 1] = \mathbb{E}[\delta_0 \mid D = 1] = \delta_0 \times \mathbb{E}[1 \mid D = 1] = \delta_0, \end{aligned}$$

thus, when the treatment/causal effects are homogeneous, $\delta = \delta^{\mathrm{T}} = \delta_0$, which is logical: if the causal effect is the same for everyone and equal to δ_0, then the average effect over the whole population or the average effect over any sub-group of the population (the treated individuals for δ^{T}) is also equal to δ_0.

2 Causality and regressions with a single binary covariate

We consider a single binary treatment $D \in \{0,1\}^{\Omega}$ and the simple linear regression of Y on D.

What is the minimal (that is, weakest) assumption for that regression to identify δ^{T}, the average causal effect on the treated? In other words,[2] what is the minimal assumption for the OLS estimator of the slope in the simple linear regression of Y on D to converge in probability to δ^{T}?

As explained above, under the standard moment conditions of Chapter 1, Proposition 5, the limit in probability of the OLS estimator of the slope in the simple linear regression of Y on D is β_D. **Thus, the question becomes: under which conditions $\beta_D = \delta^{\mathrm{T}}$?**

For a binary treatment D, **Proposition 1 of Chapter 4** answers this question:

$$\boxed{\beta_D = \delta^{\mathrm{T}} + B}$$

and, therefore,

$$\boxed{\beta_D = \delta^{\mathrm{T}} \iff B = 0 \iff \mathbb{E}[Y(0)\,|\,D=1] = \mathbb{E}[Y(0)\,|\,D=0] \iff \mathbb{C}\mathrm{ov}(D, Y(0)) = 0}.$$

Remark that

$$D \perp\!\!\!\perp (Y(0), Y(1)) \implies \forall d \in \{0,1\}, \mathbb{C}\mathrm{ov}(D, Y(d)) = 0 \implies \mathbb{C}\mathrm{ov}(D, Y(0)) = 0,$$

The condition of answer 4 implies that of 3 which, in turn, implies the condition of 2.

The condition of Answer 2. $\mathbb{C}\mathrm{ov}(D, Y(0)) = 0$ is thus the minimal/weakest condition to consistently estimate the average treatment/causal effect on the treated, denoted δ^{T}, by the OLS slope estimator in the simple linear regression of Y on D (and, as always absent stated otherwise, a constant): $\widehat{\beta}_D$.

Furthermore, remember that, since D is binary, $\widehat{\beta}_D$ is also equal to $\overline{Y}_1 - \overline{Y}_0$ (Chapter 1, slide 9). Thus, another formulation of Proposition 1, Chapter 4 is the following: provided i.i.d. observations and mild moment conditions, the direct/naive difference of the averages of the observed outcome among the treated $\overline{Y}_1$ minus the average among the non-treated $\overline{Y}_0$ is a consistent estimator of δ^{T} if and only if D and $Y(0)$ are uncorrelated; equivalently, if and only if there is no selection bias.

1. None, it is always the case

   – **FALSE**, this is false and, if a single one, *the* thing you should remember from this course: in general, the OLS slope estimator of Y on D does not converge in probability to the causal parameter of interest δ^{T} (neither to the causal parameter δ).

   Another formulation in the setting of a binary instrument: **in general, absent specific assumptions or arguments ensuring that the selection bias is null (or, at least, limited or controlled in some way), the difference of the average observed outcome between the treated and the control ($\overline{Y}_1 - \overline{Y}_0$) does *not* provide any information about the causal effects of the treatment.**

   You can read newspapers, listen to the radio, watch the TV, browse the Internet: very soon, you will see that confusion is done; be careful . . . Even if we know it, if not attentive, we will also make the confusion. You should understand it and explain it to your mother, your father, your sister, your grandmother, or anyone who does not know it already: it is probably one of the biggest mistakes (among others) we make that prevent our understanding of reality.

   In this course, you study it in a formal mathematical set-up, but it is also essential to understand this result concretely. Another formulation of this result is the so-called **Simpson's paradox**.

[2] This is another equivalent formulation of the question to explain the meaning of "identify" here.

Here are some references and thus alternative formulations of Proposition 1 with different concrete examples:[3]

- the Wikipedia page about Simpson's paradox (*link*);
- a nice short video by David Louapre (Science Étonnante) about Simpson's paradox (*link*) or the associated blog post (*link*);
- another video about Simpson's paradox and confounding factors by Lê Nguyên Hoang (Science4All) (*link*).

2. $\mathbb{C}\text{ov}(D, Y(0)) = 0$

   – **True**, it is the crucial result of Proposition 1, Chapter 4.

3. $\mathbb{C}\text{ov}(D, Y(d)) = 0$ for all $d \in \{0, 1\}$

   – **False**, because this condition is not minimal; it is stronger than the condition of Answer 2.

   Actually, under this condition, we have, for any $d \in \{0, 1\}$

   $$\mathbb{C}\text{ov}(D, Y(d)) = 0 \iff Y(d) \text{ is mean-independent of } D,$$

   because D is binary (see the solutions to a problem set (TD6?) for details and the proof of this assertion). Therefore,

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

   As a consequence, by definition of δ and δ^{T}, using those equalities (and the linearity of expectations), we obtain

   $$\delta^{\mathrm{T}} := \mathbb{E}[\Delta \,|\, D = 1] = \mathbb{E}[Y(1) \,|\, D = 1] - \mathbb{E}[Y(0) \,|\, D = 1] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[\Delta] =: \delta,$$

   and, using the definition of Y and the conditioning events,

   $$\begin{aligned} \delta = \delta^{\mathrm{T}} &= \mathbb{E}[Y(1) \,|\, D = 1] - \mathbb{E}[Y(0) \,|\, D = 1] \\ &= \mathbb{E}[Y(1) \,|\, D = 1] - \mathbb{E}[Y(0) \,|\, D = 0] \\ &= \mathbb{E}[\underbrace{DY(1) + (1 - D)Y(0)}_{=:Y=Y(1) \text{ when } D=1} \,|\, D = 1] - \mathbb{E}[\underbrace{DY(1) + (1 - D)Y(0)}_{=:Y=Y(0) \text{ when } D=0} \,|\, D = 0] \\ &= \mathbb{E}[Y \,|\, D = 1] - \mathbb{E}[Y \,|\, D = 0] = \beta_D. \end{aligned}$$

   Consequently, we have: if $\mathbb{C}\text{ov}(D, Y(d)) = 0$ for all $d \in \{0, 1\}$, then $\beta_D = \delta^{\mathrm{T}} = \delta$.

4. $D \perp\!\!\!\perp (Y(0), Y(1))$

   – **False**, because this condition is not minimal; it is stronger than 3. and than 2.

   If $D \perp\!\!\!\perp (Y(0), Y(1))$, then $\beta_D = \delta^{\mathrm{T}} = \delta$ (see slide 12, Chapter 4, ideal set-up of **randomized experiments**).

   Indeed, $D \perp\!\!\!\perp (Y(0), Y(1))$ implies, a fortiori, that D and $Y(d)$, for $d \in \{0, 1\}$ are uncorrelated. Consequently, as shown above, for Answer 3, we have $\delta^{\mathrm{T}} = \delta$, and those causal parameters are identified by the slope of the linear regression of Y on D (and a constant).

[3] *Warning:* the terminology used in those other sources may not exactly be that of the course. Nonetheless, I think those references can be interesting to see other presentations and apprehend this crucial result for Econometrics 1 and, above all and more generally, to be able to act as a rational citizen understanding statistics/econometrics – Remark: the term confounding factor ("facteur de confusion") often used can be connected with the "omitted variable" of the "omitted variable bias" result (Chapter 4, Proposition 5)

3 Examples to apply the notions of Chapter 4

This question differs from standard Quiz questions: they are rather open questions for you to search, think about, and apply the notions of Chapter 4 in some concrete examples.

(a) 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:

Banks with More Women on Their Boards Commit Less Fraud

by Scott Berinato

From the Magazine (May–June 2021)

Enlightened by your Econometrics 1 course, what questions should you ask yourself?
Some examples/hints:[4]

1. To begin with, you can simplify the analysis by considering a binary treatment D that you will precisely define; in a second step, you can extend your reasoning to a non-binary, ordered, and quantitative treatment that you will specify. Give the concrete meanings in words of the potential outcomes $Y(d)$ for this example.
2. Do you think showing such a causal effect is, a priori, easy? Think about possible omitted variable bias issues. Would a controlled randomized experiment be easy to implement in this case?
3. In words, for this concrete application, discuss what would be a statistically significant but practically insignificant effect.

Some elements of solution for (a) *These are some elements of solution; the questions remain open; do not hesitate to think further about them.*

1.(a) The first thing to do is to specify the treatment and the outcome variable (after describing the sample and how it was collected, the data-generating process, but here it is an illustration without using data).

In an initial analysis, to simplify, we can consider a *binary* treatment D, defined for each bank on a given sample we can observe, as equal to 1 if at least one woman is a member of the board of directors (“conseil d’administration”), and 0 otherwise, when there are only men on the board.

In a second step, we can consider a non-binary, ordered, and quantitative treatment D. A first idea would be to define D_i as the number of women on the board of bank i. Yet, with that definition, if different banks have less or more numerous boards, the same absolute variation in the number D of women could correspond to different real-life meanings: for instance, moving from $D = 1$ to $D = 2$ might not have the same impact in a bank where the board is made up of 20 people compared to a

[4] Feel free to imagine other questions to apply the concepts of Chapter 4 to this concrete example (or another example you can find).

board of 4 people. In practice, we should further investigate the number of board members across banks (is there a lot of variation? On the contrary, do some legal obligations fix the same number of members of the board for all banks?) and, depending, probably favor a definition of D as a *relative* measure of the presence of women at the board. For instance, the percentage of women expressed between 0 and 100 on the board).

The outcome variable would be some quantitative measure of fraud (it should be further investigated and discussed in a genuine analysis, of course), say some given monetary amount of fraud. We skip this discussion here, but again, we should discuss how we can observe fraud (the outcome variable): in general and all the more so in this specific case since, by construction, fraud is something that should present some difficulty to be observed! Maybe we only observe a subsample of banks that are not good at committing fraud, and, compared to efficient, inconspicuous fraudster banks, those observed banks make less fraud and are only the hidden part of the iceberg. In that case, the analysis will suffer from an important flaw; maybe it is hard to do otherwise, but at least we should be aware of it.

1.(b) Then, given the treatment and outcome definitions, we can specify the concrete meaning of the potential outcomes.

(**Binary** treatment) When D is defined as the aforementioned binary indicator, $Y(0)$ is the monetary amount of fraud a bank would have committed if there was no woman on its board; while $Y(1)$ is the monetary amount of fraud a bank would have committed if there was at least one woman on the board. For a given bank, the realization of D is either 0 or 1; thus, *one of those two potential outcome variables is observed: $Y = Y(D)$ and the other is counterfactual.*

(**Multi-valued**, quantitative, ordered treatment) If D represents the percentage of women members of the board, then, for any $d \in [0, 100]$, $Y(d)$ is the monetary amount of fraud a bank would have committed had it d% of women on its board of directors.

2.(a) Discussions about identifying an average causal effect and possible omitted variable bias.

Here, if we imagine we have at our disposal a sample of banks assumed to be i.i.d. and representative of the population of interest we want to study (for instance, the banks in the European Union), we could regress Y on D (and possibly some control variables G).

That regression would identify some average causal effect provided there is no selection bias (and assuming linear effects in the case of a continuous treatment variable); more precisely and formally, see

- Proposition 1 in the case of a binary treatment;
- Proposition 2 in the case of a continuous treatment;
- Proposition 4 in the case of a continuous (or binary) treatment with controls.

Qualitatively, to investigate the presence of a selection bias, you should wonder: *how far/close the actual situation, that is, the way the data and in particular the treatment variable D was determined here, is from the set-up of a randomized experiment in which the treatment D is allocated randomly, independently of any other variable?*

Note: the answer to this question is not mathematical; it depends on a knowledge of society, the world as it is, the way the data is obtained, etc. *Difficulty (and interest) of econometrics*: make links between real situations and probabilistic properties of random variables used to model.

Imagine we were in a situation related to the labor market, cultural values, etc. (the point here is to illustrate econometrics concepts, not to do an authentic analysis) where men and women are as likely to be members of a bank's board of directors. In that situation, we could consider that the number or proportion of women within a board is as if randomly assigned.

Let us take a toy example to convey the underlying idea. Imagine that, instead of man or woman, we consider whether the birthday date of an individual is odd (“impair”) or even (“pair”), so that D is the fraction of individuals (men or women) within the board whose birthday dates are even (for instance, someone born on December 24th). With that definition (and given the situation of the labor market and designation of banks' boards), we can reasonably argue that D is allocated as if randomly: not because we organize an experiment to guarantee it, but because the natural way the data is generated ("natural experiment") could be deemed as if it generates D randomly; formally, we would assume $D \perp\!\!\!\perp \{Y(d)\}_{d \in \text{Support}(D)}$. In that case, a simple linear regression of Y on D would identify a causal parameter (an average effect; its exact definition depends on the assumed restriction about the heterogeneity of treatment effects: see Propositions 2 (δ^W and 4 (δ_0), and Question 10 of the quiz).[5]

Given the current labor market, we can argue that the characteristic "man or woman" is quite different from the characteristic "odd or even birthday date": *if we take at random a member of a bank's board, it is more likely this member is a man instead of a woman.* ***Important****: yet, this point itself (the fact that the proportion is not 50%-50%) is* not *a problem in terms of selection bias.*

To understand that, imagine that instead of odd or even birthday dates (whose repartition is close to 50%-50%), D is the fraction of individuals (men or women) within the board whose birthday dates are a multiple of 10 (born on 10th, 20th, or 30th). If we take at random members of banks' boards, we can imagine the probability that they satisfy this condition is close[6] to $3/30 = 1/10$, and not half-half. Despite that, the crucial point is that it is likely that this probability is the same across banks; that is, that D is drawn independently, does not correlate with any banks' characteristics, and thus can be considered as being randomly determined, as if in an experiment.

Regarding men/women, for various reasons, we can reasonably think it is *not* the same situation; we can think that due to some specific characteristics (observed or not), in some banks, the probability that a board member is a woman is larger than in other banks. In other words, men or women are not allocated as if randomly across banks' boards of directors. We can think that the data-generating process from a natural experiment, without any particular organization like an experiment,[7] is relatively far from the situation of a randomized experiment. Thus, if no experiment was organized,[8] the data was just collected as it comes without a specific organization, and there is a risk of selection bias.

A possible story behind that: maybe the banks that have more women on their boards are banks that are more concerned about equity issues and, as a consequence, also tend to commit less fraud.

The selection bias can be interpreted in terms of omitted variable bias (see Proposition 5 of Chapter 4). In that case, the omitted variable bias would be some measure G (hard to define precisely, even more to measure in practice!) of the equity concern of a bank:

- G is correlated with the outcome: banks more concerned about equity/justice/fairness etc. are less likely to commit fraud;
- AND (we need BOTH for an omitted variable bias) G is correlated with D: because the labor market is such that the allocation of men and women in boards is not random, banks more concerned about equity/fairness may tend to have more women in their boards.

In that case, G ("a confounding factor" – you can find in various fields that alternative terminology) would entail an omitted variable bias: a naive difference in the amount of fraud between banks with women on their board ($D = 1$) and banks without ($D = 0$) does not capture the causal effect of having more women on their boards but is contaminated by the omitted variable bias of the equity concern of banks (see Proposition 5, Chapter 4).

2.(b) Would a controlled randomized experiment be easy to implement in this case? Although connected to the previous interrogation, this question is different:

[5] Maybe, in this case, we would obtain a statistically significant effect by chance and conclude: "Banks with more even-birthday-date members commit more fraud." However, from that result, we should not conclude something is interesting here; rather, it illustrates that statistically significant results should not be pursued per se and are not always interesting/relevant – see the discussions and criticisms about p-value, Question 15 of Quiz 2.

[6] We could do the exact computation (with the different number of days per month and leap years), but it is not important here.

[7] A quotation (abusive but interesting nonetheless) linked to that idea: "no causation without manipulation"

[8] The previous discussion does not say that, in the specific case of this article, the data-generating process used is not an experiment. Maybe researchers organized a controlled randomized trial to study this question, and the article is written based on this experiment (you can read the article to know).

a. One thing is to ask ourselves, and it is an important reflex to have: *given the actual way the data was collected (data-generating process – d.g.p), how far/close is it to the set-up of an experiment?* The answer to this question qualitatively informs about the presence/absence/importance of selection bias and whether a regression identifies an average causal parameter of interest.

b. Another thing is to wonder: given the treatment and outcome variables, is it possible, easy to organize, and effectively implement an experiment?

The last part of this question relates to Question b.

The point is to remark that, in this setting, it would probably be quite challenging to organize and implement an experiment. In concrete terms, it doesn't seem easy to imagine that a researcher can convince and enforce a bank to randomize the composition of its board of directors in terms of men and women.

Another example: if we consider the set-up of tennis, sleep, and alcohol consumption of one of the problem sets, it will entail some legal and ethical difficulties to draw the number of hours slept and quantity of alcohol drunk at random and force some participants to respect them. It is another case where organizing an experiment in practice seems complicated.

Questions a. and b. are questions. Most of the time, we can expect that when the answer to b. is negative (that is, it would be complicated to organize an experiment), it is likely that the answer to a. is negative too in the sense that the natural d.g.p is likely to be far from a randomized experiment. *Yet, it is not always the case, and a. and b. remain two distinct questions.* If we resume the silly but illustrative example of the treatment "even birthday-date":

- it would be as difficult (if not even more than for man/woman!) to organize and implement an experiment. Concretely, it would entail saying to the board of directors of some big bank: "no, no, you should not choose this member even if you would like to; instead, you need to flip a coin and play heads or tails to determine if you choose member A born the 13th (an odd date) or member B born the 6th (an even date);
- yet, as discussed before, it is likely that in real life, the natural way the data is generated is such that this characteristic odd/even birthday is drawn as if randomly.

Thinking about question b. is also the occasion to say that, although a theoretical comparison between the real data-generating process and the benchmark of an experiment is interesting conceptually to investigate the presence of selection bias, randomized experiments cannot be an answer for all questions. Some (if not most) interesting questions could probably *not* be answered through randomized experiments.[9] One reason is that implementing such experiments is impossible; hence the interest of pursuing Econometrics 1 course with Chapter 5 about instrumental variables and then Econometrics 2 that will present other methods to identify causal effects.

3. Statistically versus practically significant effects. This question relates to the distinction introduced in Chapter 2, slides 27 and 28; see also the reference about the "(likely) effect sizes" in Quiz 2, Question 15.

Here, an example would be to obtain a result (*admitting we can interpret it as an estimate of an average causal effect*) that

- is statistically significant at usual levels, for instance, at 1%;
- but, happens to be **practically** insignificant. To detect that, you need to *make a literal sentence to interpret the result of a regression.* To do so, you must be careful about: (*i*) the form of the model (level-level, log-level, etc.); (*ii*) the units and meaning of the variables entering the regression (see Question 5 of this quiz for details).

[9] On this topic, linked to the criticisms about doing science mainly through p-values, you can look at the references presented in Quiz 2, Question 15.

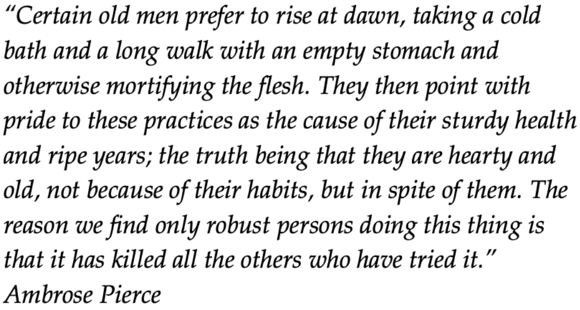
Example: imagine $D \in ([0, 100])^{\Omega}$ represents the percentage of women on banks' boards, and Y is the *annual* monetary amount of fraud expressed in *euros*, and the associated estimate for the coefficient of D is -200.

At first sight, one might think that -200 is a "big effect"; *but without considering the units, we cannot know!*.

Here, in this case, the interpretation (assuming it can be causal) is the following: all other things equal, an additional percentage point of the fraction of women on boards decreases the annual fraud committed by the bank by 200 euros. Thus, even if we move from a situation with 0% of women to that of 100% of women on board, it would imply (assuming linear effects) a decrease of $100 \times 200 = 20,000$ euros per year. For a bank, it is tiny, insignificant.[10] This would be an example of a statistically significant yet practically insignificant effect.

If instead, while the associated coefficient remains estimated at -200, the unit of Y was the *daily* amount of fraud expressed *in millions of euros*, this would be another story and, in this case, a pretty significant effect.

(b) Below is a quotation[11] from a former version of ENSAE Statistics 1 course:



> *"Certain old men prefer to rise at dawn, taking a cold bath and a long walk with an empty stomach and otherwise mortifying the flesh. They then point with pride to these practices as the cause of their sturdy health and ripe years; the truth being that they are hearty and old, not because of their habits, but in spite of them. The reason we find only robust persons doing this thing is that it has killed all the others who have tried it."*
> *Ambrose Pierce*

1. In this example, what would be the binary treatment D? Specify in words the concrete meaning of the potential outcomes $Y(0)$ and $Y(1)$ here.

2. Discuss the selection bias problem evoked in this example, notably by the phrase: "*not because of their habits, but in spite of them*".[12] The quotation also evokes another problem of selection, namely the selection of units *into the sample*: which units do we observe?[13] as opposed to the selection of observations *into the treatment*: which observations receive the treatment? You shall forget this additional dimension here (see next semester in Econometrics 2) to focus on the selection bias into the treatment studied in Chapter 4, in particular in slides 8 and 9.

Some elements of solution for (b) *The solutions for (a) were supposed to be quick, but already too long eventually; so concise elements of solution for (b) below.*

1. Treatment and potential outcomes Here, we could define D as a binary indicator of doing the following routine: "rise at dawn, taking a cold bath and a long walk with an empty stomach and otherwise mortifying the flesh". The outcome Y would be some index of health (see, for instance, the illustration at the beginning of Chapter 4) with higher values for Y indicating better health condition.

[10] For information, the total assets of the largest French bank (you can find which one it is) are estimated at about 2.6 trillion euros in 2021; the French GDP (in values) in 2021 is 2.9 trillion US dollars – one trillion = one thousand billion. Warning: a billion in English = "un milliard" in French (10^9), but in French "un billion" = one trillion in English (10^{12}).

[11] The author is probably Ambrose *B*ierce rather than Pierce – Wikipedia page *link*.

[12] For a French literary analogue: "Ma mère s'émerveillait qu'il [M. de Norpois] fût si exact quoique si occupé, si aimable quoique si répandu, sans songer que *les « quoique » sont toujours des « parce que » méconnus*", Marcel Proust, *À l'ombre des jeunes filles en fleur* (je mets en italique, et non le texte original).

[13] See, for instance, Quiz 1, Question 13 on that topic.

$Y(0)$ is the potential health condition index an individual would have if she or he had not followed the routine as mentioned above.

$Y(1)$ is the potential health condition an individual would have if she or he had followed the routine.

2. Selection bias Here, the treatment (following the routine) is very far from being randomly assigned. In other words, we are very far from the situation of a randomized experiment to generate the data we see.[14] Indeed, individuals freely decide to follow or not this routine: *most of the time, when agents can decide something, they do not do it randomly, and it induces some selection bias* (for an extreme example, see the model 2 of self-selection in one of the Problem Sets). Indeed, they decide based on their specific (unobserved) characteristics that also affect their potential outcomes: hence D and the potential outcomes are likely to be correlated.

Here, it is suggested that we would identify $\mathbb{E}[Y \mid D = 1] = \mathbb{E}[Y(1) \mid D = 1]$ and estimate a large value: the treated people following the routine we see have "sturdy health"; while we would identify and estimate $\mathbb{E}[Y \mid D = 0] = \mathbb{E}[Y(0) \mid D = 0]$ at a lower value probably: the people who do not follow the routine have less sturdy health. Yet, the quotation suggests that

$$\mathbb{E}[Y(0) \mid D = 1] > \mathbb{E}[Y(0) \mid D = 0].$$

The people with $D = 1$ are the more robust[15] and, did they not follow their routine, they would have had better health than those who do not follow the routine anyway.

Essentially, here, the routine/treatment is not responsible for their high value for the outcome. On the contrary, the quotation suggests the causal effect of D is negative. Yet, it also suggests that the selection bias might be sufficiently important that, despite the negative causal impact of the treatment, the treated individuals appear in better health than the control:

$$\mathbb{E}[Y \mid D = 1] = \mathbb{E}[Y(1) \mid D = 1] > \mathbb{E}[Y(0) \mid D = 0] = \mathbb{E}[Y \mid D = 0]. \tag{1}$$

To simplify, we assume that the causal/treatment effect is homogeneous and that there exists τ, a positive non-stochastic real number, such that $\Delta = -\tau$.

We thus have

$$Y(D) \underbrace{:=}_{\text{def. } Y} DY(1) + (1 - D)Y(0) \underbrace{=}_{\text{computation}} Y(0) + D\big[\underbrace{Y(1) - Y(0)}_{:=\Delta \text{ (def)}}\big] = Y(0) + D\Delta \underbrace{=}_{\text{ass. } \Delta=-\tau} Y(0) - D\tau.$$

For the inequality in (1) to hold, we should have

$$\begin{aligned} &\mathbb{E}[\underbrace{Y(1)}_{=Y(0)-D\tau} \mid D = 1] = \mathbb{E}[Y(0) - \tau \mid D = 1] > \mathbb{E}[Y(0) \mid D = 0] \\ \iff &\underbrace{\mathbb{E}[Y(0) \mid D = 1] - \mathbb{E}[Y(0) \mid D = 0]}_{=: B} > \tau, \end{aligned}$$

namely a selection bias B sufficiently large (compared to the negative impact of the routine).

4 Correlation between D and $Y(0)$

We are in the setting of Chapter 4, Section 1: in particular, D is the binary treatment, and $Y(0)$ is the potential outcome absent the treatment.

In this context, Proposition 1 of Chapter 4 states a sufficient and necessary condition for the simple linear regression of Y on D to identify the quantity denoted δ^{T} (that is, a sufficient and necessary condition for the OLS estimator of the slope in that regression to converge in probability to δ^{T}). This condition can be written $\mathbb{C}\mathrm{ov}(D, Y(0)) = 0$.

[14] As indicated, I neglect here the selection *in the sample* issue: the fact that we only observe the more robust individuals. A chapter of Econometrics 2 next semester will be devoted to this question.

[15] For simple modeling of the situation (to be developed), we can assume there exists an unobserved binary variable R that indicates the robustness of the individual, whether she is robust ($R = 1$) or not ($R = 0$).

(a) Write the relationship between β_D, δ^{T}, and B given by Proposition 1 and deduce another equivalent formulation of the condition $\mathbb{C}\mathrm{ov}(D, Y(0)) = 0$.

Reminders/repetitions (but very important):

$\widehat{\beta}_D$ denotes the OLS estimator of the coefficient associated with D in the simple linear regression of Y on D, and β_D is its probability limit (under the usual moment conditions of Chapter 1, Proposition 5), that is, the coefficient associated with D in the theoretical linear regression:

$$\widehat{\beta}_D := \frac{\widehat{\mathbb{C}\mathrm{ov}}(Y, D)}{\widehat{\mathbb{V}}[D]} \xrightarrow[n\to+\infty]{P} \beta_D := \frac{\mathbb{C}\mathrm{ov}(Y, D)}{\mathbb{V}[D]} \overset{\text{because } D \text{ binary}}{=} \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0]$$

$$\overset{\text{by def. of } Y := Y(D) \text{ and the conditioning}}{=} \mathbb{E}[Y(1) \mid D = 1] - \mathbb{E}[Y(0) \mid D = 0].$$

δ^{T} is the average treatment effect on the treated:

$$\delta^{\mathrm{T}} := \mathbb{E}[\Delta \mid D = 1] = \mathbb{E}[Y(1) \mid D = 1] - \mathbb{E}[Y(0) \mid D = 1].$$

B is the selection bias:

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

where the quantities in orange are *counterfactual* and cannot be identified with the data.

Result of **Chapter 4, Proposition 1** (case of a single binary covariate/treatment D):

$$\beta_D = \delta^{\mathrm{T}} + B.$$

Consequently, we have $\beta_D = \delta^{\mathrm{T}}$ if and only if $B = 0$. An equivalent formulation of the condition $\mathbb{C}\mathrm{ov}(D, Y(0)) = 0$ is thus $B = 0$, which means that

$$\mathbb{E}[Y(0) \mid D = 1] = \mathbb{E}[Y(0) \mid D = 0],$$

in words: what would have been the outcome absent the treatment ($Y(0)$) is the same, on average (expectations $\mathbb{E}$), between the treated individuals (knowing $D = 1$) and the control individuals (knowing $D = 0$).

(b) The condition $\mathbb{C}\mathrm{ov}(D, Y(0)) = 0$ is thus fundamental to determine whether a linear regression identifies a causal effect. Hence, we would like to test this condition against the alternative $\mathbb{C}\mathrm{ov}(D, Y(0)) \neq 0$.

Is it possible to do so? Choose the unique correct assertion below and justify your answer.

The answer is given explicitly in Chapter 4, first bullet-point, slide 15:

$$\boxed{\textit{We cannot test } \mathbb{C}\mathrm{ov}(D, Y(0)) = 0 \textit{ because we do not observe } Y(0) \textit{ when } D = 1.}$$

The key point is that $Y(0)$ is “partly unobserved/counterfactual”; indeed, we have only access to $Y := Y(D) = DY(1) + (1 - D)Y(0)$ when D is binary, and Y **is equal to** $Y(0)$ **only when** $D = 0$.

Intuitively, testing $\mathbb{C}\mathrm{ov}(D, Y(0)) = 0$ requires looking at how $Y(0)$ varies in the population/data-generating process (by looking at what happens in a representative sample of observation) when D varies. Yet, it is *impossible to do so* because, by construction, we can only observe $Y(0)$ when $D = 0$; when $D \neq 0$ (that is, $D = 1$), $Y(0)$ is unobserved.

Thus, by definition of the observed outcome variable

$$Y := Y(D) \overset{\text{when Support}(D) = \{0,1\}}{=} DY(1) + (1 - D)Y(0),$$

we *always* have this problem, that is, we can *never* test

$$H_0 : \mathbb{C}\mathrm{ov}(D, Y(0)) = 0 \text{ against } H_1 : \mathbb{C}\mathrm{ov}(D, Y(0)) \neq 0,$$

whatever the context (natural experiment, randomized control trials, etc.).

In particular, even if, in "reality", in other words, in the actual data-generating process (d.g.p) that generates the observations, we have $D \perp\!\!\!\perp (Y(0), Y(1))$ (because of the organization of a randomized controlled trial, for instance), we cannot make the test because it does not change in any way that $Y(0)$ can only be observed when $D = 0$ and is counterfactual otherwise when $D = 1$.

As mentioned in Chapter 4, slide 15, it is possible, nonetheless, to test close conditions. I refer to a complementary document available in Pamplemousse for further details on that topic – more advanced but quite interesting.

1. We can *always* test $\mathbb{C}\mathrm{ov}(D, Y(0)) = 0$; it amounts to looking at the empirical covariance between D and Y in the subsample of non-treated units (namely, with D equal to 0).

   – **False**.

2. In general, we cannot test $\mathbb{C}\mathrm{ov}(D, Y(0)) = 0$; however, we can test it in the specific set-up of randomized experiments.

   – **False**.

3. We can *never* test $\mathbb{C}\mathrm{ov}(D, Y(0)) = 0$ (even in the set-up of randomized experiments).

   – **True**.

4. None of the previous assertions; if so, indicate below the correct one.

   – **False**.

5 Accounting for nonlinearities

We consider the model

$$\forall p \in \mathbb{R}_+^*,\ \log Y(p) = \gamma_0 - \delta_0 \log p + \eta, \tag{2}$$

where,

- for any positive price $p \in \mathbb{R}_+^*$, the potential outcome variable $Y(p) \in (\mathbb{R}_+^*)^\Omega$ is the demand (assumed positive to take its logarithm) for some good when the price of the good is equal to p, that is, the quantity asked (measured in some physical unit of measure; for instance, 3 tons of apples) when the price is equal to p;
- $\gamma_0 \in \mathbb{R}$ and $\delta_0 \in \mathbb{R}$ are two scalar non-stochastic parameters;
- and $\eta \in \mathbb{R}^\Omega$ is a real random variable.

In other words, more formally, we assume the following proposition holds:

$$\exists\, (\delta_0, \gamma_0) \in \mathbb{R}^2, \exists\, \eta \in \mathbb{R}^\Omega :\ \forall p \in \mathbb{R}_+^*,\ \log Y(p) = \gamma_0 - \delta_0 \log p + \eta.$$

In model (2), how can we interpret the parameter δ_0?

There are two elements to answer the question.

First, *be careful*, here the model is written with a *minus*: $-\delta_0$. Hence, the following formulations in the different possible answers: an increase of the price will lead to a *decrease* of the demand by something linked to δ_0 (with a precise formulation that is the topic of the second point).

Second, the question relates to slide 19 of Chapter 4 and the distinct interpretation between level-level, log-level, level-log, and log-log models.

Question 12 of Quiz 1 already provides explanations.

Below is a summary of the different usual linear regressions that can link an outcome variable of interest Y and an explanatory variable of interest D.

Remark: the interpretations below are formulated in terms of prediction, which is always possible for any linear regression. If the linear regression does identify causal effects, *which is not always the case!* (the core issue of Chapter 4), the interpretation can also be made in terms of causal effects.

Remark: the regression may also include additional control variables G; they did not change the interpretation except by adding that the interpretation is to be understood all else (namely, the other covariates G) being equal (henceforth abbreviated to "a.e.b.e").

Let β_D denote the coefficient associated with D (or with $\log(D)$) in the theoretical linear regression of Y (or of $\log(Y)$) on D (or on $\log(D)$) and, possibly, additional control variables G.

Below, the conventional fuzzy notation Δ denotes a variation; it is used to write a shortcut symbolic expression to memorize the interpretation.[16]

The precision "approximately" refers to the fact that this is only an approximation, valid for a small variation of D (see the handwritten solutions to TD1 for details or Figure 1).

- **level-level**: regression of Y on D (and G): $\boxed{\Delta Y = \beta_D \times \Delta D}$

  If D increases by 1 unit (absolute change), we predict that, a.e.b.e, Y changes by β_D units (absolute change).

- **log-level**: regression of $\log(Y)$ on D (and G): $\boxed{\%\Delta Y \approx 100\beta_D \times \Delta D}$

  If D increases by 1 unit (absolute change), we predict that, a.e.b.e, Y approximately changes by $100\beta_D\%$ (relative change).

  Reminder: imagine Y changes from 200 to 220, the *absolute* change is $220 - 200 = 20$ while the *relative* change is $(220 - 200)/200 = 0.1 = 10/100 = 10\%$.

  Remark: this specification is sometimes called a semi-log model and β_D a semi-elasticity.

- **level-log**: regression of Y on $\log(D)$ (and G): $\boxed{\Delta Y \approx (\beta_D/100) \times \%\Delta D}$

  If D increases by 1% (relative change), we predict that, a.e.b.e, Y approximately changes by $\beta_D/100$ units (absolute change).

- **log-log**: regression of $\log(Y)$ on $\log(D)$ (and G): $\boxed{\%\Delta Y \approx \beta_D \times \%\Delta D}$

  If D increases by 1% (relative change), we predict that, a.e.b.e, Y approximately changes by $\beta_D\%$ (relative change).

  β_D is called the elasticity of Y with respect to D (or the D-elasticity of Y). However, remark that the log-log model imposes a *constant* elasticity, which is a (strong) assumption.[17]

Here, Y is the demand, D is the price, and we consider a log-log model; hence correct answer 4.

1. An increase of 1% of the price decreases the demand by δ_0 units.

   – **False**, provided δ_0 is replaced by $\delta_0/100$ and with the precision "approximately", it would correspond to a level-log model.

2. An increase of 1 euro of the price decreases the demand by δ_0 units.

   – **False**, it would correspond to a level-level model.

3. An increase of 1 euro of the price decreases the demand by $100\delta_0\%$ approximately.

   – **False**, it would correspond to a log-level model.

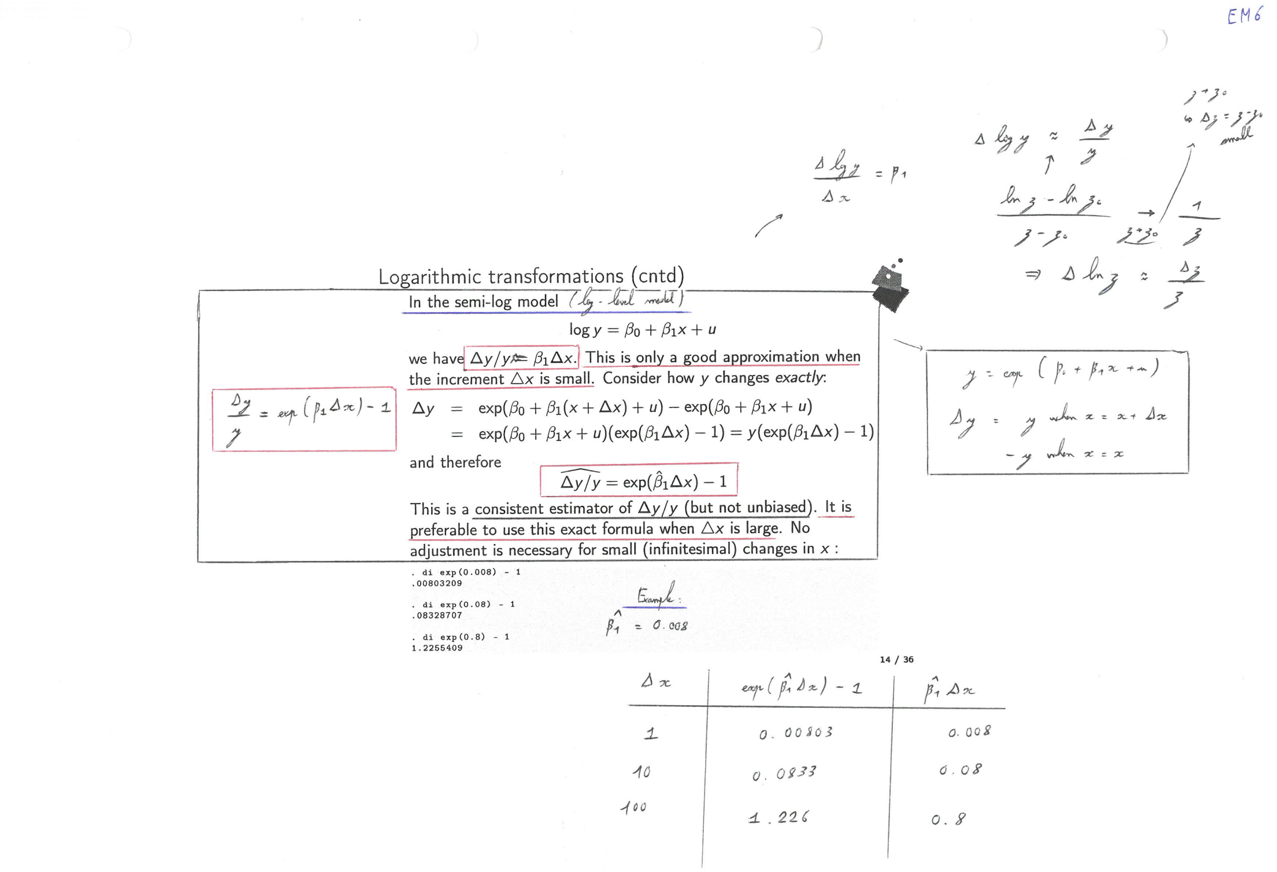
4. δ_0 is the elasticity of the demand with respect to the price; that is, an increase of 1% of the price decreases the demand by $\delta_0\%$.

   – **True**, remark that we assume here a constant elasticity (this is why we can speak of "the" elasticity of ...). Depending on the support of D and the possible variations we consider, it might be a strong, if not unrealistic, assumption: in general, your price-elasticity of the demand for some good depends on the quantity of the good you already have.

[16] Not to be confused with the individual causal effect $\Delta := Y(1) - Y(0)$ for a binary treatment; nothing to do here, of course.

[17] In general, there is no reason that the elasticity between two variables is constant. Indeed, *an elasticity is a local notion*, it is a function, similar to a derivative, and, in general, should be considered/evaluated at a given point: the elasticity of Y with respect to D for $D = d$, for some d in the support of D.

Figure 1: Handwritten notes about the interpretation of log-level models.



EM6

$$\frac{\Delta \log y}{\Delta x} = \beta_1$$

$$\Delta \log y \approx \frac{\Delta y}{y}$$

$$\frac{\ln z - \ln z_0}{z - z_0} \xrightarrow[z \to z_0]{} \frac{1}{z}$$

$z \to z_0$ $\Leftrightarrow \Delta z = z - z_0$ small

$$\Rightarrow \Delta \ln z \approx \frac{\Delta z}{z}$$

Logarithmic transformations (cntd)

In the semi-log model (log-level model)

$$\log y = \beta_0 + \beta_1 x + u$$

we have $\Delta y / y \approx \beta_1 \Delta x$. This is only a good approximation when the increment $\triangle x$ is small. Consider how y changes *exactly*:

$$\begin{aligned} \Delta y &= \exp(\beta_0 + \beta_1(x + \Delta x) + u) - \exp(\beta_0 + \beta_1 x + u) \\ &= \exp(\beta_0 + \beta_1 x + u)(\exp(\beta_1 \Delta x) - 1) = y(\exp(\beta_1 \Delta x) - 1) \end{aligned}$$

$$\boxed{\frac{\Delta y}{y} = \exp(\beta_1 \Delta x) - 1}$$

$$y = \exp(\beta_0 + \beta_1 x + u)$$

$\Delta y = y$ when $x = x + \Delta x$ $- y$ when $x = x$

and therefore

$$\boxed{\widehat{\Delta y / y} = \exp(\hat{\beta}_1 \Delta x) - 1}$$

This is a consistent estimator of $\Delta y / y$ (but not unbiased). It is preferable to use this exact formula when $\triangle x$ is large. No adjustment is necessary for small (infinitesimal) changes in x :

```
. di exp(0.008) - 1
.00803209

. di exp(0.08) - 1
.08328707

. di exp(0.8) - 1
1.2255409
```

Example:

$\hat{\beta}_1 = 0.008$

14 / 36

Δx	$\exp(\hat{\beta}_1 \Delta x) - 1$	$\hat{\beta}_1 \Delta x$
1	0.00803	0.008
10	0.0833	0.08
100	1.226	0.8

6 Interpretation of a linear regression

We consider the simple linear regression, where an observation is an American town, of Y on D, where Y is the share of votes expressed in % for the Republican party in the town, and D is a binary variable equal to 1 if the media Fox News is available in the town, 0 otherwise.

As in the course (see, for instance, the non-causal linear representation in slide 14), we denote by α_0 and β_D the intercept and the slope coefficient in the theoretical regression. As a reminder: they are the limit in probability of the OLS estimators $(\widehat{\alpha}, \widehat{\beta}_D)$ under the appropriate moment conditions and i.i.d sampling of Chapter 1, Proposition 5.

What can we say from this regression?

The answer to this question relies on three points.

First, by construction, a linear regression can *always* be interpreted in terms of best (in terms of Mean Square Error) linear predictions. On the contrary, *additional conditions are required for a linear regression to estimate/identify causal effects of interest* (This is the core issue of Chapter 4, see in particular slide 14 and the last bullet-point).

Here, the statement of the question does not assume such additional conditions, and there is no reason a priori to believe/assume that they hold. In other words, here, there is no reason a priori that β_D is equal to a causal parameter of interest.

Methodological reflex to argue that: we should wonder *to what extent the situation of interest, the d.g.p under study is far or close to the set-up of a randomized experiment in which the treatment is randomly assigned, independently of any other variable?*

Here, in this concrete example, there is no reason to think that Fox News decides to settle randomly across American towns; or, for another formulation, it seems plausible that towns with Fox News $(D = 1)$ differ from towns without Fox News $(D = 0)$ in terms of what would have been the share of votes for the Republican Party in the town if Fox News was not available (which potential outcome $Y(0)$ – *Additional exercise*: what is the likely sign of the selection bias in this context? Typical quiz question).

But, here, we do not care because answers 1 to 3 are formulated in terms of *prediction*: "We predict ...", and *we can always do it*. Hence, answer 4 is false.

Second, for a correct interpretation, as in Question 5 of this Quiz, we need (*i*) to determine whether the model is a level-level model, a log-level model, etc., and (*ii*) be careful about the units of measurement of the variables Y and D.

Here, D is a binary variable that enters as such in the regression.[18] Y is the share of votes expressed in % for the Republican Party (say, at the previous presidential election). It is a continuous, quantitative variable with values ranging between 0 and 100 (because it is expressed in percentages). Again, as described in the question, it enters the regression as such without a logarithm. In conclusion, we are in a level-level model.

Third, we need to be careful about the expression of an absolute change for a variable measured in percentage. We consider absolute changes because we are in a level-level model.

For example, imagine the percentage of vote change from 50% to 60%. Describing this variation as "an increase of 10%" is ambiguous and should be avoided (although you will hear such expressions in the news daily ...[19]). The proper way is to say: the percentage of the vote has increased by 10 percentage points[20] (p.p) (10 p.p. = 0.10, a percentage point is simply equal to 0.01; it is a unit used to compare percentages); 10 percentage points is the *absolute* change, and corresponds to a *relative* change equal to $(0.60 - 0.50)/0.50 = 0.1/0.5 = 1/5 = 2/10 = 20\%$.

Hopefully, this should clarify the difference between answers 1 and 2 and why the former is the correct one.

1. We predict an increase of β_D percentage points (1 percentage point = 0.01) in favor of the Republican party in a town with Fox News compared to a town without Fox News.

   – **True**.

2. We predict an increase of $\beta_D\%$ in the share of the vote in favor of the Republican party in a town with Fox News compared to a town without Fox News.

   – **False**.

3. We predict an increase of β_D percentage points in favor of the Republican party if the audience of Fox News increases by 1% in the town.

   – **False**.

4. None of the previous interpretations is valid in general: to make them, we have to assume that the selection bias is null.

   – **False**.

[18] Meaning without logarithm. Note that taking the logarithm would make no sense for a binary variable. Indeed, a binary variable is essentially something *qualitative* (encoded as 0 and 1), and applying some mathematical function (doing computation) to it makes no sense. Furthermore, the logarithm is not properly defined at 0.

[19] Remark that you can hear much worse, namely assertions that have not yet understood Chapter 4, Proposition 1 or, in other words, the notion of selection bias (see also Simpson's paradox).

[20] "point de pourcentage" – `https://fr.wikipedia.org/wiki/Point_de_pourcentage`.

7 Regressions and causal effects with a non-binary treatment

This question relates to Section 2 of Chapter 4: the case of a *single* non-binary but ordered and quantitative treatment D.

(a) Write the definitions of W and δ^W as defined in Chapter 4.

This is directly the first elements of Chapter 4, slide 20. Provided D admits a finite second-order moment, we define

$$W := \frac{(D - \mathbb{E}[D])^2}{\mathbb{V}[D]},$$
$$\delta^W := \mathbb{E}[W\Delta],$$

where Δ is defined in Equation (Lin. effects), slide 18: Δ is the linear causal effect of D on Y, that is, we assume the following holds:

$$\boxed{\exists!\,\Delta \in \mathbb{R}^\Omega,\ \exists\, d_0 \in \text{Support}(D):\ \forall d \in \text{Support}(D),\ Y(d) = Y(d_0) + \Delta(d - d_0)} \qquad \text{(LCE)}$$

where $Y(d)$ is the potential outcome variable associated with the value d of the treatment D.

Assumption (LCE) (for Linear Causal Effect) gives the existence and uniqueness of the random variable Δ, thus defined as “the” (it has a sense due to linearity) causal effect of the treatment D on the outcome Y.

Remarks about (LCE)**.**

As explained in the course, if (LCE) holds, it holds as well for any other reference point $d_1 \in \text{Support}(D)$; d_0 (or d_1) only plays the role of a reference point, from which the slope (the linear causal effect) is defined.

Remember that, a priori, Δ varies across individuals (heterogeneous causal effects); mathematically, it is defined as a real random variable: $\Delta \in \mathbb{R}^\Omega$.

If the treatment D is a continuous variable (admitting a density with respect to Lebesgue measure), Δ can be interpreted as “the” derivative of the potential outcome function $\text{Support}(D) \ni d \mapsto Y(d)$. Due to the linearity assumption, the derivative is constant equal to Δ (for a given individual). In that case, Δ is also called the causal marginal effect of the treatment, denoted (with the notation of slide 18) $\Delta = \dfrac{\partial Y(d)}{\partial d}$.

In particular, remark that for any d in the support of D such that $d+1 \in \text{Support}(D)$ too, we have

$$\Delta = Y(d+1) - Y(d).$$

If it holds for $d = 0$, we formally retrieve the definition of Δ in the special case of a binary treatment ($\text{Support}(D) = \{0, 1\}$), $\Delta := Y(1) - Y(0)$. Note that, in this case, we do not assume linearity: it holds by construction as we can only move between two points. On the contrary, *with non-binary treatments,* (LCE) *is a linearity assumption.* Nonetheless, remember that we can consider transformations of D and Y as explanatory and explained variables (see slide 19 “Accounting for nonlinearities” and Question 5).

(b) Give a sufficient condition (but possibly strong assumption) that implies $\delta^W = \delta$.

We propose three conditions: (A), (B), and (C).

Note that, in the case of a non-binary treatment, we do not consider δ^{T}, because the definition of δ^{T} is only relevant for binary treatments: restriction to the treated sub-population, conditional on $D = 1$; for non-binary treatment D, $D = 1$ is a treatment status among others, without particular meaning in general.

(A) A sufficient condition that implies the equality $\delta^W = \delta$ is to assume *homogeneous causal effects,* namely $\exists\, \delta_0 \in \mathbb{R}$ non-stochastic : $\Delta = \delta_0$, that is, Δ is a constant random variable.

Note that Equation (Lin. effects) of Chapter 4, or (LCE) above, assumes a linear effect of D on Y, but the causal effects are *heterogeneous: Δ is a random variable, a priori specific to each individual.* In other words, for each individual, the effect of an increase of D to $D+1$ on Y does not depend on the initial value of D we start from (linear effect), but can be individual-specific (heterogeneous effect).

Remark that the linearity assumption is not necessarily a very strong one in the sense that *we only need a linear relationship between a known transform of $Y(d)$ and a known transform of d* (see slide 19); plus, we can allow d and Δ to be multivariate (see Model (Lin. mod. 2), slide 31).

In contrast, in most settings, the assumption of homogeneous causal effects is a strong hypothesis. Under this assumption, we already know that $\delta = \delta_0$ (see Question 1 (c) of this quiz).

It also implies

$$\begin{aligned}
\delta^W &:= \mathbb{E}[W\Delta] \quad \text{(definition of } \delta^W\text{)} \\
&= \mathbb{E}[W\delta_0] \quad \text{(homogeneous causal effects assumption)} \\
&= \mathbb{E}[W] \times \delta_0 \quad \text{(linearity of expectation – } \delta_0 \text{ is a parameter, non-stochastic)} \\
&= \mathbb{E}\left(\frac{(D - \mathbb{E}[D])^2}{\mathbb{V}[D]}\right) \times \delta_0 \quad \text{(definition of } W\text{)} \\
&= \frac{\mathbb{E}\big((D - \mathbb{E}[D])^2\big)}{\mathbb{V}[D]} \times \delta_0 \quad \text{(linearity of expectation – } \mathbb{V}[D] \text{ is not stochastic)} \\
&= \frac{\mathbb{V}[D]}{\mathbb{V}[D]} \times \delta_0 \quad \text{(definition of a variance for the numerator)} \\
&= \delta_0.
\end{aligned}$$

Thus, the assumption of homogeneous causal effects implies $\delta^W = \delta$.

This result is logical. Indeed, δ^W can be interpreted as a *weighted* average of individual causal effects (with larger weights for individuals whose treatment value D is further from the expectation $\mathbb{E}[D]$), and δ as the unweighted average of individual causal effects. If we assume that the individual causal effects are all equal, those two averages coincide (and, more generally, with any other weighted average of individual causal effects).

(B) In the special case of a *binary treatment* ($D \in \{0, 1\}$), the condition $\mathbb{C}\text{ov}(D, Y(d)) = 0$ for all $d \in \text{Support}(D)$ (that is, for $d \in \{0, 1\}$), implies $\delta^W = \delta^{\text{T}} = \delta$ (first bullet-point of slide 22, Chapter 4).

(C) An alternative interesting answer relates to the special case of randomized controlled trials. The fundamental assumption behind randomized controlled trials (RCT), or, more precisely, the assumption that is made plausible if the data comes from an RCT is $D \perp\!\!\!\perp (Y(0), Y(1))$ in the case of a binary treatment D, or, more generally $D \perp\!\!\!\perp \{Y(d)\}_{d\in\text{Support}(D)}$; in words, the treatment variable is independent of the potential outcome variables.

This assumption implies in particular that Δ, as a function (the difference) of $Y(d_0+1)$ and $Y(d_0)$, for some d_0 in the support of D, is independent of D.

Then, note that W is a function of D. Therefore, $\Delta \perp\!\!\!\perp D \implies \Delta \perp\!\!\!\perp W$.

Finally, $\Delta \perp\!\!\!\perp W$ implies that a fortiori Δ and W are uncorrelated: the expectation of their product is equal to the product of their expectation. Thus,

$$\delta^W := \mathbb{E}[W\Delta] \overset{\text{if } W \text{ and } \Delta \text{ are uncorrelated}}{=} \mathbb{E}[W]\,\mathbb{E}[\Delta] = 1 \times \mathbb{E}[\Delta] = \mathbb{E}[\Delta] =: \delta.$$

In conclusion, the assumption $D \perp\!\!\!\perp \Delta$ implies $\delta^W = \delta$.

(c) Is the following assertion true or false?
The parameter δ^W does not vary if the distribution P_D of the treatment D changes.

By definition,

$$\delta^W := \mathbb{E}[W\Delta], \text{ with } W := \frac{(D - \mathbb{E}[D])^2}{\mathbb{V}[D]}.$$

Consequently, δ^W depends on the distribution P_D of D through W.

As a consequence, if the distribution of D changes (for instance, if we consider another population of interest), even if the distribution P_Δ of Δ remains unchanged[21], the causal parameter of interest δ^W changes; it is a way to suggest that maybe this parameter δ^W is not so interesting, but the only one we can hope to identify in this case under already strong assumptions (Proposition 2, Chapter 4):

- linear causal effects (LCE) (Equation (lin. effects) of slide 18);
- and absence of selection bias: $\mathbb{C}\text{ov}(D, Y(d)) = 0$ for all d in the support of D.

(d) As a comparison, same question for δ:
Does the parameter δ vary if the distribution P_D of the treatment D changes?

In contrast, by definition, under the assumption (LCE) (Equation (lin. effects) of slide 18):

$$\delta := \mathbb{E}[\Delta] = \mathbb{E}[Y(d_0 + 1) - Y(d_0)],$$

for some d_0 in the support of D; such that $d_0 + 1$ belongs to the support too.

Thus, δ does not depend on the distribution of D: the causal parameter δ (average treatment effect) does not change if the distribution of the treatment D changes; it only depends on the distribution of the individual causal effect Δ.

Bonus question: in the case of a binary treatment D, same question for the parameter δ^{T}. You can find the solution in slide 15 of the document TD6_correction_slides.pdf available on Pamplemousse.

(e) Give the conditions stated in Chapter 4 (Proposition 2) for the limit in probability β_D of the OLS estimator $\widehat{\beta}_D$ of the slope in the simple linear regression of Y on D to be equal to the causal parameter δ^W.

This is directly **Proposition 2 of Chapter 4**.
Under the two following conditions:

- Equation (Lin. effects) of slide 18 ((LCE)); *in words*, the causal effect of D on Y is *linear*;
- $\forall\, d \in \text{Support}(D), \mathbb{C}\text{ov}(D, Y(d)) = 0$; *in words*, the treatment variable is uncorrelated with all the potential outcome variables (no selection bias),

we have $\beta_D = \delta^W$, where β_D is the probability limit of the OLS estimator $\widehat{\beta}_D$ of the slope in the simple linear regression of Y on D. In other words, under those two conditions, the regression of Y on D identifies the causal parameter δ^W.

Remark that it does not identify δ. Absent further conditions that regression only identifies a *weighted* average causal effect δ^W (whose exact interpretation and usefulness for a policymaker might not be evident in practice).

[21] In that sense, we could say that “the treatment effect is the same”; but, be careful, remember that Δ is stochastic, hence “the” refers, in fact, to P_Δ, a probability distribution, not a number.

8 The absence of (conditional) selection

We are in the setting of Section 2 of Chapter 4 with D the treatment variable (a real random variable, not necessarily binary), G the control variables (a vector of real random variables), and, for any $d \in \text{Support}(D)$, $Y(d)$ the potential outcome corresponding to the value d of D ($Y(d)$ is a real random variable).

We consider the following two assumptions:

(i) **the absence of selection**: $\mathbb{C}\text{ov}(Y(d), D) = 0$ for all $d \in \text{Support}(D)$;

(ii) **the absence of conditional selection**: $\mathbb{C}\text{ov}(Y(d), D \mid G) = 0$ for all $d \in \text{Support}(D)$.

This is directly the discussion of slides 25, 26, and 27 of Chapter 4: there is no implication between the two assumptions; in a way, it says the two assumptions are really different, it is another thing to condition with respect to control variables G compared to no conditioning.

The absence of conditional selection is often more credible. The intuition behind this relates to the **critical issue of selection bias** and the question: is D determined/chosen/generated independently (or at least such that there is no correlation, possibly conditional on G) of the potential outcome variables $Y(d)$?

In many cases, notably when the agents can choose their realization of D, it is not the case: it is likely that agents know more than the econometrician and will choose D depending on (knowledge/expectation about) their potential outcome variables (see notably Model 2 for D of TD6: self-selection of the agents into the treatment – see also “Roy model”).

When conditioning by control variables G, we still wonder whether D is allocated randomly/ independently/without correlation with the potential outcomes $Y(d)$, but we ask this question, not over the entire population but *looking only at agents with the “same”*[22] *variables G, that is, similar agents in terms of the control variables G.* **The intuition behind** is that if we have the good or enough controls G, we can make the individuals with the same G so similar regarding the choice/determination of D that, essentially, among that subset of individuals, the realization of the treatment variable D *becomes as if randomly assigned, hence uncorrelated with the potential outcome variables* $\{Y(d)\}_{d \in Support(D)}$.

This is basically the meaning of the second line of Models (Lin. mod. 1) and (Lin. mod. 2) of Chapter 4 (slides 28 and 31):

$$Y(d_0) = \zeta_0 + G'\gamma_0 + \eta, \text{ with } \mathbb{E}[\eta \mid D, G] = 0.$$

Or rather, to say it more precisely, **Models (Lin. mod. 1) and (Lin. mod. 2) imply the absence of conditional selection** (see bullet-point 4, slide 28 of Chapter 4).

Example: repeating (“redoubler”) a school year Imagine D is an indicator of repeating the first year of senior high school (“la classe de seconde au lycée”), and the outcome Y is the average grade on the final exam (“baccalauréat”). Repeating is not something drawn randomly; the decision (the way to generate/determine D) depends on the level of the students through their grades during the first year of senior high school. In particular, we can reasonably expect that

$$\mathbb{E}[Y(0) \mid D = 1] < \mathbb{E}[Y(0) \mid D = 0] = \mathbb{E}[Y \mid D = 0],$$

that is, the average grade among the students who should have repeated ($D = 1$) if, in fact, they had not repeated is lower than the average grade among students who do not repeat ($D = 0$): it is merely

[22] This interpretation, looking at individuals cell by cell defined by the modalities of G, holds exactly if G is discrete (see Proposition 3, slide 27 of Chapter 4 also in this case). When G is continuous, it is not exactly that interpretation cell by cell; yet, the intuition remains similar.

saying that students who repeat tend to have lower grades than those who repeat. Thus, there is a selection bias.

Now, imagine G is a binary indicator equal to 1 if the general grade during the first year of senior high school is between 9.9 and 10.1 and that a student repeats if and only if its general grade is strictly below 10. Conditional on $G = 1$, we can think that being just below or above 10 is as if random and thus we can assume:

$$\mathbb{E}[Y(0) \,|\, D = 1, G = 1] = \mathbb{E}[Y(0) \,|\, D = 0, G = 1]$$

It would be an example of the absence of selection conditional on $G = 1$.

Note: this example aims to convey the idea of the absence of conditional selection on a concrete case; yet, it is a bit different here in the last section of Chapter 4. The previous idea suggests to restrict attention to observations with $G = 1$ (see Regression Discontinuity Design in Econometrics 2 next semester) while in Chapter 4, we would like to find adequate controls G such that we keep all observations and for any modality g of G:

$$\mathbb{E}[Y(0) \,|\, D = 1, G = g] = \mathbb{E}[Y(0) \,|\, D = 0, G = g],$$

which, when D is binary, is equivalent to having $\mathbb{C}\mathrm{ov}(Y(d), D \,|\, G = g) = 0$.

Explanations about the second equation of Models (Lin. mod. 1) and (Lin. mod. 2) We recall that it writes:

$$Y(d_0) = \zeta_0 + G'\gamma_0 + \eta, \text{ with } \mathbb{E}[\eta \,|\, D, G] = 0.$$

First, remember that the equation is written for a given d_0 in the support of D but, assuming linear effects (which is implied by the first line of Models (Lin. mod. 1) and (Lin. mod. 2)), we can choose any d_0 in the support of D. Consequently, the rest of the reasoning holds for any value d_0 in the support of D.

Second, note that when conditioning by D and G, hence, in particular, by G, the only remaining randomness in $Y(d_0)$ is η.

Third, the assumption $\mathbb{E}[\eta \,|\, D, G] = 0$ implies that η is centered $\mathbb{E}[\eta] = 0$ (law of iterated expectations) and also that $\mathbb{E}[\eta \,|\, G] = 0$ (composition of projections)[23].

Combining the second and third remarks yield

$$\begin{aligned}
\mathbb{C}\mathrm{ov}(Y(d_0), D \,|\, G) &= \mathbb{C}\mathrm{ov}(\eta, D \,|\, G) \qquad \text{because only } \eta \text{ is stochastic knowing } G \\
&= \mathbb{E}[D\eta \,|\, G] - \mathbb{E}[D \,|\, G] \times \underbrace{\mathbb{E}[\eta \,|\, G]}_{=0} \qquad \text{by property of covariances (cond. or uncond.)} \\
&= \mathbb{E}[D\eta \,|\, G] \qquad \text{computations + hypothesis } \mathbb{E}[\eta \,|\, D, G] = 0 \implies \mathbb{E}[\eta \,|\, G] = 0 \\
&= \mathbb{E}(\mathbb{E}[D\eta \,|\, D, G] \,|\, G) \qquad \text{by composition of projections} \\
&= \mathbb{E}(D \times \mathbb{E}[\eta \,|\, D, G] \,|\, G) \qquad \text{by property of cond. exp. } (D \text{ as if non-stochastic}) \\
&= \mathbb{E}(D \times 0 \,|\, G) = 0, \qquad \text{by the assumption } \mathbb{E}[\eta \,|\, D, G] = 0
\end{aligned}$$

where the first equality comes from the second remark, the second equality is the definition of the conditional covariance with respect to G (same as the unconditional standard covariance but conditioning all expectations by G), and the fourth equality comes from the composition of projections.

Finally, remember that the reasoning hold for an arbitrary $d_0 \in \mathrm{Support}(D)$. Thus, we do obtain as a consequence of Linear Models 1 or 2,

$$\forall\, d \in \mathrm{Support}(D),\ \mathbb{C}\mathrm{ov}(Y(d), D \,|\, G) = 0, \qquad \text{(NCBPOTCC)}$$

[23] Composition des projections : généralisation de la loi des espérances itérées en un sens. Intuitivement, en voyant l'espérance conditionnelle comme une projection, si on projette plusieurs fois successivement, on se retrouve sur le "plus petit" espace. Exemple : si Y, X_1 et X_2 sont des variables aléatoires (potentiellement des vecteurs aléatoires) telles que les espérances considérées sont bien définies, on a $\mathbb{E}[Y \,|\, X_1] = \mathbb{E}(\mathbb{E}[Y \,|\, X_1, X_2] \,|\, X_1)$, X_1 est une fonction de (X_1, X_2). Plus largement, pour un vecteur aléatoire X, si $g(X)$ est une fonction mesurable de X, on a $\mathbb{E}[Y \,|\, g(X)] = \mathbb{E}\{\mathbb{E}[Y \,|\, X] \,|\, g(X)\}$. Remarque : si $X_1 = 1$, la constante, on retombe sur la loi des espérances itérées : projeter sur l'espace des constantes, c'est juste prendre l'espérance: $\mathbb{E}[Y \,|\, 1] = \mathbb{E}[Y]$.

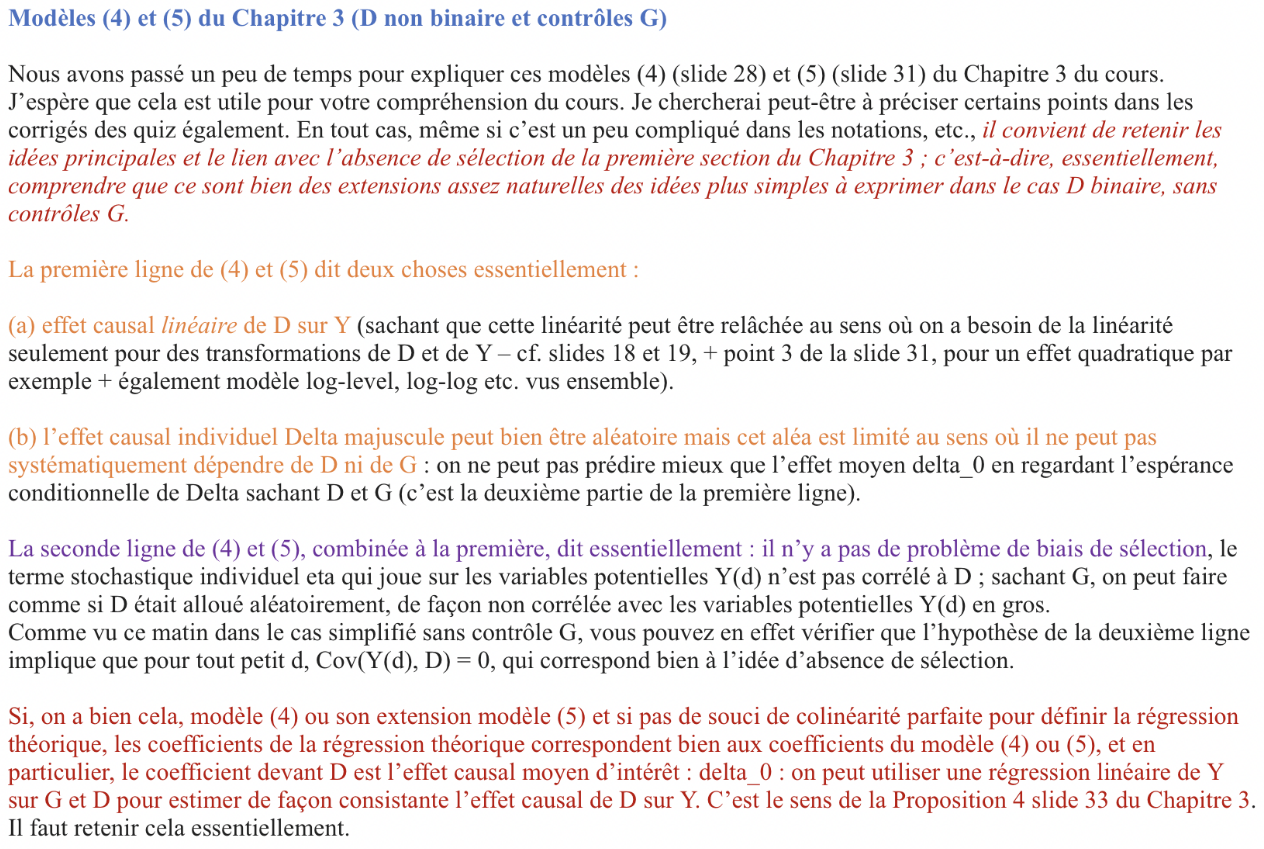
with (NCBPOTCC) standing for Null Correlation Between Potential Outcomes and Treatment Conditional on Controls.

Ci-dessous (Figure 2) quelques autres explications reprises d'un mail envoyé en français sur les modèles (Lin. mod. 1) et (Lin. mod. 2) du Chapitre 4.

What can you say about those two conditions?

1. $\mathbb{C}\text{ov}(Y(d), D) = 0$ is the weakest: it is implied by $\mathbb{C}\text{ov}(Y(d), D \,|\, G) = 0$ – **False**.
2. $\mathbb{C}\text{ov}(Y(d), D \,|\, G) = 0$ is the weakest: it is implied by $\mathbb{C}\text{ov}(Y(d), D) = 0$ – **False**.
3. The two conditions are equivalent when D is binary – **False**, nonsense.
4. $\mathbb{C}\text{ov}(Y(d), D) = 0$ neither implies nor is implied by $\mathbb{C}\text{ov}(Y(d), D \,|\, G) = 0$, but the latter (conditional on G) is often more credible – **True**.
5. $\mathbb{C}\text{ov}(Y(d), D) = 0$ neither implies nor is implied by $\mathbb{C}\text{ov}(Y(d), D \,|\, G) = 0$, but the former (unconditional) is often more credible – **False**.

Figure 2: Quelques explications en français sur les modèles (Lin. mod. 1) et (Lin. mod. 2) du Chapitre 4 (anciennement le Chapitre 3).



Modèles (4) et (5) du Chapitre 3 (D non binaire et contrôles G)

Nous avons passé un peu de temps pour expliquer ces modèles (4) (slide 28) et (5) (slide 31) du Chapitre 3 du cours. J'espère que cela est utile pour votre compréhension du cours. Je chercherai peut-être à préciser certains points dans les corrigés des quiz également. En tout cas, même si c'est un peu compliqué dans les notations, etc., *il convient de retenir les idées principales et le lien avec l'absence de sélection de la première section du Chapitre 3 ; c'est-à-dire, essentiellement, comprendre que ce sont bien des extensions assez naturelles des idées plus simples à exprimer dans le cas D binaire, sans contrôles G.*

La première ligne de (4) et (5) dit deux choses essentiellement :

(a) effet causal *linéaire* de D sur Y (sachant que cette linéarité peut être relâchée au sens où on a besoin de la linéarité seulement pour des transformations de D et de Y – cf. slides 18 et 19, + point 3 de la slide 31, pour un effet quadratique par exemple + également modèle log-level, log-log etc. vus ensemble).

(b) l'effet causal individuel Delta majuscule peut bien être aléatoire mais cet aléa est limité au sens où il ne peut pas systématiquement dépendre de D ni de G : on ne peut pas prédire mieux que l'effet moyen delta_0 en regardant l'espérance conditionnelle de Delta sachant D et G (c'est la deuxième partie de la première ligne).

La seconde ligne de (4) et (5), combinée à la première, dit essentiellement : il n'y a pas de problème de biais de sélection, le terme stochastique individuel eta qui joue sur les variables potentielles Y(d) n'est pas corrélé à D ; sachant G, on peut faire comme si D était alloué aléatoirement, de façon non corrélée avec les variables potentielles Y(d) en gros.
Comme vu ce matin dans le cas simplifié sans contrôle G, vous pouvez en effet vérifier que l'hypothèse de la deuxième ligne implique que pour tout petit d, Cov(Y(d), D) = 0, qui correspond bien à l'idée d'absence de sélection.

Si, on a bien cela, modèle (4) ou son extension modèle (5) et si pas de souci de colinéarité parfaite pour définir la régression théorique, les coefficients de la régression théorique correspondent bien aux coefficients du modèle (4) ou (5), et en particulier, le coefficient devant D est l'effet causal moyen d'intérêt : delta_0 : on peut utiliser une régression linéaire de Y sur G et D pour estimer de façon consistante l'effet causal de D sur Y. C'est le sens de la Proposition 4 slide 33 du Chapitre 3. Il faut retenir cela essentiellement.

9 Short or long regressions in a randomized experiment

We are interested in the causal effect of a training program[24] on jobseekers' income. To do so, we implement a *randomized controlled experiment* (*RCT*) on a representative sample of jobseekers: we

[24] like the NSWD Program we will see in a Problem Set with personalized support to write CVs, make online job searches, prepare for interviews, etc.

randomly provide the training to some of them, then, one year later, we measure their monthly income as well as other individual covariates (age, gender, education, having or not a college degree, etc.)

We contemplate two linear regressions:

(i) a "short" regression of the monthly income Y on the indicator D of receiving the training;

(ii) a "long" regression of the monthly income Y on the indicator D of receiving the training and the indicator C of having a college degree (diplôme post-bac).

In order to *estimate* the average causal effect of the training program, it is preferable to do

Although the exercise about Tennis and Sleep (impact of sleep on tennis performance) in the Problem Sets is not written explicitly in terms of causal effects, the question is related to question 5 of that exercise: you can also look at it.

It is important to note that we are interested in *estimation* only here (as opposed to inference: the precision of the estimation).

The setting is that of the first section of Chapter 4: a binary treatment D equal to 1 in case of effective participation in the training, 0 otherwise. In addition, we assume to be in the set-up of a randomized controlled trial (RCT).

As a consequence, we can assume that D is independent of the potential outcome variables $Y(0)$ and $Y(1)$, but also, more generally, of any other covariates, in particular, here, the measure of education as an indicator C of having a college degree.

Since $D \perp\!\!\!\perp (Y(0), Y(1))$, we have $\beta_D^{\text{S}} = \delta = \delta^{\text{T}}$ (see the result under RCT of slide 12, Chapter 4) where β_D^{S} is the theoretical slope coefficient (that is, the limit in probability of the OLS estimator under the usual moment conditions) in the simple linear regression (SLR) of Y on D (the superscript "S" is added to underscore that it is the theoretical coefficient associated with D in the short, simple linear regression of Y on D).

As a consequence, the "short" regression of Y on D is valid to identify and consistently estimate the average causal effect δ of the training program.

The result also eliminates Answer 4: in the RCT set-up, the average causal effect δ is equal to the average causal effect *on the treated* δ^{T}.

It also explains why Answer 2 is false: there is no omitted variable bias here.

It is crucial to remember that an omitted variable bias requires TWO CONDITIONS (in the simple case of a univariate omitted variable G, but it is also the case more generally – (see slides 35 and 36 of Chapter 4 for the notation, notably point 2 of slide 36)):

(i) the omitted variable G has to be indeed omitted in the sense of "being in the error term", that is, being correlated to the outcome Y, ($\gamma_0 \neq 0$), AND ALSO

(ii) the omitted variable G needs to be correlated with the included variable D ($\lambda_0 \neq 0$).

Here, we can be confident that (i) holds: C belongs to the error term in the short, simple linear regression of Y on D because having or not having a college degree influences income.

However, (ii) does *not* hold because, by assumption, the treatment D is allocated randomly (RCT set-up), hence D is independent, a fortiori uncorrelated (conditional on D), with C.

Although the previous discussions show that the first part of Answer 1 is correct, Answer 1 is false because the justification of Answer 1 is false.

The "included variable bias" discussed at the end of Chapter 4 (slides 37 and 38) relates to a setting where a control variable G is itself influenced by the treatment D. In such cases, including G may induce a bias. We are not in this setting here since the control variable C is already determined ex-ante and is not influenced by participation in the training program.

By elimination, we can conclude that Answer 3 is the correct one.

A positive justification for Answer 3 is the following result, a direct consequence of the "omitted variable bias" formula (Proposition 4, Chapter 1; or rather here, the theoretical result of Proposition 7, Chapter 1) using the notation of Chapter 1: if G and D are uncorrelated, then $\beta_D^{\mathrm{S}} = \beta_D$ (see below for a more precise statement and proof).

Here, it applies to $G = C$ as $D \perp\!\!\!\perp C$ due to the RCT set-up: the OLS estimators of the coefficient of D on the short, simple linear regression of Y on D and on the long multiple linear regression of Y on D and C have the same probability limit and, for large sample sizes, the estimates will tend to be very close to each other.

Therefore, for *estimation*, it is indifferent to use the short or the long regression. Remark that taking into account *inference*, we would prefer the *long* regression (see the solutions of the exercise about Tennis and Sleep, question 5 for details).

1. the "short" regression because the "long" one suffers from the included variable bias

   – **False**.

2. the "long" regression because the "short" one suffers from the omitted variable bias

   – **False**.

3. it is indifferent: we can do the "short" regression or the "long" regression since the estimates of the coefficients of D will tend to be close in the two regressions (more precisely, the estimators of the two regressions converge to the same probability limit)

   – **True**.

4. none of the previous answer: neither the "short" nor the "long" regressions are useful to estimate the average causal effect of the training; they can only estimate the average causal effect *on the treated*

   – **False**.

***Détails (en français) sur le résultat utilisé pour justifier la réponse 3.** *Corollaire de la "formule du biais de variable omise" : identité du coefficient dans les régressions linéaires théoriques simple et multiple lorsque les deux régresseurs sont non corrélés.*

On considère ici les régressions théoriques (projections linéaires). Soient Y, X_1, X_2 trois variables aléatoires réelles ayant les propriétés nécessaires pour réaliser la régression multiple de Y sur X_1 et X_2 (et sous-entendu une constante) – c'est-à-dire $\mathbb{E}[|X_1|^2] < +\infty$, $\mathbb{E}[|X_2|^2] < +\infty$, $\mathbb{E}[XX']$ inversible avec $X := (1, X_1, X_2)'$ – et les régressions simples de Y sur X_1 et X_1 sur X_2 – $\mathbb{V}(X_1) > 0$, $\mathbb{V}(X_2) > 0$.

Si X_1 et X_2 sont non corrélées, alors le coefficient de X_1 dans la régression multiple de Y sur X_1 et X_2 est identique au coefficient de X_1 dans la régression simple de Y sur X_1, et donc égal à $\mathbb{C}\mathrm{ov}(Y, X_1)/\,\mathbb{V}(X_1)$.

Pour relier régression simple et régression multiple, les outils de base sont le théorème de Frisch-Waugh et la formule dite du "biais de variable omise".

Preuve 1. Ici, la proposition 7 du Chapitre 1 donne directement le résultat du lemme puisque le coefficient (noté λ dans le cours) dans la régression de X_2 (l'omise) sur X_1 (l'incluse) est nul car il vaut $\mathbb{C}\mathrm{ov}(X_2, X_1)/\,\mathbb{V}(X_1)$ et on a supposé X_2 et X_1 non corrélées: $\mathbb{C}\mathrm{ov}(X_2, X_1) = 0$.

Preuve 2. Pour s'exercer, on propose une autre démonstration du lemme en utilisant à la place Frisch-Waugh.

D'après le Chapitre 1, on sait que le coefficient de X_1 dans la régression simple de Y sur X_1 est égal à $\mathbb{C}\mathrm{ov}(Y, X_1)/\,\mathbb{V}(X_1)$. De plus, par Frisch-Waugh, on sait que le coefficient de X_1 dans la régression multiple de Y sur X_1 et X_2, qu'on note, par exemple, β_1, est identique au coefficient dans la régression linéaire simple de Y sur le résidu de la régression de X_1 sur X_2.

On écrit donc cette dernière régression. D'après le Chapitre 1, Proposition 5, il existe des réels (non stochastiques, des paramètres) λ_0 et γ_2, et une variable aléatoire réelle R_1 tels que

$$X_1 = \lambda_0 + \gamma_2 X_2 + R_1, \text{ avec } \mathbb{E}(R_1) = \mathbb{E}(X_2 R_1) = 0.$$

De plus, on sait que $\gamma_2 = \mathbb{C}\mathrm{ov}(X_1, X_2)/\,\mathbb{V}(X_2)$. Si on suppose $\mathbb{C}\mathrm{ov}(X_1, X_2) = 0$ comme dans l'énoncé du lemme, alors $\gamma_2 = 0$ et on a donc $X_1 = \lambda_0 + R_1$.

Or, par Frisch-Waugh, comme écrit plus haut, on a :

$$\begin{aligned}
\text{coeff. de } X_1 \text{ dans la régression multiple : } \beta_1 &\overset{\text{Frisch-Waugh}}{=} \text{coeff. dans la régression de } Y \text{ sur } R_1 \\
&= \frac{\mathbb{C}\mathrm{ov}(Y, R_1)}{\mathbb{V}(R_1)} \\
&= \frac{\mathbb{C}\mathrm{ov}(Y, X_1 - \lambda_0)}{\mathbb{V}(X_1 - \lambda_0)} \text{ (si } \mathbb{C}\mathrm{ov}(X_1, X_2) = 0) \\
&= \frac{\mathbb{C}\mathrm{ov}(Y, X_1)}{\mathbb{V}(X_1)} \quad (\lambda_0 \text{ est une constante}) \\
&= \text{coeff. dans la régression de } Y \text{ sur } X_1,
\end{aligned}$$

ce qui donne donc le résultat voulu.

10 Heterogeneous or homogeneous causal effects

We consider the linear causal model (Lin. mod. 1) (slide 28) and its multivariate extension (Lin. mod. 2) (slide 31) of Chapter 4. Remember that m is the dimension of D and of Δ.

We focus here on one of the conditions of those models, namely:

$$\exists\, \delta_0 \in \mathbb{R}^m \text{ (that is, non-stochastic)} : \mathbb{E}[\Delta \,|\, D, G] = \delta_0. \quad (*)$$

Are the following assertions true or false?

(a) Condition $(*)$ allows Δ to be random, namely to vary from one individual to another (case of *heterogeneous* causal effects). – **True**.

Remember (see for instance Equation (Lin. effects) of Chapter 4 in slide 18 or the first line of Models (Lin. mod. 1) and (Lin. mod. 2) of Chapter 4) that Δ is the *individual* linear causal effect of D on Y. A priori, causal effects can vary across individuals: Δ is modeled as a random variable; it is stochastic (case of *heterogeneous* causal effects). To resume the concrete example about repeating a class (see Question 8), this authorizes the effect of repeating a class on the final grade at baccalauréat to be different across students; in particular, it might be positive for some students, but negative for others. Condition $(*)$ does allow Δ to be random.

Nonetheless, it somewhat limits the amount of treatment effect heterogeneity in the sense that the individual causal effect Δ is mean-independent (thus, a fortiori, uncorrelated) with the variables D and G. In other words, the knowledge of D and G for a given individual is useless to predict his or her individual causal effect Δ.

Remark that it would be possible to relax that assumption and have individual causal effects that depend on G by adding interactions terms $D \times G$ in the regression: this is the model at the end of slide 29 of Chapter 4.

(b) If Δ is a degenerate constant random variable, that is, Δ is equal across individuals (case of *homogeneous* causal effects), then condition $(*)$ is necessarily satisfied. – **True**.

The opposite of the *heterogeneous* causal effects set-up is the *homogeneous* effects case: Δ is assumed to be degenerate, non-stochastic, a constant (formally, $\mathbb{V}[\Delta] = 0$): the causal effects is the same for all individuals.

In many settings, this assumption is a strong assumption. For instance, in the case of repeating classes, it assumes the effect is the same for all students. In particular, it is either positive (causally increases the grade on the final exam of high school) or negative (causally decreases the grade), or null for all students.

Models (Lin. mod. 1) and (Lin. mod. 2) of Chapter 4 avoid such a strong assumption, precisely by stating condition $(*)$ in a sense.

Formally, the homogeneous causal effects assumption writes:

$$\exists\, \delta_0 \in \mathbb{R}^m \text{ (that is, non-stochastic)} : \Delta = \delta_0.$$

Such an assumption implies condition $(*)$ since the expectation (be it unconditional or conditional) of a constant is equal to this constant itself:

$$\mathbb{E}[\Delta \mid D, G] \overset{\text{if homogeneous effects}}{=} \mathbb{E}[\delta_0 \mid D, G] = \delta_0.$$

(c) Condition $(*)$ rules out nonlinear effects of D on Y. – **(Rather) false**.

Condition $(*)$ is the second part of the first line of Models (Lin. mod. 1) and (Lin. mod. 1) of Chapter 4, that combines (LCE), reproduced below

$$\boxed{\exists!\, \Delta \in \mathbb{R}^\Omega,\ \exists\, d_0 \in \text{Support}(D) :\ \forall d \in \text{Support}(D),\ Y(d) = Y(d_0) + \Delta(d - d_0)}$$

and the assumption $(*)$

$$\boxed{\exists\, \delta_0 \in \mathbb{R}^m : \mathbb{E}[\Delta \mid D, G] = \delta_0} \tag{IHCE}$$

In that sense, given that condition $(*)$ involves Δ that arises to define the linear effect of D on Y, we could argue condition $(*)$ is linked to the linearity of the effects of D on Y (hence the "(rather)"). Yet, the answer remains "false" since condition $(*)$ is somewhat orthogonal. Indeed, we could have the first part of the assumption, namely linear effects of D on Y, without condition $(*)$, typically by authorizing correlation between Δ and G or D, for instance. In that sense, the condition $(*)$ by itself does not say anything regarding linear or nonlinear effects of D on Y.

It is rather the role of the first part of the first line of the Linear Models 1 and 2 considered in Chapter 4.

At first sight, the linearity of the effects might appear as a strong condition. However, as explained in slide 19 of Chapter 4, the assumption is weaker in so far as the crucial point is the linearity between a *known* transform of $Y(d)$ and a *known* transform of d. Hence, linear models encompass nonlinear effects between the original (before the transforms) variables of interest (see notably models log-level, level-log, log-log, and Question 5 of this quiz).

(d) Condition $(*)$ implies that $\mathbb{E}[\Delta] = \delta_0$ and $\mathbb{E}[\Delta \mid D] = \delta_0$. – **True**.

By the law of iterated expectations,

$$\mathbb{E}[\Delta] = \mathbb{E}(\mathbb{E}[\Delta \mid D, G]) = \mathbb{E}(\delta_0) = \delta_0.$$

Likewise, by the composition of projection (a generalization of the law of iterated expectations – see footnote 23 for further details),

$$\mathbb{E}[\Delta \mid D] = \mathbb{E}(\mathbb{E}[\Delta \mid D, G] \mid D) = \mathbb{E}(\delta_0 \mid D) = \delta_0.$$

Similarly, we also have

$$\mathbb{E}[\Delta \mid G] = \mathbb{E}(\mathbb{E}[\Delta \mid D, G] \mid G) = \mathbb{E}(\delta_0 \mid G) = \delta_0.$$

Remember that, for a given individual, we can never observe more than one potential outcome, which is the one corresponding to the actual outcome: **we only observe the realized outcome $\boldsymbol{Y := Y(D)}$.**

Consequently, absent further assumptions, it is impossible to identify and estimate *individual* causal effects. **That is why we restrict to some features, which we call *causal parameters*, of the distribution P_Δ of Δ.**

When D is binary (first section of Chapter 4), we are often interested in *the average causal effect*:

$$\delta := \mathbb{E}[\Delta],$$

or *the average causal effect on the treated*:

$$\delta^{\mathrm{T}} := \mathbb{E}[\Delta \mid D = 1].$$

Remark that δ^{T} makes sense (the fact of calling it the average causal effect *on the treated*, more precisely) only in the case of a binary D with $D = 1$ means being treated.

When D is not binary (second and following sections of Chapter 4), there is no more any special meaning of $D = 1$ (actually, 1 may even not be in the support of D). Therefore, we do not consider δ^{T} in the case of non-binary treatment D. Rather, we are interested in the average causal effect δ or, possibly, but less frequently in this course, the average causal effect for some specific value d_0 of D: $\mathbb{E}[\Delta \mid D = d_0]$.

Also, we might be interested in *the weighted average causal effect* introduced in Chapter 4, slide 20:

$$\delta^W := \mathbb{E}[W\Delta], \text{ where } W := \frac{(D - \mathbb{E}[D])^2}{\mathbb{V}[D]}.$$

In fact, the interpretation of δ^W might not be so easy. We may be more interested in δ than in δ^W if we had the choice for policy-making concerns (deciding whether or not to implement a given policy/treatment).

Yet, under general heterogeneous effects (“general” meaning without the additional restriction of condition $(*)$), the point is that, essentially, we cannot do better than identify δ^W when we assume the absence of selection bias.

This is the result of Proposition 2, Chapter 4, slide 20: if we assume linear effects of D on Y ((LCE)) and the absence of selection: $\mathbb{C}\mathrm{ov}(D, Y(d)) = 0$ for all d in the support of D, then the OLS estimator of the slope coefficient in the simple linear regression of Y on D converges in probability to δ^W (in other words, the linear regression of Y on D identifies δ^W).

Condition $(*)$ implies that $\delta^W = \delta_0$. Indeed, by the law of iterated expectation, we have

$$\begin{aligned}
\delta^W &:= \mathbb{E}[W\Delta] \qquad \text{(definition of } \delta^W\text{)} \\
&= \mathbb{E}(\mathbb{E}[W\Delta \mid D, G]) \qquad \text{(law of iterated expectations)} \\
&= \mathbb{E}(W\mathbb{E}[\Delta \mid D, G]) \qquad (W \text{ is a function of } D \text{ and linearity of cond. expect.)} \\
&= \mathbb{E}(W\delta_0) \qquad \text{(assumption } (*)\text{)} \\
&= \mathbb{E}[W]\delta_0 \qquad (\delta_0 \text{ is not stochastic and linearity of expect.)} \\
&= \mathbb{E}\left(\frac{(D - \mathbb{E}[D])^2}{\mathbb{V}[D]}\right)\delta_0 \qquad \text{(definition of } W\text{)} \\
&= \frac{\mathbb{E}\left[(D - \mathbb{E}[D])^2\right]}{\mathbb{V}[D]}\delta_0 \qquad \text{(linearity of expect.)} \\
&= \frac{\mathbb{V}[D]}{\mathbb{V}[D]}\delta_0 = \delta_0, \qquad \text{(definition of a variance)}
\end{aligned}$$

details for the third equality: W is a function of D only (and non-stochastic constants); hence, W is as a constant when conditioning by D (and G).

Thus, by limiting the possible heterogeneity of the (unattainable) individual causal effects Δ, condition $(*)$ implies that *there is only a single average causal parameter of interest δ_0, equal to δ, δ^W, or any conditional expectation of Δ by D or G.* In that sense, it is a convenient assumption to restrict the heterogeneity of causal effects and get back to a single average causal parameter δ_0.

11 *Causal linear models and linear conditional expectations

We consider the linear causal model (Lin. mod. 1) (slide 28) or its multivariate extension (Lin. mod. 2) (slide 31) of Chapter 4.

Let us consider the univariate case (Lin. mod. 1) for simplicity.

(a) Given an outcome real random variable $Y \in \mathbb{R}^\Omega$, a univariate treatment $D \in \mathbb{R}^\Omega$, and control variables $G \in (\mathbb{R}^{\dim(G)})^\Omega$, write Linear Model 1 (Lin. mod. 1) formally with quantifiers ($\exists$ and $\forall$).

(Lin. mod. 1) writes

$$\begin{aligned}
&\exists\, \Delta \in \mathbb{R}^\Omega,\ \exists\, \eta \in \mathbb{R}^\Omega,\ \exists\, \delta_0 \in \mathbb{R},\ \exists\, \zeta_0 \in \mathbb{R},\ \exists\, \gamma_0 \in \mathbb{R}^{\dim(G)},\ \exists\, d_0 \in \text{Support}(D):\\
&\qquad \forall\, d \in \text{Support}(D),\ Y(d) = \zeta_0 + G'\gamma_0 + \Delta(d - d_0) + \eta && \text{(LASCE)}\\
&\qquad \mathbb{E}[\eta \,|\, D, G] = 0 && \text{(ACS)}\\
&\qquad \mathbb{E}[\Delta \,|\, D, G] = \delta_0. && \text{(IHCE)}
\end{aligned}$$

General methodology remark: even if written implicitly sometimes, it is important to have in mind the corresponding formal propositions with quantifiers (and their order!) for the different models and results seen in your course of applied mathematics.

In particular, it forces us to be precise about the nature of the objects.

Here, the first line specifies the nature of the objects. In particular, do not confuse:

(i) random variables (Δ and η);

(ii) non-stochastic parameters of interest (δ_0, ζ_0, γ_0, among which δ_0 is of particular interest);

(iii) an arbitrary non-stochastic reference value (d_0) – see the first bullet-point in the discussion of slide 29.

(b) Under (Lin. mod. 1), what can you say for the observed outcome $Y := Y(D)$? That is, write an equation satisfied by Y.

It is important to remark that, for $Y := Y(D)$ the observed outcome, (Lin. mod. 1) implies that

$$Y = \zeta_0 + G'\gamma_0 + \Delta(D - d_0) + \eta, \text{ with } \mathbb{E}[\eta \,|\, D, G] = 0 \text{ and } \mathbb{E}[\Delta \,|\, D, G] = \delta_0.$$

Symbolically, we can replace lowercase d (a non-stochastic free variable) with uppercase D, a real random variable in equation (LASCE).

That comes from the fact that $Y(d) = \zeta_0 + G'\gamma_0 + \Delta(d - d_0) + \eta$ holds for all $d \in \text{Support}(D)$ and reasoning ω by ω, with $\omega \in \Omega$, where $(\Omega, \mathcal{A}, \mathbb{P})$ is the underlying probability space over which all random variables are defined.

(c) Comment each of the assumptions (LASCE), (ACS), and (IHCE) of (Lin. mod. 1).

In particular, try to guess why the related equations are called as such; that is, for which words do the acronyms LASCE, ACS, and IHCE stand? Each letter represents one word, and link words, such as "of", "the", etc., are not represented by a letter. An example: ATE = Average Treatment Effect.

(LASCE) stands for **Linear Additively Separable Causal Effects** and combines the two equations of (Lin. mod. 1) as expressed in slide 28 of Chapter 4.

(LASCE) assumes two crucial things about the causal effect of D on Y: (i) it is a linear effect; (ii) that effect is additively separable from the "effect"[25] of the controls G on the outcome Y.

For (ii), in other words, we consider an effect of D on Y with G held constant; or, rather, more precisely, *the assumption is* that the causal effect of D on Y does not interact with G.

Example. To be concrete, imagine G is an indicator of being a woman, Y is the income, and D is the number of years of education (a classical example seen in class). Furthermore, to simplify the reasoning, let us assume homogeneous causal effects: $\Delta = \delta_0 \in \mathbb{R}$.

(LASCE) says (i) the effect of the number of years of education on income is linear: it is the same effect for each additional year of education (which is arguably unrealistic; hence rather, the use of the logarithm of income as Y – see also Question 5 of this Quiz); (ii) the effect of one additional year of education is the same for women and men. Point (ii) can also be unrealistic; hence the extension of (Lin. mod. 1) with interactions between treatment and controls (see the fourth bullet-point of slide 29). For heterogeneous causal effects, (ii) would be similar, but in distribution.

The key consequence of (LASCE) is that, *for one given individual*, there is a *unique* real random variable, Δ, that summarizes the causal effect of D on Y. It would not be the case otherwise:

- Without linearity (i), we would need a function linking d to $Y(d)$.
- Without additive separability (ii), but keeping (i) as a start, we would need one linear effect that we could denote $\Delta_g \in \mathbb{R}^\Omega$ by possible value $g \in \text{Support}(G)$.
- Without (i) nor (ii), we would need a function linking the couple (d, g) to $Y(d, g)$.

Granted, without homogeneous causal effects, that unique number Δ remains a random variable that is allowed to vary across individuals. Yet, it remains possible to summarize the causal effect of D on Y through a single number per individual.

(ACS) stands for **Absence of Conditional Selection**. As explained in Question 8, (ACS), combined with (LASCE), implies a Null Correlation Between Potential Outcomes and Treatment Conditional on Controls (NCBPOTCC):

$$\forall\, d \in \text{Support}(D),\ \mathbb{C}\text{ov}(Y(d), D \,|\, G) = 0,$$

that is, the absence of conditional selection. Remember that, in contrast, the absence of selection formally writes with an *unconditional* covariance: $\forall\, d \in \text{Support}(D),\ \mathbb{C}\text{ov}(Y(d), D) = 0$.

See the solutions to Question 8 for intuition about the meaning of that crucial hypothesis to identify causal effects: *essentially, conditional on G, that is, among units having the same*[26] *G, the determination of the treatment D can be thought of as if it was drawn randomly in an experiment.*

(IHCE) stands for **Idiosyncratic Heterogeneity of Causal Effects**. That assumption is discussed in Question 10.

It preserves heterogeneous causal effects (as opposed to homogeneous effects – identical for all individuals), but *restricts the heterogeneity* to be idiosyncratic: it does not depend on the controls G nor the treatment D, only on some "random" (in ordinary language) unobserved individual features. In that sense, the heterogeneity cannot be anything; the heterogeneity is not unrestricted: there are only "random" individual variations around δ_0.

As explained in Question 10-(d), a decisive consequence of (IHCE) is that there is only a *single average*[27] causal effect parameter: δ_0. In particular, $\delta = \delta^W = \delta_0$.

[25] I put "effect" into parentheses as it is not here a causal effect of G on Y since we do not model potential outcomes depending on the values of G: $Y(d, g)$, for $d \in \text{Support}(D)$ and $g \in \text{Support}(G)$ – see further comments below in question (d). Remember that *causal effects and potential outcomes are defined jointly*: causal effects are differences of potential outcomes.

[26] More or less in the case of continuous controls (that is, when G admits a density with respect to Lebesgue measure).

[27] Parameters other than averages/expectations can also be considered (see advanced econometric courses in ENSAE's third year).

An equivalent[28] formulation of (IHCE), $\mathbb{E}[\Delta \mid D, G] = \delta_0$ is

$$\exists \xi \in \mathbb{R}^{\Omega} : \Delta = \delta_0 + \xi \text{ with } \mathbb{E}[\xi \mid D, G] = 0. \qquad \text{(equiv. form. IHCE)}$$

With the explicit individual index i, (IHCE) implies that, for all individual i, $\Delta_i = \delta_0 + \xi_i$. In other words, the individual causal effect can be decomposed additively[29] into a common population component $\delta_0 \in \mathbb{R}$ (non-stochastic) and an individual fluctuation $\xi_i \in \mathbb{R}^{\Omega}$ (stochastic) that is "random"[30], idiosyncratic, in the sense that ξ is mean-independent of D and G: the fluctuations cannot depend on the treatment nor the control variables.

Thus, the three assumptions of Linear Model 1 can be grouped in two types.

(LASCE) **and** (IHCE) **could be called jointly "simplification hypotheses".**

At the level of *one* individual/unit, (LASCE) simplifies the modeling of the causal effect of the treatment D on the outcome Y by assuming an effect additively separable from the "effect" of G and, in addition, linear. As a consequence, a single real random variable Δ, of the dimension of a number 1×1, contains all the information of the individual causal effect of D on Y.

There is some structure and real restrictions in this assumption (LASCE). However, it is important to understand that, *without deep modifications of the model (although with more caution for its interpretation), it is possible to relax those restrictions*:

(for AS) by adding interactions between treatment and controls;

(for L) by considering transformations of the initial treatment and outcome (notably logarithms) or introducing a multivariate treatment (formally, it corresponds to Linear Model 2) with D, D^2 to model a quadratic effect, for instance, etc.

In that sense, assumption (LASCE) can be deemed less constraining or strong than others (see the discussion of (ACS) below).

At the level of the entire population of interest, that is, across individuals, (IHCE) also simplifies "the" (it is a distribution P_{Δ}, not a parameter) causal effect of D on Y. It is not homogeneous, but the magnitude/amount or, rather, type of heterogeneity is restricted.

Indeed, as explained above, under (IHCE), Δ is additively decomposed as $\Delta = \delta_0 + \xi$, with ξ "almost independent" (only, mean-independent) of D and G. Thus, at the level of the population, we can focus on the single average causal parameter $\delta_0 \in \mathbb{R}$.

Similarly as (LASCE)*, but to a lesser extent,* (IHCE) *is a "simplification" hypothesis that can be partly relaxed without substantial impact on the model.*

First, as for the AS part of (LASCE), it is possible to add interactions between treatment and controls to authorize a modelled heterogeneity of causal effects: the causal effects depend on the value of D and G. The interaction is introduced at the individual level.

However, at the population level, relaxing (IHCE) is less innocuous. If the heterogeneity of causal treatment is fully unrestricted, essentially, the best a linear regression can retrieve is a weighted average causal effect with weights linked to the dispersion of the treatment (conditionally on controls). Without controls, Proposition 2 formalizes that result: under linear effects and the absence of selection, the theoretical slope coefficient in the simple linear regression of Y on D, $\beta_D = \delta^W := \mathbb{E}[W\Delta]$. Proposition 3 extends that result *with controls* and the absence of *conditional* selection.

In terms of mathematical results, there is no genuine loss in the sense that we obtain a similar identification result: a linear regression identifies a causal parameter in the sense of an average (possibly

[28] *It is important to understand and know that equivalence*: see the solutions for Question 13 of Quiz 2 for a detailed explanation (with other variables, but the reasoning is exactly the same). Also, remark that the assumption (IHCE) actually writes: $\mathbb{E}[\Delta \mid D, G] = \mathbb{E}[\Delta]$, δ_0 is only a notation for the unconditional expectation of Δ: $\delta_0 := \mathbb{E}[\Delta]$.

[29] That *additive* separation is linked behind to the modeling with conditional expectations (and the linearity of expectations).

[30] "random" in the acceptation of ordinary language; formally, in the mathematical language, ξ_i is also random in the sense of stochastic: it is a real random variable (that is, a measurable function from the underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$ into $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ – with $\mathcal{B}(E)$ denoting the Borel sigma-algebra of a topological space $(E, \mathcal{T})$).

weighted) of the individual causal effect Δ. However, in applied mathematics, we are also concerned with the practical implication of a result. We remark that, *in the perspective of a decision-maker or social planner contemplating whether or not implementing a treatment or a public policy, the parameter δ^W is not relevant in general.*[31] Indeed, the weights, although at least all non-negative, may be hard to interpret. On comparison, the average treatment effect (ATE), denoted δ, is a natural benchmark for an egalitarian (attributing the same weight to anyone) and utilitarian (focusing on the total or average welfare) social planner.[32] *In that sense, the ATE is a more precious causal parameter than other weighted averages of Δ.* A key interest of (IHCE), and not easily relaxed, is that the ATE is identified. Indeed, (IHCE) implies $\delta = \delta_0$ and, under the hypothesis of Lin. mod. 1 and the usual moment conditions, a linear regression identifies δ_0 (Proposition 4).

As opposed to those "simplification hypotheses", (ACS) **can be called an "identification" hypothesis.**[33]

The absence of conditional selection is indeed crucial for a linear regression of the outcome Y on the treatment D and the controls G (plus a constant) to identify the causal parameter of interest δ_0.

Question (f) below explains how the hypothesis (ACS) *can be partly relaxed: instead of the mean-independence, a null correlation is enough*:

$$\mathbb{E}[\eta] = 0, \quad \mathbb{E}[D\eta] = 0, \quad \mathbb{E}[G\eta] = 0 = 0_{\mathbb{R}^{\dim(G)}}. \qquad \text{(ACS relaxed)}$$

Remark that the first equality $\mathbb{E}[\eta] = 0$ of (ACS relaxed) is without loss of generality since there is a constant ζ_0 in the model. Consequently, as η is centered, $\mathbb{C}\text{ov}(D, \eta) = \mathbb{E}[D\eta]$ and $\mathbb{C}\text{ov}(G, \eta) = \mathbb{E}[G\eta]$. (ACS) implies (ACS relaxed). The latter assumption is thus weaker. Mathematically, we move one step away from independence within the scale: independence, mean-independence, and null correlation.

However, in concrete meaning, it remains the same idea of the absence of conditional selection: *conditional on the controls G, the treatment D is determined "randomly", "independently" (in the meaning or ordinary language) with respect to the potential outcomes $\{Y(d)\}_{d \in Support(D)}$*, as-if it was drawn at random in an experiment – formally, there are uncorrelated conditional on G; see (NCBPOTCC).

As tentatively explained several times in this document (see notably the examples in Question 3), THIS IDEA OF ABSENCE OF CONDITIONAL SELECTION (OR, WITHOUT CONTROLS, ABSENCE OF SELECTION) IS ABSOLUTELY KEY FOR A LINEAR REGRESSION TO IDENTIFY A CAUSAL PARAMETER, namely, some average, possibly weighted, of individual causal effects. Such an assumption cannot be relaxed anyway: it is the fundamental identification assumption that always needs to be argued in real settings. In that sense, it differs from the "simplification" hypotheses (LASCE) and (IHCE).

***(d)** Why do we *not* consider $Y(d, g)$, for $d \in \text{Support}(D) \subseteq \mathbb{R}$ and $g \in \text{Support}(G) \subseteq \mathbb{R}^{\dim(G)}$?

First, specify what the notation $Y(d, g)$ means. A related question that may help (or not): in the models studied in Chapter 4, is it possible to consider $Y(d)$ with d multivariate? If so, quote the associated model.

Let us start with the easy part of the question.

First, as a natural extension of the definition of $Y(d)$, for any $d \in \text{Support}(D)$ and $g \in \text{Support}(G)$, $Y(d, g)$ *can be defined as the potential outcome a given individual* (remember that index i are omitted thanks to i.i.d-ness) *would have obtained had she got the value d for the treatment and g for the controls.*

In that modeling, *the observed outcome* Y is defined as $Y := Y(D, G)$, that is, the potential outcome that corresponds to the actual value D and G of the treatment and the controls respectively; it is thus no more potential but "realized", and we assume we observe it.

[31] Besides, we neglect the fact that, in practice, we can, at best, consistently estimate and perform inference on δ^W, which remains *unknown*.

[32] The structural summary of the course (available on Pamplemousse: document `ResumeStructurel.pdf`) elaborates on this issue.

[33] *Note that the distinction mainly serves presentation purposes and is not entirely exact.* For instance, as just explained, (IHCE) could also be named an identification hypothesis insofar as it enables the identification of the ATE by imposing $\delta = \delta_0$.

Second, the models studied in Chapter 4 indeed introduce $Y(d)$ *with* d *multivariate. It is Lin. mod. 2 in slide 31.* The dimension of D is denoted m, that is, $D \in (\mathbb{R}^m)^\Omega$, and the product is replaced by a scalar product with the transpose of Δ:

$$Y(d) = Y(d_0) + \Delta'(d - d_0),$$

with

$$d = (d_1, \dots, d_m)' \in \mathbb{R}^m, \ d_0 = (d_{01}, \dots, d_{0m})' \in \mathbb{R}^m, \quad \text{and} \quad \Delta = (\Delta_1, \dots, \Delta_m)' \in (\mathbb{R}^m)^\Omega,$$

the individual causal effects associated with each component of the multivariate treatment D on the outcome Y. That is, for any $j \in \{1, \dots, m\}$, Δ_j is the individual causal effect of D_j, the j-th component of D, on the outcome.

Now, let us try to answer the difficult part.

In short, the answer is that we do *not* model a causal effect of G on Y in the course. Remember that *causal effects and potential outcomes are defined jointly: causal effects are differences of potential outcomes.* Thus, the previous sentence means essentially the same as the proposition asked by the question: we do not model potential outcomes $Y(d, g)$, for $d \in \text{Support}(D)$ and $g \in \text{Support}(G)$, which depend on the values of the controls G.

I try to explain the idea behind that with two elements.

(First element) If we are interested in several treatments, that is, a *multivariate treatment:* $m := \dim(D) > 1$ (see slide 30 for an example), we can do so by considering Linear Model 2 with a multivariate treatment and potential outcomes with multiple arguments:

$$Y(d) = Y(d_1, \dots, d_m) \in (\mathbb{R}^m)^\Omega.$$

In other words, it is not because we restrict to univariate treatment that we do not consider $Y(d, g)$. The setup does authorize the analysis of several treatments.

(Second element) It is rather that we do not consider $Y(d, g)$ because we do not consider D and G on the same level. Mathematically, D and G are both random variables (*same nature*); yet, they have *different*, say, *statuses*.

The starting point is that, given outcome Y and treatment(s) D (univariate or multivariate), we are interested in “the” causal effect of D on Y. At the level of a given individual or unit, that “the” makes sense due to (LASCE).

Said differently, D and Y exist before the controls G.

At the beginning, we are interested in the effect of D on Y holding other variables fixed in a way, or, rather, we consider there is a meaning of considering the causal effect of D on Y among some specified population of interest. If there is the absence of (unconditional) selection, for instance, if we are able to organize a randomized experiment (with perfect compliance) where D is allocated at random, we could stop there: there is no need to introduce control variables G.

As explained and made more visible in the structural summary of the course (available in Pamplemousse), control variables are one of *several possible strategies* to identify causal effects. *We could do other things*: instrumental variables (Chapter 5), difference-in-differences (Econometrics 2), regression discontinuity designs (RDD) (Econometrics 2), matching (Microeconometric Evaluation of Public Policies in 3A), etc.

At that stage, we are only interested in the causal effect of D on Y, and there is no controls G. A priori, the individual causal effects Δ are heterogeneous, and we are trying to recover some average of the effects Δ.

Perhaps, for instance, men and women have quite different causal effects; that is, the distribution of Δ conditional on gender = man differs from the distribution of Δ conditional on gender = woman. But, *at that stage, it is not an issue* here for the target of an average causal effect, say δ or δ^{T}. Again, imagine

D is binary and we can organize an RCT with perfect compliance, then $\mathbb{E}[Y \mid D=1]-\mathbb{E}[Y \mid D=0]=\delta$, and we identify (and can estimate consistently) the average treatment effect (ATE) $\delta := \mathbb{E}[\Delta]$.

Then, we decide to use the control variables strategy to identify a causal effect.

As further explained in the structural summary (see also Question 8 of this Quiz), the main idea is to add relevant controls G so as to obtain the absence of conditional selection: among the individuals with the same controls, the determination of D can be considered as-if "random", in particular, uncorrelated with the potential outcomes.

Formally, we need Linear Model 1 (univariate) or 2 (multivariate treatment) to apply Proposition 4 in order to identify δ. Thus, we need:

(LASCE) the causal effect of D on Y is linear and additively separable from the controls G,

(IHCE) the heterogeneity of the causal effect is restricted to be "idiosyncratic", mean-independent of D and G, and

(ACS) the absence of conditional selection.

We could relax (IHCE) at the cost of identifying only some (complicated) weighted average causal effect (in the same vein as δ^W without controls); yet, that parameter may not have a real interest for a public policy maker (compared to the ATE δ).

The control variables G thus appear in a *second step*: not in the definition of the target (what we are trying to do: identify a causal effect of D on Y), but on the way to achieve it (identification strategy through adequate control variables G). Hopefully, that should explain why, with the notation of the course, there is no sense of considering $Y(d, g)$.

Final remark. We resume the hypothetical example where the causal effects depend on gender. Imagine that gender is part of the control variables G; that is, we need to control by gender to have credibly the absence of conditional selection.

In that case, in the basic (without interaction between treatment and controls) linear models 1 or 2, the assumption (IHCE) does *not* hold. As a consequence, a linear regression of Y on D and G does *not* identify the average causal effect δ. In such a case, the issue can be fixed by adding interactions of the treatment D with the controls G, or some components thereof only, as explained at the end of slide 29.

A consequence of that reasoning: distinction between selection bias and omitted variable bias The previous reasoning may also be a way to explain the difference between selection bias and omitted variable bias (often shortcut as OVB).

First, remember Chapter 1, Propositions 4 (empirical, for OLS estimators) and 7 (theoretical, for parameters): they state an *algebraic, mechanical relation between the coefficients of a "short" regression and a "long" regression.* At that stage, in Chapter 1, there was nothing more as we were not trying to identify any specific parameter, causal or not.[34]

In contrast, Chapter 4, Proposition 5 says that *if* (Lin. mod. 2) holds, then, the simple linear regression of Y on D identifies, through its theoretical slope coefficient denoted β_D^{S},

$$\beta_D^{\mathrm{S}} = \delta_0 + \lambda_0' \gamma_0, \tag{OVB}$$

where $\lambda_0 = (\lambda_{01}, \ldots, \lambda_{0\,\dim(G)})' \in \mathbb{R}^{\dim(G)}$ is the vector of the theoretical slope coefficients associated with D in the different simple linear regressions of G_j (the omitted variables) on D (the single variable of the "short" regression), for $j \in \{1, \ldots, \dim(G)\}$.

Consequently, as long as the product $\lambda_0' \gamma_0 \neq 0$, the short linear regression does *not* identify the average causal effect of interest δ_0. We say that there is an omitted variable bias insofar as the bias can

[34] Rather, we were trying to understand linear regressions and Ordinary Least Squares (OLS) estimators. That relation can be seen as a "property" of those objects, presented in order to develop their understanding.

be suppressed (meaning that we could identify and consistently estimate the targeted parameter δ_0) if we include the omitted correct controls G, those such that linear model 1 or 2 holds.

Remark that (OVB) is a direct corollary of Chapter 1, Proposition 7 and Chapter 4, Proposition 4 since, with the notation of Chapter 1, under (Lin. mod. 2), we have $\beta_D = \delta_0$ and $\beta_G = \gamma_0$, which yields the result.

As tentatively explained above, D and Y pre-exist G in some way. *Similarly, we can say that, if any, the selection bias pre-exists the omitted variable bias.*

If, for *a given data-generating process (d.g.p)*, there is a selection bias in the sense that we cannot identify the causal parameter of interest, it says we need to find another identification strategy. At that time, there is no OVB.

Now, if we can think of and measure controls G such that Linear model 1 (or 2) holds (identification strategy with adequate controls – see Section 3.1 of the structural summary), then we can identify the ATE $\delta := \mathbb{E}[\Delta]$ (let us imagine it is our targeted parameter here). In that case, for the specific d.g.p considered and the adequate controls G, we have (OVB).

As far as I understand, we cannot go really further *without specifying a quantity that we call selection bias (SB)*. The course does so for a binary treatment (Support$(D) = \{0, 1\}$)

$$\text{With binary } D\text{, selection bias: } B := \mathbb{E}[Y(0) \,|\, D = 1] - \mathbb{E}[Y(0) \,|\, D = 0]. \qquad \text{(SB – binary } D)$$

Also, that definition is in fact *specific to the targeted parameter* $\delta^{\mathrm{T}} := \mathbb{E}[\Delta \,|\, D = 1]$ in the sense that the key result of Chapter 4, Proposition 1 says that

$$\beta_D^{\mathrm{S}} = \delta^{\mathrm{T}} \iff B = 0 \iff \mathbb{C}\mathrm{ov}(D, Y(0)) = 0. \tag{3}$$

With hindsight, this is not surprising as a "bias" needs to be defined relative to a targeted parameter and an estimator or its limit in probability. In (3), the limit in probability of the estimator (under classical moment conditions) is β_D^{S}, the slope coefficient in the theoretical simple linear regression of Y on D; the target is δ^{T}, so that the bias is the identified quantity minus the target: $B = \beta_D^{\mathrm{S}} - \delta^{\mathrm{T}}$.

If we are interested in δ, (3) is not enough!

A sufficient condition for identifying the ATE, δ by a simple linear regression is

$$\big(\mathbb{C}\mathrm{ov}(D, Y(0)) = 0 \text{ and } \mathbb{C}\mathrm{ov}(D, Y(1)) = 0\big) \implies \beta_D^{\mathrm{S}} = \delta. \tag{4}$$

Indeed, as $D \in \{0, 1\}^{\Omega}$ is binary, for any random variable A, $\mathbb{C}\mathrm{ov}(D, A) = 0$ (null-correlation) is equivalent to $\mathbb{E}[A \,|\, D] = \mathbb{E}[A]$ (mean-independence of A from D). Therefore, the sufficient condition of the implication (4) (the left-hand side) is equivalent to

$$\mathbb{E}[Y(0) \,|\, D] = \mathbb{E}[Y(0)] \text{ and } \mathbb{E}[Y(1) \,|\, D] = \mathbb{E}[Y(1)].$$

In particular, we thus have

$$\mathbb{E}[Y(0) \,|\, D = 0] = \mathbb{E}[Y(0)] \text{ and } \mathbb{E}[Y(1) \,|\, D = 1] = \mathbb{E}[Y(1)].$$

Finally, we thus obtain

$$\begin{aligned}
\beta_D^{\mathrm{S}} &= \mathbb{E}[Y \,|\, D = 1] - \mathbb{E}[Y \,|\, D = 0] && \text{property of SLR with a binary regressor} \\
&= \mathbb{E}[Y(1) \,|\, D = 1] - \mathbb{E}[Y(0) \,|\, D = 0] && Y := Y(D) \text{ and use of the conditioning events} \\
&= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] && \textit{if } \mathbb{C}\mathrm{ov}(D, Y(0)) = \mathbb{C}\mathrm{ov}(D, Y(1)) = 0 \\
&= \mathbb{E}[Y(1) - Y(0)] && \text{linearity of the expectation} \\
&= \mathbb{E}[\Delta] =: \delta && \text{by definition of } \Delta.
\end{aligned}$$

In the case of interest in δ, the selection bias is therefore not the same as for δ^{T}: we can have $B = 0$ (which is equivalent to $\mathbb{C}\mathrm{ov}(D, Y(0)) = 0$, but $\beta_D^{\mathrm{S}} \neq \delta$.

Rather, by analogy, we would define the following quantity as the selection bias with respect to δ:

$$B_\delta := \beta_D^{\mathrm{S}} - \delta = (\mathbb{E}[Y(1) \mid D = 1] - \mathbb{E}[Y(1)]) - (\mathbb{E}[Y(0) \mid D = 0] - \mathbb{E}[Y(0)])$$

the difference between the limit in probability of the estimator of the target (considering here a simple linear regression of the outcome Y on the treatment D) and the targeted parameter of interest.

A necessary and sufficient condition (NSC) for β_D^{S} be equal to δ (in other words, for a simple linear regression to identify the average treatment or causal effect) is $B_\delta = 0$.

That condition is less readable than the equivalent for δ^{T} (Equation (3)); yet, it is as it is. A more readable condition, but only sufficient, is expressed in Equation (4): a null covariance between both D and $Y(0)$, and D and $Y(1)$.

Remark and reminder: we could rename

$$\begin{aligned} B = B_{\delta^{\mathrm{T}}} := \beta_D^{\mathrm{S}} - \delta^{\mathrm{T}} &= (\mathbb{E}[Y(1) \mid D = 1] - \mathbb{E}[Y(1) \mid D = 1]) - (\mathbb{E}[Y(0) \mid D = 0] - \mathbb{E}[Y(0) \mid D = 1]) \\ &= \mathbb{E}[Y(0) \mid D = 1] - \mathbb{E}[Y(0) \mid D = 0]. \end{aligned}$$

Summary of our findings from the previous reasoning and computations. In general, the notion of the "bias" of an estimator (here asymptotic, not to be confused with the finite-sample bias studied in Statistics 1) is defined relatively to two elements:

1. an estimation procedure and related conditions giving: an estimator and its limit in probability; let us denote it generically β;

2. a targeted parameter of interest, like δ, or δ^{T} with a binary treatment, or δ^W with a non-binary treatment; let us denote it generically θ.

The bias is then defined simply as $\beta - \theta$: what we can identified minus the target.

Now, in the course, the term "selection bias" is linked to the estimation through a simple linear regression of the outcome Y on the treatment D; that is: $\beta = \beta_D^{\mathrm{S}}$ (the theoretical coefficient of the slope).

The notion remains dependent on the targeted parameter: δ (for binary or non-binary treatments), δ^{T} when the treatment is binary, δ^W for non-binary treatments?

In addition, in a way, there happen to be *two different things* behind the term "selection" or "selection bias":

(quant.) A quantity that, by definition, is equal to $B_\theta := \beta_D^{\mathrm{S}} - \theta$ (the identified object minus the target). We could call it the "selection *bias*". By construction, remark that a NSC to identify the target θ is therefore $B_\theta = 0$.

(cond.) A condition that we could name "the absence of selection" that can be (again!) different things.

- The previous NSC: $B_\theta = 0$.

  Example : $\theta = \delta^{\mathrm{T}}$ when the treatment is binary. In this setting, the absence of selection bias or simply, rather, *the absence of selection* means $B_{\delta^{\mathrm{T}}} = 0 \iff \mathbb{C}\mathrm{ov}(D, Y(0)) = 0$.
- Or a readable, meaningful, simple *sufficient* condition that implies $B_\theta = 0$

  Example 1: $\theta = \delta$ when the treatment is binary. The absence of selection (bias) as a condition would rather mean $\mathbb{C}\mathrm{ov}(D, Y(0)) = \mathbb{C}\mathrm{ov}(D, Y(1)) = 0$ (and not, the quantity $B_\delta = 0$, which is only a consequence of that condition).

  Example 2: $\theta = \delta^W$ when the treatment is non-binary (assuming linear causal effects as in the course). Here, the absence of selection bias typically means: for all $d \in \mathrm{Support}(D)$, $\mathbb{C}\mathrm{ov}(D, Y(d)) = 0$. This is so because, by Proposition 2 (where β_D^{S} is just denoted β_D), with linear effects, it is a sufficient condition to have $\beta_D^{\mathrm{S}} = \delta^W$.

  Remark that, in the binary case, $\mathbb{C}\mathrm{ov}(D, Y(0)) = \mathbb{C}\mathrm{ov}(D, Y(1)) = 0$ can equivalently also be stated as $\forall\, d \in \mathrm{Support}(D), \mathbb{C}\mathrm{ov}(D, Y(d)) = 0$. In this case, $\delta = \delta^W = \delta^{\mathrm{T}}$. As a consequence, without further precision, the terminology "absence of selection" usually means that condition[35]

$$\forall\, d \in \text{Support}(D), \mathbb{C}\text{ov}(D, Y(d)) = 0.$$

Back to the comparison between SB and OVB. As already said, the selection bias pre-exists the OVB.

Given only a treatment D and an outcome Y, the selection bias or the absence of selection is connected with a simple linear regression of Y on D: does it identify some causal parameter?

The OVB arises in a second step conceptually when we choose the identification strategy (typically because there is a selection bias precisely and a simple linear regression is not enough!) of adding adequate controls.

If there are controls G such that (Lin. mod. 1) holds, then $\delta = \delta^W = \delta_0$. As explained above, the associated selection bias would be defined as β^{S}_D minus the target δ_0: $B_{\delta_0} = \beta^{\text{S}}_D - \delta_0$.

Given (OVB), it thus coincides with the omitted variable bias: $\lambda_0'\gamma_0$.

The (OVB) can also be seen as a guide to select the adequate controls G: any variable that is *both* correlated to the treatment and to the potential outcome should be introduced as controls.

That paragraph, §"A consequence of that reasoning: distinction between selection bias and omitted variable bias", is finally quite long. Below is a short summary in French[36]: *Bias de sélection et biais de variable omise. Explication: biais de sélection, plus général, avant de réfléchir à la stratégie d'identification. Biais de variable omise: dans le cas où stratégie d'identification avec des contrôles adéquats. La formule du biais de variable omise permet aussi de réfléchir à quels contrôles adéquats sont requis pour avoir l'absence de sélection conditionnelle*

(e) What is the implication of (Lin. mod. 1) for the conditional expectation of $\mathbb{E}[Y \mid D, G]$, or, equivalently, of $\mathbb{E}[Y \mid X]$ with $X := (1, D, G')'$?
Hint: look at the title of this question and use your result to question (b).

(Lin. mod. 1) implies that the conditional expectation of $\mathbb{E}[Y \mid D, G]$ is linear in D and G. As a consequence, the theoretical linear projection (also called theoretical linear regression) coincides with the conditional expectation. Remember that, *in general, there is no reason that the linear projection be equal to the conditional expectation.* It is thus a rather strong assumption implied by (Lin. mod. 1).

We use (equiv. form. IHCE) and (Lin. mod. 1) to write, for all $d \in \text{Support}(D)$,

$$\begin{aligned} Y(d) &= \zeta_0 + G'\gamma_0 + (\delta_0 + \xi)(d - d_0) + \eta \\ &= \underbrace{\zeta_0 - \delta_0 d_0}_{=:\,\alpha_0 \in \mathbb{R}} + G'\gamma_0 \;+\; \delta_0 d \;+\; [\eta + \xi(d - d_0)] \\ &= \alpha_0 \;+\; G'\gamma_0 \;+\; \delta_0 d \;+\; [\eta + \xi(d - d_0)]. \end{aligned}$$

Consequently, as explained in question (b), for the observed outcome $Y := Y(D)$, we have

$$\begin{aligned} Y &= \alpha_0 \;+\; G'\gamma_0 \;+\; \delta_0 D + \underbrace{\eta + \xi(D - d_0)}_{=:\,\nu \in \mathbb{R}^\Omega} \\ &= \alpha_0 \;+\; G'\gamma_0 \;+\; \delta_0 D \;+\; \nu. \end{aligned}$$

Furthermore, ν is mean-independent of (D, G). Indeed, by properties of the conditional expectation (remember that $d_0 \in \text{Support}(D) \subseteq \mathbb{R}$ is non-stochastic, and that, conditional on D, D behaves as-if non-stochastic too), we have

$$\mathbb{E}[\nu \mid D, G] = \mathbb{E}[\eta \mid D, G] + \mathbb{E}[\xi \mid D, G](D - d_0).$$

[35] See also Question 8 of that quiz.

[36] The memory notes from which I elaborate the paragraph; in case it could be useful too.

Now, on the one hand, (ACS) is exactly $\mathbb{E}[\eta \mid D, G] = \mathbb{E}[\eta] = 0$ (the second equality is without loss of generality as there is an intercept ζ_0 in the model). On the other, (IHCE) says that $\mathbb{E}[\xi \mid D, G] = 0$. Thus, we obtain

$$\mathbb{E}[\nu \mid D, G] = 0 + 0(D - d_0) = 0.$$

Note that the previous equality is an equality among random variables: $\mathbb{E}[\nu \mid D, G] \in \mathbb{R}^{\Omega}$ is a degenerate null random variable (technically, the equality a.s will be enough).

Eventually, (Lin. mod. 1) yields

$$Y = \alpha_0 + G'\gamma_0 + \delta_0 D + \nu, \text{ with } \mathbb{E}[\nu \mid D, G] = 0. \tag{5}$$

As a direct consequence (using the linearity of conditional expectation),

$$\mathbb{E}[Y \mid D, G] = \alpha_0 + G'\gamma_0 + \delta_0 D,$$

and is thus linear in D and G.

By the way, that reasoning is also a proof of Proposition 4 of Chapter 4.

The proof seen in the course is very similar: it is the same (for a multivariate treatment in the course), in fact, simply with the initial formulation of (IHCE), using Δ (instead of $\Delta = \delta_0 + \xi$ from (equiv. form. IHCE)).

Indeed, (5) directly says that the linear projection of Y on $X = (1, D, G')'$ is $\alpha_0 + G'\gamma_0 + \delta_0 D$, because, *if* the conditional expectation is linear, we know it coincides with the “linear conditional expectation”, that is, with the linear projection or theoretical linear regression.

Another way to see it is the following. By the law of iterated expectations and the composition of projections, we have the implication (but, in general, not the converse!)

$$\mathbb{E}[\nu \mid D, G] = 0 \implies \big(\mathbb{E}[\nu] = 0,\, \mathbb{E}[D\nu] = 0,\, \mathbb{E}[G\nu] = 0\big).$$

Therefore, (5) implies (6)

$$Y = \alpha_0 + G'\gamma_0 + \delta_0 D + \nu, \text{ with } \mathbb{E}[\nu] = 0,\, \mathbb{E}[D\nu] = 0,\, \mathbb{E}[G\nu] = 0, \tag{6}$$

which is precisely the writing of the theoretical linear regression of Y on $X = (1, D, G')'$.

Assuming the usual moment conditions, the linear projection being unique (remember, as always, Chapter 1, Proposition 5), $\beta_0 \in \mathbb{R}^{\dim(X)}$, the theoretical coefficient in the linear regression of Y on X, satisfies

$$\beta_0 = \begin{pmatrix} \alpha_0 \\ \gamma_0 \\ \delta_0 \end{pmatrix} = \begin{pmatrix} \zeta_0 - \delta_0 d_0 \\ \gamma_0 \\ \delta_0 \end{pmatrix},$$

which is the result of Proposition 4, in the univariate treatment case – the reasoning would be exactly the same with a multivariate treatment $D \in (\mathbb{R}^m)^{\Omega}$. *In words*, under (Lin. mod. 1), the linear regression of Y on D and G (and a constant) identifies δ_0, and thus the average treatment effect $\delta := \mathbb{E}[\Delta] = \delta_0$ (by (IHCE)).

(f) Is that condition (the one of question (e)) on $\mathbb{E}[Y \mid D, G]$ necessary to identify the causal parameter of interest δ_0? If not, propose an alternative to (Lin. mod. 1) that does not imply the condition of question (e) on $\mathbb{E}[Y \mid D, G]$.

The answer is negative: the linearity of the conditional expectation $\mathbb{E}[Y \mid D, G]$ is *not* necessary to identify δ_0.

Indeed, for a linear regression of Y on D and G to identify δ_0, a sufficient (and necessary – by the uniqueness of the projection) condition is the representation (6), which is both the “causal” (the parameter δ_0 intervenes) and the “simple projection” representation (they coincide, and this is exactly why the regression can identify δ_0). The linearity of the conditional expectation (5) is not required.

Another way to justify the answer is “constructive”, by proposing a variant of Lin. mod. 1 with the relaxed assumptions and checking that, under these hypotheses, a linear regression of Y on D and G identifies δ_0.

(Relaxed lin. mod. 1) writes

$$\begin{aligned}
&\exists\, \Delta \in \mathbb{R}^{\Omega},\, \exists\, \eta \in \mathbb{R}^{\Omega},\, \exists\, \delta_0 \in \mathbb{R},\, \exists\, \zeta_0 \in \mathbb{R},\, \exists\, \gamma_0 \in \mathbb{R}^{\dim(G)},\, \exists\, d_0 \in \text{Support}(D) : \\
&\qquad \forall\, d \in \text{Support}(D),\; Y(d) = \zeta_0 + G'\gamma_0 + \Delta(d - d_0) + \eta && \text{(LASCE)}\\
&\qquad \mathbb{E}[\eta] = 0, \quad \mathbb{E}[D\eta] = 0, \quad \mathbb{E}[G\eta] = 0 = 0_{\mathbb{R}^{\dim(G)}}, && \text{(ACS relaxed)}\\
&\qquad \mathbb{E}[\Delta \,|\, D, G] = \delta_0. && \text{(IHCE)}
\end{aligned}$$

The only modification relative to (Lin. mod. 1) is that (ACS) is replaced by (ACS relaxed): the unobserved heterogeneity / the error term affecting the potential outcome, η, is only (instead of mean-independent) uncorrelated with the treatment D and the controls G. The assumption is weaker. In particular, (Relaxed lin. mod. 1) does not imply a linear conditional expectation of Y given D and G.

Nonetheless, those assumptions are sufficient to identify δ_0.

Indeed, under (Relaxed lin. mod. 1), as previously, we obtain using (equiv. form. IHCE), for all $d \in \text{Support}(D)$,

$$Y(d) = \alpha_0 \,+\, G'\gamma_0 \,+\, \delta_0 d \,+\, [\eta + \xi(d - d_0)],$$

and thus, for the observed outcome $Y := Y(D)$, with $\nu := \eta + \xi(D - d_0)$,

$$Y = \alpha_0 \,+\, G'\gamma_0 \,+\, \delta_0 D \,+\, \nu. \tag{7}$$

Now, since $\mathbb{E}[\xi \,|\, D, G] = 0$, by the law of iterated expectations, we have $\mathbb{E}[\xi \times f(D, G)] = 0$ for any measurable function f. In particular, we have (where the symbol “0” can denote zeros of several dimension, be careful)

$$\mathbb{E}[\xi] = \mathbb{E}[\xi D] = \mathbb{E}[\xi G] = \mathbb{E}[\xi GD] = \mathbb{E}[\xi D^2] = 0. \tag{8}$$

Consequently,

$$\begin{aligned}
\mathbb{E}[\nu] &= \mathbb{E}[\eta + \xi(D - d_0)]\\
&= \mathbb{E}[\eta] + \mathbb{E}[\xi D] - \mathbb{E}[\xi] d_0\\
&= 0 + 0 - 0 = 0;\\
\mathbb{E}[G\nu] &= \mathbb{E}[G(\eta + \xi(D - d_0))]\\
&= \mathbb{E}[G\eta] + \mathbb{E}[\xi GD] - \mathbb{E}[G\xi] d_0\\
&= 0 + 0 - 0 = 0;\\
\mathbb{E}[D\nu] &= \mathbb{E}[D(\eta + \xi(D - d_0))]\\
&= \mathbb{E}[D\eta] + \mathbb{E}[\xi D^2] - \mathbb{E}[D\xi] d_0\\
&= 0 + 0 - 0 = 0;
\end{aligned}$$

where, for each of the three computations, the first equality comes from the definition of ν, the second from the linearity of expectations (and computations), and the third equality from (8) (in blue) and (ACS relaxed) (in red).

We thus have $\mathbb{E}[\nu] = \mathbb{E}[G\nu] = \mathbb{E}[D\nu] = 0$, so that (7) is the theoretical linear projection of Y on $X = (1, G', D)'$. Under the classical moment conditions, that projection being unique, by identification, we have that the theoretical coefficient β_0 of that regression satisfies

$$\beta_0 = \begin{pmatrix} \alpha_0 \\ \gamma_0 \\ \delta_0 \end{pmatrix} = \begin{pmatrix} \zeta_0 - \delta_0 d_0 \\ \gamma_0 \\ \delta_0 \end{pmatrix}.$$

In particular, the causal parameter of interest δ_0, which, in particular, is equal to the ATE, $\delta := \mathbb{E}[\Delta]$, is identified under (Relaxed lin. mod. 1).

(e and f bis) The same questions as (e) and (f) in a model without control variables G. You can start by writing the corresponding model (formally, with quantifiers, as in question (a)).

There is nothing really new in this question, but it provides the opportunity (i) to connect with the linear model without controls considered in Chapter 5, (ii) to make formal links of the assumptions of Lin. mod. 1 with the absence of selection expressed as $\forall\, d \in \text{Support}(D), \mathbb{C}\text{ov}(D, Y(d)) = 0$.

With some advance, Lin. model 1 of Chapter 5.

Given the primitives of an outcome Y, a treatment D, and considering the potential outcomes $\{Y(d)\}_{d \in \text{Support}(D)}$, we consider Lin. model 1 *of Chapter 5*:

$$\exists\, \delta_0 \in \mathbb{R}, \exists\, d_0 \in \text{Support}(D) \subseteq \mathbb{R} : \forall\, d \in \text{Support}(D),\ Y(d) - Y(d_0) = \delta_0(d - d_0). \tag{HLCE}$$

(HLCE) stands for Homogeneous Linear Causal Effect. Indeed, as in Chapter 4, it assumes a linear causal effect of D on Y, plus, for simplicity, it assumes homogeneous causal effects: $\Delta = \delta_0$ almost surely. Also as in Chapter 4, d_0 is only a reference point: if the relation holds for a given d_0, it also holds for any other reference point $d_1 \in \text{Support}(D)$.

As explained in Question 10, an important consequence as regards identification is that the average treatment or causal effect (ATE) satisfies $\delta := \mathbb{E}[\Delta] = \delta_0$.

As explained in Chapter 5, slide 20, (HLCE) is equivalent to

$$\exists\, (\zeta_0, \delta_0) \in \mathbb{R}^2, \exists\, \eta \in \mathbb{R}^\Omega : \forall\, d \in \text{Support}(D),\ Y(d) = \zeta_0 + \delta_0 d + \eta,\ \text{with}\ \mathbb{E}[\eta] = 0. \tag{HLCE bis}$$

The proof of that equivalence is the following:

- add and subtract the quantity $\mathbb{E}[Y(d_0)]$ (we implicitly assume that Y admit an expectation);
- define $\zeta_0 := \mathbb{E}[Y(d_0)] - \delta_0 d_0$ and $\eta := Y(d_0) - \mathbb{E}[Y(d_0)]$ (or in a reverse way to go from the second to the first formulation).

We obtain, going, for instance, from (HLCE) to (HLCE bis),

$$\begin{aligned} Y(d) &= Y(d_0) + \delta_0 d - \delta_0 d_0 \\ &= Y(d_0) + \delta_0 d - \delta_0 d_0 + (\mathbb{E}[Y(d_0)] - \mathbb{E}[Y(d_0)]) \\ &= \underbrace{\mathbb{E}[Y(d_0)] - \delta_0 d_0}_{=:\zeta_0} + \delta_0 d + \underbrace{Y(d_0) - \mathbb{E}[Y(d_0)]}_{=:\eta}. \end{aligned}$$

Now, it remains to notice that $\mathbb{E}[\eta] = 0$ because η is the real random variable $Y(d_0)$ re-centered (by subtracting its expectation).

As always, it is crucial to be aware of the nature (notably stochastic or not) of the different objects. Notably, in the right-hand side of

$$Y(d) = \zeta_0 + \delta_0 d + \eta,$$

remark that *the only stochastic object is* η. Consequently, for any other random variable A, we have, for any $d \in \text{Support}(D)$

$$\mathbb{C}\text{ov}(A, Y(d)) = \mathbb{C}\text{ov}(A, \eta) = \mathbb{E}[A\eta],$$

since η is centered for the second equality. In particular, for $A = D$, we have

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

This holds for any d (a free, non-stochastic variable) in the support of D (a real random variable).

In conclusion, **under** (HLCE)**, the hypothesis of the absence of selection has the following equivalent formulations**:

$$\big(\forall\, d \in \text{Support}(D), \mathbb{C}\text{ov}(D, Y(d)) = 0\big) \iff \mathbb{E}[D\eta] = 0 \iff \mathbb{C}\text{ov}(D, \eta) = 0.$$

Back to Chapter 4: Lin. mod. 1 without controls.

Compared to Chapter 5, in Chapter 4, we do not assume homogeneous causal effects but restrict the heterogeneity to be idiosyncratic in the sense of (IHCE).

The reasoning about the linearity of the conditional expectation and the fact that it is not a necessary condition to identify the average causal effect work exactly as in questions (e) and (f) with controls.

Therefore, let us write directly the relaxed linear model 1 without controls.

With primitives outcome Y and treatment D, (Relaxed lin. mod. 1 w/o G) writes

$$\exists\,\Delta \in \mathbb{R}^{\Omega},\ \exists\,\eta \in \mathbb{R}^{\Omega},\ \exists\,\zeta_0 \in \mathbb{R}, \exists\,d_0 \in \text{Support}(D):$$

$$\forall\,d \in \text{Support}(D),\ Y(d) = \zeta_0 + \Delta(d-d_0) + \eta \qquad \text{(LCE)}$$

$$\underbrace{\mathbb{E}[\eta]=0}_{\text{w.l.o.g since there is a constant } \zeta_0} \quad , \quad \mathbb{E}[D\eta]=0, \qquad \text{(AS)}$$

$$\mathbb{E}[\Delta \mid D] = \mathbb{E}[\Delta] \overset{\text{noted}}{=} \delta_0, \qquad \text{(IHCE w/o } G\text{)}$$

where (LCE) stands for Linear Causal Effect (not supposed homogeneous) and (AS) for Absence of Selection (unconditional now, since there are no controls G).

As in question (e), using (equiv. form. IHCE), we obtain

$$\begin{aligned} Y(d) &= \zeta_0 + (\delta_0+\xi)(d-d_0) + \eta \\ &= \underbrace{\zeta_0 - \delta_0 d_0}_{=:\,\alpha_0 \in \mathbb{R}} + \delta_0 d + [\eta + \xi(d-d_0)] \\ &= \alpha_0 \ + \delta_0 d + [\eta + \xi(d-d_0)]. \end{aligned}$$

In the right-hand-side of that equality, the only stochastic part is $\nu := \eta + \xi(d-d_0)$. Therefore, we have (and more generally replacing D by any random variable A), for any $d \in \text{Support}(D)$,

$$\mathbb{C}\text{ov}(D, Y(d)) = \mathbb{C}\text{ov}(D, \nu). \qquad (9)$$

Now, using the same reasoning as in question (f), since $\mathbb{E}[\xi \mid D] = 0$ under (IHCE w/o G), ξ has zero expectation and, furthermore, $\mathbb{E}[D\xi(d-d_0)] = \mathbb{E}[D\xi](d-d_0) = 0$. Consequently,

$$\mathbb{C}\text{ov}(D, \nu) = \mathbb{E}[D\nu] = \mathbb{E}[D\eta]. \qquad (10)$$

Finally, combining (9) and (10), we obtain that, under (Relaxed lin. mod. 1 w/o G), more precisely, under the two other assumptions, (LCE) and (IHCE w/o G), of the model rather, the absence of selection writes

$$\boxed{\big(\forall\,d \in \text{Support}(D), \mathbb{C}\text{ov}(D, Y(d)) = 0\big) \iff \mathbb{E}[D\eta] = 0 \iff \mathbb{C}\text{ov}(D,\eta) = 0.} \qquad \text{(equiv. AS)}$$

In word, when interested in the causal effect of a treatment D on an outcome Y, if we assume linearity of the effect (assumption (LCE)) and restrict its heterogeneity to be mean-independent of D (assumption (IHCE w/o G)), then the absence of (unconditional) selection is equivalent to a null correlation between the realized treatment D and the additive error term η affecting the potential outcomes.

Consequence: explanations about the definition of endogeneity / exogeneity.

The previous result (equiv. AS) also shed light on the definition of the *exogeneity* (and of its opposite, *endogeneity*) of some variable A, with respect to a given linear causal model as stated in (LCE), regarding its link with the absence of selection: it is the same notion essentially.

Indeed, a variable A is said to be *exogenous* (respectively *endogenous*) with respect to the causal model (LCE) (without controls thus, as of now) when $\mathbb{C}\text{ov}(A,\eta) = 0$ (resp. $\mathbb{C}\text{ov}(A,\eta) \neq 0$), that is when there is no selection, absence of selection (resp. when there is a selection bias).

Example: Chapter 5, slide 23, the exogeneity of the instrument Z is defined as $\mathbb{C}\text{ov}(Z, Y(d_0)) = 0$ (implicitly, for any $d_0 \in \text{Support}(D)$ since, as already recalled, the reference point d_0 is arbitrary). Note that it is the same definition by the fact that η is the only stochastic element in $Y(d_0)$, so the condition is indeed equivalent to $\mathbb{C}\text{ov}(Z, \eta) = 0$.

Remark: those definitions are specific to econometrics and, within that field, *only makes sense relative to a given causal model!*. There exist other acceptations of the words "endogenous" and "exogenous" in other fields: macroeconomics, for instance, but more generally[37].

And with controls G?

We can wonder whether a result similar to (equiv. AS) holds with control variables.

Remember that we have seen two models with control variables, given the following primitives: a treatment D (univariate here, but could be extended without issue to a multivariate treatment), an outcome Y, control variables G.

(Lin. mod. 1) of Chapter 4, slide 28:

$$\exists\, \Delta \in \mathbb{R}^\Omega,\ \exists\, \eta \in \mathbb{R}^\Omega,\ \exists\, \zeta_0 \in \mathbb{R},\ \exists\, \gamma_0 \in \mathbb{R}^{\dim(G)},\ \exists\, d_0 \in \text{Support}(D):$$

$$\forall\, d \in \text{Support}(D),\ Y(d) = \zeta_0 + G'\gamma_0 + \Delta(d - d_0) + \eta \qquad \text{(LASCE)}$$

$$\boxed{\mathbb{E}[\eta \,|\, D, G] = 0} \qquad \text{(strong ACS)}$$

$$\mathbb{E}[\Delta \,|\, D, G] = \mathbb{E}[\Delta] \overset{\text{noted}}{=} \delta_0, \qquad \text{(IHCE)}$$

where we have simply[38] change the labeling of the equation (ACS) by (strong ACS), emphasizing that it is a strong version of the Absence of Conditional Selection (ACS) insofar as we have seen that a weaker assumption is sufficient to identify the causal parameter of interest δ_0: that was the relaxed linear model 1.

(Relaxed lin. mod. 1) introduced in question (f):

$$\exists\, \Delta \in \mathbb{R}^\Omega,\ \exists\, \eta \in \mathbb{R}^\Omega,\ \exists\, \zeta_0 \in \mathbb{R},\ \exists\, \gamma_0 \in \mathbb{R}^{\dim(G)},\ \exists\, d_0 \in \text{Support}(D):$$

$$\forall\, d \in \text{Support}(D),\ Y(d) = \zeta_0 + G'\gamma_0 + \Delta(d - d_0) + \eta \qquad \text{(LASCE)}$$

$$\boxed{\underbrace{\mathbb{E}[\eta] = 0}_{\text{w.l.o.g since there is a constant } \zeta_0},\quad \mathbb{E}[D\eta] = 0,\quad \mathbb{E}[G\eta] = 0 = 0_{\mathbb{R}^{\dim(G)}}}, \qquad \text{(weak ACS)}$$

$$\mathbb{E}[\Delta \,|\, D, G] = \mathbb{E}[\Delta] \overset{\text{noted}}{=} \delta_0, \qquad \text{(IHCE)}$$

where, similarly, we have simply relabelled (ACS relaxed) as (weak ACS). Remember that the names are coherent in the sense that we have (immediately using the law of iterated expectations)

$$\text{(strong ACS)} \implies \text{(weak ACS)}.$$

We also remember and label what we have called so far, for instance in Question 8, "the absence of conditional selection" (and noted in Equation (NCBPOTCC): Null Correlation Between Potential Outcomes and Treatment Conditional on Controls, but it was a complicated name):

$$\boxed{\forall\, d \in \text{Support}(D),\ \mathbb{C}\text{ov}(Y(d), D \,|\, G) = 0} \qquad \text{(Cov. ACS)}$$

with the label (Cov. ACS) to refer to a formulation of the Absence of Conditional Selection (ACS) written with covariances (Cov.)

[37] Example: for Hegelians (for instance, Marx and Engels, when applied to economy, politics, and social life), a genuine solution to a problem is necessarily *endogenous*.

[38] And in addition, introduce δ_0 only as a notation rather than as an object. Indeed, as already explained, the substantial claim in (IHCE) is that Δ is mean-independent of (D, G), that is, by definition, that the conditional expectation of Δ knowing D and G, $\mathbb{E}[D \,|\, D, G] \in \mathbb{R}^\Omega$ (a stochastic object, a real random variable), is equal to the unconditional expectation, $\mathbb{E}[\Delta] \in \mathbb{R}$ (a non-stochastic real number – when the treatment is univariate).

We wonder what are the possible links between (weak ACS), (strong ACS), and (Cov. ACS).

Using (equiv. form. IHCE), (LASCE), and the bilinearity of conditional covariances, it is possible to show (good exercise to check) that, for any $d \in \text{Support}(D)$,

$$\mathbb{C}\text{ov}(D, Y(d) \,|\, G) = \mathbb{C}\text{ov}(D, \eta \,|\, G) + \mathbb{C}\text{ov}(D, \xi \,|\, G)\,(d - d_0).$$

Now, as for the unconditional covariance, the conditional variance is equal to the expectation of the product minus the product of expectations, but with *conditional* expectations. Thus,

$$\begin{aligned}\mathbb{C}\text{ov}(D, \xi \,|\, G) &= \mathbb{E}[D\xi \,|\, G] - \mathbb{E}[D \,|\, G]\,\mathbb{E}[\xi \,|\, G] \\ &= 0 - \mathbb{E}[D \,|\, G] \times 0,\end{aligned}$$

by the composition of projections (remember footnote 23 details) and $\mathbb{E}[\xi \,|\, D, G] = 0$ (see (equiv. form. IHCE)). Therefore, we obtain, still for any $d \in \text{Support}(D)$ and under (LASCE) and (IHCE),

$$\mathbb{C}\text{ov}(D, Y(d) \,|\, G) = \mathbb{E}[D\eta \,|\, G] - \mathbb{E}[D \,|\, G]\,\mathbb{E}[\eta \,|\, G]. \tag{11}$$

From (11), again by composition of projections, it is easy to see that $\mathbb{E}[\eta \,|\, D, G] = 0$ implies $\mathbb{E}[D\eta \,|\, G] = 0$ and $\mathbb{E}[\eta \,|\, G] = 0$. Therefore, we have the first result of our investigation:

$$\big[(\text{LASCE}) \wedge (\text{IHCE})\big] \implies \big[(\text{strong ACS}) \implies (\text{Cov. ACS})\big].$$

This shall not be a surprise[39] since it is a result of the course (Chapter 4, slide 28) that Linear Model 1 implies the absence of conditional selection formulated as (Cov. ACS). The previous reasoning is thus an alternative demonstration of that result.

Also, remember that, without any assumption (no need to locate ourselves in Linear Model 1), we have the implication

$$(\text{strong ACS}) \implies (\text{weak ACS}).$$

Hence, it remains to investigate the links between (Cov. ACS) and (weak ACS). We do so under the assumptions (LASCE) and (IHCE), that is, placing within the setting of Lin. mod. 1. Remember that, under (LASCE) and (IHCE),

$$\forall\, d \in \text{Support}(D),\, \mathbb{C}\text{ov}(D, Y(d) \,|\, G) = \mathbb{C}\text{ov}(D, \eta \,|\, G)\,. \tag{12}$$

(12) is a way to see that (weak ACS) does not imply (Cov. ACS). The key intuition behind that result is the following: *conditional moment conditions (a conditional expectation or covariance equal to zero, for instance) are (far) stronger than unconditional moment conditions*. This is again due to the law of iterated expectations or its generalization (composition of projections): essentially, *conditional* moment conditions imply the corresponding *unconditional* moment conditions, but the converse is generally false.

Now, the final question is: does (Cov. ACS) imply (weak ACS)? The short answer is negative. Remember (11), the tricky point is that, due to the difference, we could have $\mathbb{C}\text{ov}(D, Y(d) \,|\, G) = 0$, but $\mathbb{E}[D\eta \,|\, G] \neq 0$ and $\mathbb{E}[\eta \,|\, G] \neq 0$: in this case, it is necessary and sufficient for that to have:

$$\mathbb{E}[D \,|\, G] = \frac{\mathbb{E}[D\eta \,|\, G]}{\mathbb{E}[\eta \,|\, G]}. \tag{13}$$

Now, remember that we are in applied mathematics, not pure mathematics. The goal is less doing an exhaustive, full-of-generality analysis than applying rigorous mathematical tools to analyze real situations. In that perspective, the point is that, arguably, (13) has essentially no real-life meaning in the sense that it could happen by chance but is totally specific to particular distributions of (D, G, η). Thus, to move forward in our analysis[40], it makes sense to assume

$$\big(\forall\, d \in \text{Support}(D), \mathbb{C}\text{ov}(D, Y(d) \,|\, G) = 0\big) \iff \big(\mathbb{E}[D\eta \,|\, G] = 0 \wedge \mathbb{E}[\eta \,|\, G] = 0\big). \tag{14}$$

[39] Combined with that logical result, for three propositions p, q, and r, the two following expressions are tautologically equivalent (denoted with $\equiv$): $p \implies (q \implies r) \equiv (p \wedge q) \implies r$, where $\wedge$ denotes the AND logical boolean operator.

[40] Which, by the way, thinking of the previous sentence about applied mathematics, might be a bit too mathematical... The objective is to make clearer the different modeling of the important concrete absence of conditional selection (see Question 3 for examples!)

Note that the universal quantifier is not important here. Under (LASCE) and (IHCE), due to the equality (11), assumption (14) is equivalent to

$$\mathbb{E}[D\eta \,|\, G] - \mathbb{E}[D \,|\, G]\,\mathbb{E}[\eta \,|\, G] = 0 \iff \big(\mathbb{E}[D\eta \,|\, G] = 0 \wedge \mathbb{E}[\eta \,|\, G] = 0\big).$$

The "if" part ($\Longleftarrow$) is evident. The assumption is on the "only if" ($\Longrightarrow$) part: it rules out very specific settings where we have (13).

Under that assumption, we eventually obtain an answer to our question: (Cov. ACS) implies (weak ACS)! Indeed, by the law of iterated expectation, it is easy to see that if the right-hand side of (14) holds, then

$$\mathbb{E}[D\eta] = \mathbb{E}\left(\underbrace{\mathbb{E}[D\eta \,|\, G]}_{=0}\right) = \mathbb{E}(0) = 0,$$

$$\mathbb{E}[G\eta] = \mathbb{E}(\mathbb{E}[G\eta \,|\, G]) = \mathbb{E}\left(G \times \underbrace{\mathbb{E}[\eta \,|\, G]}_{=0}\right) = \mathbb{E}(G \times 0) = 0,$$

$$\mathbb{E}[\eta] = \mathbb{E}\left(\underbrace{\mathbb{E}[\eta \,|\, G]}_{=0}\right) = \mathbb{E}(0) = 0.$$

We can state our final formal result on this investigation of the different formulations of the absence of conditional selection:

$$\big[(\text{LASCE}) \wedge (\text{IHCE}) \wedge (14)\big] \implies \big[(\text{Cov. ACS}) \implies (\text{weak ACS})\big].$$

***(g)** Discuss if and why the condition of question (e) on $\mathbb{E}[Y \,|\, D, G]$ might be interesting nonetheless.

All in all, we have seen that:

1. Linear model 1 with (strong ACS) implies the absence of conditional selection (Cov. ACS), and also implies the linearity of the conditional expectation $\mathbb{E}[Y \,|\, D, G]$.

2. Linear model 1 allows to identify (and consistently estimate through the linear regression of Y on D, G, and a constant, by OLS) the average causal parameter of interest, δ_0, also equal to $\delta := \mathbb{E}[\Delta]$ (a likely meaningful parameter for an egalitarian decision-maker).

3. Linear model 1 also implies Relaxed linear model 1 where (strong ACS) (written with conditional expectations) is replaced by (weak ACS) (that involves only unconditional quantities: the error term η that intervenes in the model with potential outcomes $Y(d)$ is uncorrelated with both D and G – and centered).

4. Relaxed linear model 1 does not imply the linearity of the conditional expectation $\mathbb{E}[Y \,|\, D, G]$. It is sufficient to identify δ_0 as it is the condition that enables the causal linear representation to coincide with the theoretical linear projection.

I give only an elliptical and somewhat vague answer; I am sorry for that – perhaps there will be a more satisfactory answer next year.

Basically, the idea is that (strong ACS) (that, under (LASCE) and (IHCE), is equivalent to the linearity of $\mathbb{E}[Y \,|\, D, G]$) might be interesting nonetheless because we *may be interested in more than just identifying causal parameters*: also counterfactuals analyses (what would be Y if we change the distribution of the treatment D) to inform public policy decisions.

Update December 2024: in fact, there might be a better justification for a stronger form of the Absence of Conditional Selection. Essentially, the point is that although the assumptions with null correlation, mean-independence, or independence are not mathematically equivalent, they essentially say the same thing: treatment can be considered as-if randomized (sometimes, we say "as good as random"). Those assumptions cannot be formally tested because we do not observe potential outcomes but only the observed outcome $Y := Y(D)$. They thus have to be argued on a case-by-case basis. Often, it would be weird to be able to convince someone of the absence of correlation between potential outcomes and treatment, but not mean-independence, for instance. Therefore, it might not be a real issue to have a stronger form of Absence of Conditional Selection.

Update January 2025: see the complements to quiz 4 in Pamplemousse.