

Quiz 1 – Chapter 1: the fundamentals of linear regressions

(Lucas Girard) – This version: 14 October 2023

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 learning¹.

Some words about the quiz. As always henceforth and absent contrary indication, the notation used follows that of the course's slides. *Beyond notations, try to be always 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 $\hat{\beta}$? Likewise, be careful about the dimension of the objects (vectors, matrices, numbers) in computations.

In a preamble, Question 1 is a more general question about different notions of (in)dependence between random variables; it will be useful in econometrics and, more generally, in statistics and probability theory.

Questions 2 to 11 are rather basic questions about Chapter 1. Questions 2, 3, and 4 deal with “empirical” objects like estimators (first sections of Chapter 1), while Questions 7, 8, 9, and 10 are more about the asymptotic properties of OLS estimators and thus concern theoretical non-stochastic objects (last section of Chapter 1). Questions 5, 6, and 11 deal with the links between simple (short) and multiple (long) linear regressions.

Question 12 is an important question about marginal effects. It presents the notion in general and makes you aware that the case where the components of X are not functionally dependent is the exception rather than the rule! In the solutions, I add some words about level-level, log-level, level-log, or log-log linear regressions since it also concerns the interpretation of coefficients.

Finally, Question 13 is a bit aside from the course's material by itself (hence the asterisk symbol) and, besides, is more related to Chapter 0. Nonetheless, it considers crucial questions in actual data analyses about the representativity of a sample, and I encourage you always to keep those interrogations in mind. In the solutions, I take the opportunity to present the notion of partial identification. It is outside the course, yet a simple and powerful idea to have in mind when analyzing data.

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

1 Dependence between random variables

Let ε and X be two real random variables with finite variance (that is, belonging to L^2).

(a) Verify that $\text{Cov}(X, \varepsilon)$ is well defined, in the sense that $\text{Cov}(X, \varepsilon) < +\infty$.

Hint: Cauchy-Schwarz.

By hypothesis, ε and X have a finite variance, hence a finite first-order moment. Furthermore, the covariance between two random variables is equal to the expectation of the product minus the product of the expectations. Therefore, the covariance between ε and X is well defined (in the sense of not being infinite) as long as the variable $X\varepsilon$ admits a finite first-order moment. By Cauchy-Schwarz (applied here to $|X|$ and $|\varepsilon|$; see, for instance, in Chapter 1, slide 32), this is indeed the case

$$|\mathbb{E}[|X\varepsilon|]| = \mathbb{E}[|X\varepsilon|] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[\varepsilon^2]} < +\infty,$$

and

$$\text{Cov}(X, \varepsilon) = \mathbb{E}[X\varepsilon] - \mathbb{E}[X]\mathbb{E}[\varepsilon]. \quad (1)$$

¹See “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.

(b) Show that

$$\mathbb{E}[\varepsilon | X] = \mathbb{E}[\varepsilon] \implies \text{Cov}(X, \varepsilon) = 0,$$

that is, in words, ε mean-independent of X implies that ε and X are uncorrelated.

Hint: Law of Iterated Expectations.

By the **Law of Iterated Expectations**, we have

$$\mathbb{E}[X\varepsilon] = \mathbb{E}[\mathbb{E}[X\varepsilon | X]] = \mathbb{E}[X \mathbb{E}[\varepsilon | X]] = \mathbb{E}[X \mathbb{E}[\varepsilon]] = \mathbb{E}[X] \mathbb{E}[\varepsilon],$$

where the last inequality uses the linearity of the unconditional expectation and the previous one the hypothesis. Combined with (1), this yields the result.

We thus have the following implications regarding to what extent any two random variables X and ε are (in)dependent :

$$X \perp\!\!\!\perp \varepsilon \implies \mathbb{E}[\varepsilon | X] = \mathbb{E}[\varepsilon] \implies \text{Cov}(X, \varepsilon) = 0.$$

(c) Show that, *in general, the converse of (b) is false.*

In general, the converse of (b) is false. It is enough to exhibit a counter-example. To do so, consider a real random variable X such that $\mathbb{E}[X] = 0$, $\mathbb{E}[X^3] = 0$, and $\mathbb{V}[X] = \mathbb{E}[X^2] > 0$, that is, a centered, symmetric but not degenerate random variable; for instance, $X \sim \mathcal{N}(0, \sigma^2)$, with $\sigma^2 > 0$ or $X \sim \mathcal{U}[-a, a]$ with $a > 0$ works; and define $\varepsilon := X^2$.

In this case, we have

$$\text{Cov}(X, \varepsilon) = \mathbb{E}[X\varepsilon] - \mathbb{E}[X] \mathbb{E}[\varepsilon] = \mathbb{E}[X^3] - \mathbb{E}[X] \mathbb{E}[X^2] = 0 - 0 \times \mathbb{V}[X] = 0,$$

but,

$$\mathbb{E}[\varepsilon | X] = \mathbb{E}[X^2 | X] = X^2,$$

which cannot be almost surely equal to 0 (otherwise, X would be almost surely equal to 0 too and, therefore, $\mathbb{V}[X] = 0$, which would contradict the hypothesis $\mathbb{V}[X] > 0$).

(d) Find a special case for X such that the converse of (b) is true, that is,

$$\text{Cov}(\varepsilon, X) = 0 \implies \mathbb{E}[\varepsilon | X] = \mathbb{E}[\varepsilon]$$

holds.

Hint: consider a special case of simple linear regressions studied in Chapter 1.

As a special case, the converse of (b) is true if X is binary ($\text{Support}(X) = \{0, 1\}$).

In words, if a random variable is uncorrelated with a binary variable, then it is also mean-independent of that binary variable.

Exercise: prove that result. *Hint:* you can show and use that for two real random variables A and B with B binary, we have $\mathbb{E}[AB] = \mathbb{E}[A | B = 1]\mathbb{E}[B]$. The solution will be provided in the solutions of an upcoming TD.

2 OLS estimator with a single binary regressor

The OLS estimator $\hat{\beta}_D$ of the slope coefficient in the simple linear regression of Y on² a binary covariate D (that is, $\text{Support}(D) = \{0, 1\}$), using an i.i.d. sample $(Y_i, D_i)_{i=1,\dots,n}$ is equal to

²And, as always absent contrary indication, with a constant; formally, $X = (1, D)'$.

In a simple linear regression, as long as there is some variation in the observations of the regressor D , namely (D_1, \dots, D_n) are not all equal (see assumption in Chapter 1, slide 4), or, equivalently, $\widehat{\mathbb{V}}[D] > 0$, the OLS estimator of the slope is equal to

$$\begin{aligned}\widehat{\beta}_D &= \frac{\widehat{\text{Cov}}(Y, D)}{\widehat{\mathbb{V}}[D]} = \frac{(n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y})(D_i - \bar{D})}{(n-1)^{-1} \sum_{i=1}^n (D_i - \bar{D})^2} \\ &= \frac{n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})(D_i - \bar{D})}{n^{-1} \sum_{i=1}^n (D_i - \bar{D})^2} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(D_i - \bar{D})}{\sum_{i=1}^n (D_i - \bar{D})^2}.\end{aligned}$$

Remark that the choice of $n-1$ (to have unbiased estimators) instead of n at the denominator in the definition of the empirical covariance and empirical variance has no impact on the definition of $\widehat{\beta}_D$ since they cancel out.

For this course, it is preferable that you know this result by heart: in a simple linear regression, the OLS estimator of the slope is equal to the empirical covariance between the outcome variable and the regressor divided by the empirical variance of the regressor.

In the particular case of a simple linear regression with a *binary* regressor, that is with values in $\{0, 1\}$, $\widehat{\beta}_D = \bar{Y}_1 - \bar{Y}_0$ (see Chapter 1, slide 9). This explains why the correct answer is Answer 2.

1. $\widehat{\beta}_D = \mathbb{E}[Y | D = 1] - \mathbb{E}[Y | D = 0]$ – **False**, **warning: serious error**. $\widehat{\beta}_D$ is an estimator, a statistic, namely a measurable function of the observations. Here, Answer 1 involves *unknown theoretical non-stochastic* quantities, $\mathbb{E}[Y | D = 1]$ and $\mathbb{E}[Y | D = 0]$ that cannot be used to define an estimator. Besides, such an equality would imply that a random variable $\widehat{\beta}_D$ is equal to a non-stochastic, deterministic quantity (that is, the random variable $\widehat{\beta}_D$ is degenerate) – *always think about the nature of objects*. Here, $\mathbb{E}[Y | D = 1] - \mathbb{E}[Y | D = 0]$ is, in fact, the limit in probability of $\widehat{\beta}_D$ (under proper moment conditions, see Proposition 5 of Chapter 1). Therefore, Answer 1 is like saying $\widehat{\theta} = \theta$ to define an estimator $\widehat{\theta}$ of θ . It would be very convenient but impossible, of course, since we are trying to estimate the *unknown* quantity θ by the estimator (random variable) $\widehat{\theta}$.
2. $\widehat{\beta}_D = \bar{Y}_1 - \bar{Y}_0$, where $\bar{Y}_d := \frac{1}{n_d} \sum_{i:D_i=d} Y_i$, and $n_d := \text{Card}(\{i : D_i = d\})$ for $d \in \{0, 1\}$ – **True**.
3. $\widehat{\beta}_D = \sum_{i=1}^n (D_i - \bar{D})(Y_i - \bar{Y}) / \sum_{i=1}^n (Y_i - \bar{Y})^2$ – **False**, it should be the reverse between Y and D ; that is, the right-hand side is the OLS estimator of the slope in the simple linear regression of D on Y .
4. None of the above; if so, write the correct expression below – **False**.

3 In-sample properties of linear regressions

Let the column vector of regressor $X = (1, D, G')'$, with $D \in \mathbb{R}$ (abuse of notation used in the course to say: D is a real random variable)³. We consider the linear regression of Y on X , and we denote (omitting the subscript i), \widehat{Y} the predicted value obtained from the regression of Y on X , and $\widehat{\varepsilon}$ the (estimated)⁴ residual.

- (a) Choose the correct proposition (or, equivalently, give the definition of $\widehat{\varepsilon}$):

By definition (Chapter 1, first section for the case of Simple Linear Regression (SLR) or second section for the case of Multiple Linear Regression (MLR) as here: see the last point of slide 16), we

³Instead, we shall use the notation $D \in \mathbb{R}^\Omega$. Motivation: as a real random variable, by definition, D is a measurable function from Ω (the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which is not given explicitly) in \mathbb{R} ; that is, $D \in \mathbb{R}^\Omega$.

⁴In English, the word *residual* alone generally denotes what French people would call le “résidus estimé”, $\widehat{\varepsilon}$; while the “résidus théorique” in French is rather called the *error term*, ε .

have, for any observation $i \in \{1, \dots, n\}$,

$$\begin{aligned}\hat{Y}_i &:= X'_i \hat{\beta}, \\ \hat{\varepsilon}_i &:= Y_i - \hat{Y}_i,\end{aligned}$$

where $\hat{\beta}$ is the OLS estimator in the linear regression⁵ of Y on X , that is,

$$\hat{\beta} := \left(\sum_{i=1}^n X_i X'_i \right)^{-1} \left(\sum_{i=1}^n X_i Y_i \right). \quad (2)$$

Therefore, we have, omitting the subscript i , $Y = \hat{Y} + \hat{\varepsilon} = X' \hat{\beta} + \hat{\varepsilon}$: the outcome variable Y is equal to the fitted/predicted value \hat{Y} plus the residual⁶ $\hat{\varepsilon}$.

1. $\hat{Y} = Y + \hat{\varepsilon}$ – **False.**

2. $Y = \hat{Y} + \hat{\varepsilon}$ – **True.**

(b) By definition of OLS estimators and residuals, what can you say about

This is an important result, which can be interpreted as the definition/characterization of the OLS estimator from **Proposition 2 of Chapter 1**. Provided the matrix $\sum_{i=1}^n X_i X'_i$ is invertible (Condition (Inv) of Chapter 1), the OLS estimator $\hat{\beta}$ in the linear regression of Y on X is defined as in Equation (2) and we have

$$\begin{aligned}Y_i &= X'_i \hat{\beta} + \hat{\varepsilon}_i = \hat{Y}_i + \hat{\varepsilon}_i, \text{ and} \\ \overline{X \hat{\varepsilon}} &= 0, \text{ that is, } \frac{1}{n} \sum_{i=1}^n X_i \hat{\varepsilon}_i = 0.\end{aligned}$$

Note that the last equality is *an equality of vectors*: the empirical mean of each component X_j of the regressors X times the residual $\hat{\varepsilon}$ is equal to 0; in other words, the notation “0” in the equality denotes the zero of $\mathbb{R}^{\dim(X)}$: $0_{\mathbb{R}^{\dim(X)}}$.

In particular, whenever X includes a constant, it implies (considering the component corresponding to the constant) that the residuals are centered *in the sample*, that is, for the *empirical mean*:

$$\overline{1 \times \hat{\varepsilon}} = \bar{\hat{\varepsilon}} = \hat{\mathbb{E}}[\hat{\varepsilon}] = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i = 0.$$

The empirical covariance shares the same properties as the theoretical covariance. Especially, as long as the empirical mean of $\hat{\varepsilon}$ is null, the empirical covariance of any component $X^2 j$ of X with $\hat{\varepsilon}$ is equal to the empirical mean of the product of $X^2 j$ and $\hat{\varepsilon}$.

As a consequence, for $X = (1, D, G')'$, we have the following equivalence

$$\overline{X \hat{\varepsilon}} = 0 \iff \begin{cases} \bar{\hat{\varepsilon}} = 0 \\ \widehat{\text{Cov}}(D, \hat{\varepsilon}) = 0 \\ \widehat{\text{Cov}}(G, \hat{\varepsilon}) = 0 \end{cases}.$$

1. $\bar{\hat{\varepsilon}} = 0$

⁵As always, implicitly, absent contrary indication, the moment conditions required to define the theoretical linear regression of Y on X as well as the empirical counterpart, namely the OLS estimator in the linear regression of Y on X , are assumed to be satisfied.

⁶In English, the terminology “residual” directly refers to that *estimated* quantity; in contrast, ε is generally called “the error term”. Note that, in French, there is not such a distinction: ε is often called “le résidu” or “le terme d’erreur”; hence, in French, it is required to state explicitly “le résidu estimé” to refer to $\hat{\varepsilon}$ instead of ε .

2. $\widehat{\text{Cov}}(D, \widehat{\varepsilon}) = 0$
3. $\widehat{\text{Cov}}(G, \widehat{\varepsilon}) = 0$

Deduce and explain why we have $\widehat{\text{Cov}}(\widehat{Y}, \widehat{\varepsilon}) = 0$.

By definition, $\widehat{Y} := X'\widehat{\beta}$, and is thus a linear combination of X . As the theoretical covariance, the empirical covariance is bilinear. Thus, $\widehat{\text{Cov}}(\widehat{Y}, \widehat{\varepsilon}) = 0$ (see bullet-point 4, slide 21, Chapter 1).

This form of writing the linear regression of Y on X as $Y = X'\widehat{\beta} + \widehat{\varepsilon} = \widehat{Y} + \widehat{\varepsilon}$ with all the regressors empirically uncorrelated with the residual: $\widehat{X}\widehat{\varepsilon} = 0$, hence $\widehat{\text{Cov}}(\widehat{Y}, \widehat{\varepsilon}) = 0$ is often useful as a starting point to obtain results such as Frisch-Waugh theorem or the so-called “omitted variable bias” formula (see the proofs of Propositions 3 and 4 of Chapter 1; also see Exercise 3 of 2021/22 mid-term exam).

4 Definition of the R^2

We consider the linear regression of Y on X , where Y is a real random variable and X a (column) random vector, and we denote by \widehat{Y} the predicted value of Y obtained with that regression.

The R^2 of the regression is defined as:

This is directly **the definition of the R^2** (Chapter 1, slide 21, bullet-point 4).

1. $R^2 = \widehat{V}[Y]/\widehat{V}[\widehat{Y}]$ – **False**, it should be the inverse.
2. The probability that the regression measures the causal effect of D on Y – **False**, the R^2 only relates to the usefulness of X to predict Y ; it has nothing to do with a “right” or “wrong” regressions as regards the measurement of causal effects (more to come in the rest of the course).
3. $R^2 = \widehat{V}[\widehat{Y}]/\widehat{V}[Y]$ – **True**, by definition of the R-squared.
4. $R^2 = \widehat{\text{Corr}}(Y, \widehat{Y})$ – **False**, there should be a square to be correct.
5. None of the above; if so, write the correct expression below – **False**.

5 Link between simple and multiple regressions

We consider two linear regressions: `test_ce2` on `pc` (simple linear regression), and `test_ce2` on `pc` and `red` (multiple linear regression) where:

- `test_ce2` is the grade (out of 100) obtained by a pupil for a test taken at the beginning of third grade, that is, CE2;
- `pc` is a dummy variable equal to 1 if a pupil is in a small class in first grade (CP), 0 otherwise; (`pc`: “petite classe”)
- `red` is a dummy variable equal to 1 if a pupil repeats CP, 0 otherwise (`red`: “redoublement”)

Using recent French data, we obtained the following OLS estimates:

$$\begin{aligned}\widehat{\text{test_ce2}} &= 67.4 - 0.56\text{pc}, \\ \widehat{\text{test_ce2}} &= 68.1 - 0.61\text{pc} - 11.4\text{red}.\end{aligned}$$

What can you say about the sign of the empirical covariance between `pc` and `red`?

To relate the coefficients of simple and multiple linear regressions, you should think to **Frisch-Waugh theorem** (Chapter 1, Proposition 3, slide 23) or the **so-called “omitted variable bias” formula** (Chapter 1, Proposition 4, slide 24).

If condition (Inv) of Chapter 1 holds: $n^{-1} \sum_{i=1}^n X_i X_i'$ is invertible, (that is, it is possible to define properly and uniquely the OLS estimator), Proposition 4 gives:

$$\widehat{\beta}_D^S = \widehat{\beta}_D + \widehat{\lambda}' \widehat{\beta}_G.$$

In words, to remember,

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

More precisely:⁷ the short regression is the simple linear regression of Y on D ; the long regression is the multiple linear regression of Y on D and G ; D is the “included variable” while $G = (G^1, \dots, G^p)$ is the “omitted variable(s)”; and the formula says that

- *short*: the coefficient of the included variable D in the short regression =
- *long*: the coefficient of the included variable D in the long regression + the product (scalar product if G is multivariate) of
- *omitted*: coefficients of the omitted variable G in the long regression
- by *coefficients of omitted on included*: the slope coefficients in the auxiliary simple linear regressions of each omitted variables G^j on the included variable D .

Here, it gives with $D = \text{pc}$ and G ($G = G^1 \in \mathbb{R}^\Omega$ is univariate in this example) = red

$$-0.56 = -0.61 + \hat{\lambda} \times -11.4 \iff \hat{\lambda} = \frac{0.05}{-11.4} < 0$$

with $\hat{\lambda}$ the coefficient of G on the simple linear regression of G on D , namely:

$$\hat{\lambda} = \frac{\widehat{\text{Cov}}(G, D)}{\widehat{\text{V}}[D]}.$$

Therefore, $\hat{\lambda}$ is negative and, since the sign of $\hat{\lambda}$ is the sign of the empirical covariance between pc and red , the latter is negative; hence correct Answer 1.

1. the empirical covariance between pc and red is negative – **True**.
2. the empirical covariance between pc and red is positive – **False**.
3. we cannot conclude directly here: we should regress pc on red – **False**, that regression would indeed provide the answer (yet, remark that it would not yield $\hat{\lambda}$ = coefficient of the regression of the omitted red on the included pc , but the reverse), since the OLS slope coefficient would have the same sign as $\widehat{\text{Cov}}(D, G)$. Yet, thanks to the formula of Proposition 4 and previous computations, it is not necessary to perform it.
4. we cannot conclude here because pc and red are perfectly collinear – **False**, nonsense, if that were the case, it would be impossible to make the regression of test_ce2 on pc and red ; yet, that regression is done.

6 Link between simple and multiple regressions (bis)

This question is exactly Question 1 of TD1. I refer to the solutions of that Problem Set for more details (see some handwritten solution notes in Figures 1 and 2. See also the previous Question 5 for another example, with full details, of applying the so-called “omitted variable bias formula”).

We consider two linear regressions: lnwage on eduy (simple linear regression), and lnwage on eduy and age (multiple linear regression) where

- lnwage is the logarithm of the hourly wage,
- eduy is the number of years of education (the count starts at six years old),

⁷As always, absent contrary indications, all regressions include a constant/intercept.

- `age` is the age in years.

Using the French Labor Force Survey data, we obtained the following OLS estimates:

$$\widehat{\lnwage} = 1.60 + 0.053 \text{eduy},$$

$$\widehat{\lnwage} = 0.95 + 0.063 \text{eduy} + 0.015 \text{age}.$$

- (a) What can you say about the sign of the empirical covariance between `eduy` and `age`?

1. the empirical covariance between `eduy` and `age` is negative – **True**.
2. the empirical covariance between `eduy` and `age` is positive – **False**.
3. the empirical covariance between `eduy` and `age` is necessarily null: otherwise, we could not do the multiple linear regression due to collinearity issues – **False**.
4. we cannot conclude here: we should regress `eduy` on `age` – **False**.

In addition to the previous two regressions, we compute the estimates of the expectation and standard deviation of the three variables. The results are displayed below:

Variable	Mean	Std. Dev.
<code>lnwage</code>	2.24	0.40
<code>eduy</code>	12.07	2.69
<code>age</code>	36.47	8.51

- (b) With that additional information, can you compute the value of the empirical correlation between `eduy` and `age`?

1. True – **True** (Answer “True” is True here.)
2. False – **False**.

If so, compute that value: $\widehat{\text{Corr}}(\text{eduy}, \text{age}) \approx -0.22 < 0$ (see distinction between “effet d’âge”, “effet de génération”, “effet de période”).

7 Probability limit of the OLS estimator

As in Chapter 1, Y is the outcome variable, D is the covariate or explanatory variable (setting of a simple linear regression). We assume that Y and D have a finite second-order moment, that $\mathbb{V}[D] > 0$, $\mathbb{E}[DY] = 1$, $\mathbb{E}[Y] = 1$, $\mathbb{E}[D] = 0.5$, and that we have an i.i.d. sample of observations $(Y_i, D_i)_{i=1,\dots,n}$ to compute the OLS estimator of the slope, $\widehat{\beta}_D$, in the linear regression of Y on D (and, implicitly, as always absent contrary indication, a constant, that is Y on $X = (1, D)'$).

When the sample size n goes to infinity, $\widehat{\beta}_D$ converges in probability to β_0 . What can you say about the value of β_0 here?

We check the assumptions relative to the **consistency of the OLS estimator** (Proposition 5 of Chapter 1, slide 31). Here, in a simple linear regression of Y on D (*and always implicitly a constant, also known as intercept, absent contrary indications*), the **column vector X** as defined in the course is the column vector $(1, D)'$, where ' denotes the transpose:

$$X = (1, D)' = \begin{pmatrix} 1 \\ D \end{pmatrix}.$$

- First, Y and each component of X have a finite second-order moment by assumption.

Figure 1: Solution notes – TD1 Exercice 1, Question 1 (page 1)

2022 TD2 2022-TD1
 2020 TD4 - Ex1
 EM 2010 TD3 - Ex2 - salaire, âge, éducation

[7 pages, 5+1] (voir aussi
 do-file stata)
 (21 septembre 2022)

TD
 TD
 TD
 TD
 TD
 TD

1
 1
 1
 1
 1
 1

âge éducati
 salaire
 emp2007

Q1 On a les sorties (estimation par MCO) des deux régressions suivantes :

* Une "courte": y sur x_1 (et la constante) s'écrit : $y = \hat{\alpha} + \hat{\beta}_1 x_1 + \tilde{\epsilon}$ (C) (vii, écriture des régressions théoriques)

où $E(\tilde{\epsilon}) = E(x_1 \tilde{\epsilon}) = 0$ → reprise notation du cours, link between multiple vols

lnw eduy chapitre 1, slide 24 [pour 2019]

$\hat{\alpha}, \hat{\beta}_1 \in \mathbb{R}$ paramètres (Frish-Waugh 1933 empirique) et Prop. 7

* Une "longue": y sur x_1 et x_2 (et constante) : $y, x_1, x_2, \tilde{\epsilon}$ variables aléatoires réelles : $E(x_1 \tilde{\epsilon}) = 0$ E(\tilde{\epsilon}) = 0

$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ (L) α, β₁, β₂ ∈ ℝ paramètres

où $E(\epsilon) = E(x_1 \epsilon) = E(x_2 \epsilon) = 0$ y, x₁, x₂, ε v.a. réelles

lnw eduy age E(ε) = E(x₁ ε) = E(x₂ ε) = 0

lien entre une régression "courte" et une régression "longue"?

→ formule du "bias de variable omise", version théorique (chap 2, Prop. 2) 2022, chap 1, Régression 4

on version empirique (chap 2, Prop. 7) 2020 6, 2019

terminologie théorique, (estimation MCO) 2020 3, 2019

selon la question.

mais N.B.: c'est une relation mécanique, algébrique, entre une régression "courte" et une régression "longue", que l'une des régressions puisse avoir une interprétation causale ou non.

Possos : $(\hat{\alpha}, \hat{\beta}_1)$ estimateur MCO de (C) (C: Courte) N.B. Attention, trois régressions différentes

$(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2)$ estimateur MCO de (L). (L: Longue) régression différente

et considérons la régression de x_2 sur x_1 : $x_2 = \gamma + \lambda x_1 + \eta$ (OSI) Attention, trois régressions différentes

et $\hat{\lambda}$ le coefficient estimé par MCO de la régression de x_2 sur x_1 . N.B. ici $\hat{\lambda}$ et $\hat{\beta}_2 \in \mathbb{R}$ (entre des vecteurs dans la proposition générale du cours)

Preparation 7 (2019) $\hat{\beta}_1 = \hat{\beta}_1 + \hat{\lambda} \hat{\beta}_2$

Preparation 7 (2020) $\hat{\beta}_1 = \hat{\beta}_1 + \hat{\lambda} \hat{\beta}_2$

et N.B. ici $\hat{\lambda}$ et $\hat{\beta}_2 \in \mathbb{R}$ (entre des vecteurs dans la proposition générale du cours)

count = long + effet de l'omise \times coefficient de la régression de l'omise sur l'incluse

count = long + effet de l'omise \times coefficient de la régression de l'omise sur l'incluse

Figure 2: Solution notes – TD1 Exercice 1, Question 1 (page 2)

Typiquement, si $\hat{\beta}_2$ peut être interprété causallement mais non $\hat{\beta}_1$, ou, plus logiquement, est le paramètre d'intérêt réel, celui qu'on cherche à estimer. $\hat{\lambda}'\hat{\beta}_2$ est appelé le biais → terminologie : "biais de variable omise".

BH
TD
XOH
[2]
évaluation
en 2017 selon

On remplace simplement par les estimées obtenues dans les sorties :

$$\hat{\beta}_1 = \hat{\beta}_1 + \hat{\lambda} \hat{\beta}_2 \quad \Leftrightarrow \quad \hat{\lambda} = \frac{\hat{\beta}_1 - \hat{\beta}_1}{\hat{\beta}_2}$$

$$0.053 \approx 0.063 + \hat{\lambda} \times 0.015 \quad \Leftrightarrow \quad \hat{\lambda} \approx \frac{0.053 - 0.063}{0.015} \approx -0.696 < 0$$

avant de remplacer

On sait par ailleurs [estimation MCO dans une régression linéaire simple] :

$$\hat{\lambda} = \frac{\hat{\text{Cov}}(X_1, X_2)}{\hat{\text{Var}}(X_1)} \quad (*)$$

↳ cf. Chapitre 1, définition de l'estimation des MCO

Remarque : $\text{sign}(\hat{\lambda}) = \text{sign}(\hat{\text{Cov}}(X_1, X_2)) = \text{sign}(\hat{\text{Corr}}(X_1, X_2))$

On sait donc que X_1 et X_2 sont empiriquement négativement corrélés, car $\hat{\lambda} \approx -0.696 < 0$.

"
eduy age

Ici, on peut même calculer $\hat{\text{Corr}}(X_1, X_2)$ car on a aussi quelques stats descriptives sur X_1 et X_2 .

$$\text{On a : } \hat{\text{Corr}}(X_1, X_2) = \frac{\hat{\text{Cov}}(X_1, X_2)}{\sqrt{\hat{\text{Var}}(X_1)} \sqrt{\hat{\text{Var}}(X_2)}}$$

$$= \frac{\hat{\lambda} \hat{\text{Var}}(X_1)}{\sqrt{\hat{\text{Var}}(X_1)} \sqrt{\hat{\text{Var}}(X_2)}}$$

$$= \frac{\hat{\lambda} \sqrt{\hat{\text{Var}}(X_1)}}{\sqrt{\hat{\text{Var}}(X_2)}}$$

$$\hat{\text{Corr}}(X_1, X_2) = \frac{\hat{\lambda} \hat{s}(X_1)}{\hat{s}(X_2)}$$

Application

$$\hat{\text{Corr}}(X_1, X_2) \underset{\text{numérique}}{\approx} \frac{-0.696 \times 2.69}{8.51} \approx -0.22 < 0.$$

(définition idem pour – empirique corrélation linéaire)

(autre remarque : effet d'âge ≠ effet de génération)

explication ?

pour $\hat{\text{Corr}}(\text{eduy}, \text{age}) < 0$, ≈ -0.2

- population d'individuals adultes
- démocratisation scolaire

sd : écart-type (standard deviation)

empirique $= \sqrt{\hat{\text{Var}}}$

- Second, with $X = (1, D)'$, we have⁸

$$\mathbb{E}[XX'] = \begin{pmatrix} 1 & \mathbb{E}[D] \\ \mathbb{E}[D] & \mathbb{E}[D^2] \end{pmatrix}.$$

Hence, the determinant of $\mathbb{E}[XX']$ (a 2 by 2 matrix) is equal to $1 \times \mathbb{E}[D^2] - (\mathbb{E}[D])^2 = \mathbb{V}[D]$. Thus, in the case of a simple linear regression, namely $X = (1, D)'$, the assumption $\mathbb{E}[XX']$ invertible is equivalent to $\mathbb{V}[D] > 0$, which is indeed assumed here.

- Finally, we assume to have an i.i.d. sample of observations.

We can, therefore, apply Proposition 5 of Chapter 1. In the special case of a simple linear regression (see Chapter 1, slide 34), it means for the OLS estimator of the slope coefficient:

$$\hat{\beta}_D \xrightarrow[n \rightarrow +\infty]{P} \frac{\text{Cov}(Y, D)}{\mathbb{V}[D]} = \frac{\mathbb{E}[YD] - \mathbb{E}[Y]\mathbb{E}[D]}{\mathbb{V}[D]} = \frac{1 - 1 \times 0.5}{\mathbb{V}[D]} = \frac{0.5}{\mathbb{V}[D]},$$

given the assumptions. *For this course, it is preferable that you know this result by heart: in a simple linear regression, with i.i.d. sampling and under proper moment conditions (finite second moment for the outcome variable and for the regressor and positive variance of the regressor), the OLS estimator of the slope converges in probability to the (theoretical) covariance between the outcome variable and the regressor divided by the (theoretical) variance of the regressor.*

Without more precision, we do not know $\mathbb{V}[D]$. However, if D is binary, that is, a Bernoulli variable, we know that $\mathbb{V}[D] = \mathbb{E}[D](1 - \mathbb{E}[D]) = 0.5 \times (1 - 0.5) = 0.25$, and therefore the limit in probability of $\hat{\beta}_D$ is equal to $0.5/0.25 = 2$ (Answer 1).

1. It cannot be computed with that information, but it is equal to 2 if D is binary – **True**.
2. It cannot be computed with that information (D being binary or not) – **False**.
3. It is equal to 2 (D being binary or not) – **False**.
4. It is equal to 0 (D being binary or not) – **False**.

8 Asymptotic property of the OLS estimator

We consider the simple linear regression of Y on D , where D is a binary variable, that is, $D \in \{0, 1\}$. We assume that $\mathbb{E}[Y^2] < +\infty$ and that $\mathbb{P}(D = 1) \in (0, 1)$.⁹ We denote by $(\hat{\alpha}_D, \hat{\beta}_D)$ the OLS estimator obtained from an i.i.d. sample $(D_i, Y_i)_{i=1,\dots,n}$ with $(D_i, Y_i) \sim (D, Y)$.

When the sample size n goes to infinity, the OLS estimator of the slope $\hat{\beta}_D$

The answer directly follows from the **consistency of the OLS estimator** (Proposition 5 of Chapter 1, slide 31).

Point 1. of Proposition 5 states that with an i.i.d. sample of observations and provided three (mild) moment conditions, namely

1. $\mathbb{E}[|Y|^2] < +\infty$: finite second-order moment for the outcome variable Y ;
2. $\mathbb{E}[\|X\|^2] < +\infty$: finite second-order moment for the column vector X of regressors (i.e., finite second-order moment for each component X^j of X);
3. $\mathbb{E}[XX']$ is invertible: no perfect collinearity among regressors;

⁸Reminder: the expectation of a matrix (respectively a vector) is simply the matrix (resp. the vector) of the expectations.

⁹In English, $(0, 1)$ denotes the open interval $]0, 1[$ (French notation).

the OLS estimator $\hat{\beta}$ of Y on X is well-defined with probability approaching one and

$$\hat{\beta} \xrightarrow[n \rightarrow +\infty]{P} \mathbb{E}[XX']^{-1} \mathbb{E}[XY].$$

This limit in probability is often denoted

$$\beta_0 := \mathbb{E}[XX']^{-1} \mathbb{E}[XY],$$

and referred to as the coefficient of the *theoretical* linear regression of Y on X , or the *theoretical* coefficient of the linear regression of Y on D .

Thus, the first point of Proposition 5 (and the first equality of the second point) says that, **under i.i.d. sampling and weak moment conditions, (i) the theoretical linear regression of Y on X is well-defined, (ii) the associated theoretical coefficient β_0 can be consistently estimated by the OLS estimator $\hat{\beta}$.**

However, to anticipate discussions on causality, the crucial point you will need to understand is that, despite this result, β_0 (the probability limit of the OLS estimator $\hat{\beta}$) is not always the target parameter, the parameter we want to estimate.

We need to check the three moment conditions.

1. By assumption, $\mathbb{E}[|Y|^2] < +\infty$.

2. We consider here a simple linear regression of Y on D : with Chapter 1's notation, we have $X = (1, D)'$; 1 is the constant/intercept; $\mathbb{E}[\|X\|^2] < +\infty$ is equivalent to $\mathbb{E}[\|D\|^2] < +\infty$. By assumption, D is a dummy/binary variable; that is, the support of D is $\{0, 1\}$. Consequently, D is a variable with bounded support and admits finite moment of any order; in particular, we have $\mathbb{E}[\|D\|^2] < +\infty$.

3. As detailed in Question 7, for simple linear regressions ($X = (1, D)'$), $\mathbb{E}[XX']$ invertible is equivalent to $\mathbb{V}[D] > 0$. Again, D is a dummy variable here; in other words, D is distributed according to a Bernoulli distribution. Therefore, we have $\mathbb{V}[D] = \mathbb{P}(D = 1)[1 - \mathbb{P}(D = 1)]$. By assumption, $\mathbb{P}(D = 1) \neq 0$ and $\mathbb{P}(D = 1) \neq 1$ (that is, D is not a constant random variable). It implies $\mathbb{V}[D] > 0$.

The three moment conditions are thus satisfied. Moreover, we use an i.i.d. sample of observations to compute $\hat{\beta}$. Hence, by Proposition 5, we have

$$\hat{\beta} \xrightarrow[n \rightarrow +\infty]{P} \mathbb{E}[XX']^{-1} \mathbb{E}[XY], \quad (3)$$

where $X = (1, D)'$ and $\hat{\beta}$ is the OLS estimator, i.e., a column vector of size two by one → be careful not to confuse the *column vector* $\hat{\beta}$ of estimators of each coefficient with, in the case of a simple linear regression, the OLS estimator of the slope. In the latter case of simple linear regressions, we often write $\hat{\beta} = (\hat{\alpha}, \hat{\beta}_D)'$, with $\hat{\beta}_D$ the OLS estimator of the slope and the result of Equation (3) writes (see Chapter 1, slide 34)

$$\begin{aligned} \hat{\alpha} &\xrightarrow[n \rightarrow +\infty]{P} \alpha_0 := \mathbb{E}[Y] - \frac{\mathbb{C}\text{ov}(D, Y)}{\mathbb{V}[D]} \mathbb{E}[D], \\ \hat{\beta}_D &\xrightarrow[n \rightarrow +\infty]{P} \beta_{0D} \text{ or } \beta_D := \frac{\mathbb{C}\text{ov}(D, Y)}{\mathbb{V}[D]}. \end{aligned}$$

It remains to compute β_D . In the specific case where D is binary, it can be shown that (Chapter 1, slide 34; the computation with *empirical* expectations and covariances is done on slide 9 and the computation is similar with theoretical counterparts; Figure 3 below, as an appendix, gives details on those computations if necessary.)

$$\beta_D := \frac{\mathbb{C}\text{ov}(D, Y)}{\mathbb{V}[D]} \stackrel{\text{when } D \text{ binary}}{=} \mathbb{E}[Y | D = 1] - \mathbb{E}[Y | D = 0].$$

Hence, correct Answer 1 for this question.

1. converges in probability to $\mathbb{E}[Y | D = 1] - \mathbb{E}[Y | D = 0]$ – **True**.
2. does not necessarily converge in probability: we need to further assume that D and Y are independent – **False**, such an assumption does not make much sense since it implies $\beta_{0D} = 0$.
3. does not necessarily converge in probability: we need to further assume that D and Y are uncorrelated – **False**, such an assumption does not make much sense since it implies $\beta_{0D} = 0$.
4. does not necessarily converge in probability: we need to further assume that $\mathbb{E}[D^2] < +\infty$ – **False**, it is implied here by the fact that D is binary.

9 About a moment condition

With Chapter 1's notations, the condition $\mathbb{E}[XX']$ invertible is satisfied when¹⁰

Slide 20 of Chapter 1 explains the meaning of the **full-rank condition (Inv)**:

$$n^{-1} \sum_{i=1}^n X_i X'_i \text{ is invertible.} \quad (\text{Inv})$$

Remark that condition (Inv) can be interpreted as the *sample or empirical analog* of the third moment condition:

$$\mathbb{E}[XX'] \text{ is invertible.} \quad (\text{Inv th})$$

(Inv th) (with “th” for theoretical – see Chapter 2, slide 5), in contrast, can be called *a population or distribution(al) assumption* in so far as it concerns the underlying data-generating process: the X_i are i.i.d. variables with the distribution P_X of a generic variable X ; $\mathbb{E}[XX']$ invertible depends on the distribution P_X only, irrespective of the actual realizations X_1, \dots, X_n of a given sample whereas these realizations determines whether or not condition (Inv), “ $n^{-1} \sum_{i=1}^n X_i X'_i$ invertible”, is satisfied.

The condition “ $\mathbb{E}[XX']$ invertible” or its sample counterpart “ $n^{-1} \sum_{i=1}^n X_i X'_i$ invertible”, is satisfied as long as one cannot recreate any regressor by a linear combination of the other regressors (Chapter 1, slide 20, point 6).

In particular, **when a regressor is a discrete categorical variable** (instead of a quantitative variable; for instance, gender taking two modalities, man and woman) **it is included in the regression as indicators of the different modalities**. Note that it would make no sense to include such qualitative variables in another way since they do not have any quantitative meaning. Since regressions include, by default, an intercept/constant, it is necessary to exclude one of the modalities of the discrete variable (which will be the reference category; we will see interpretations in classes) to avoid perfect collinearity.

1. we regress log(wage) on 1, $\mathbf{1}\{\text{man}\}$, $\mathbf{1}\{\text{woman}\}$ – **False**, assuming only those two values for gender, we have $\mathbf{1}\{\text{man}\} + \mathbf{1}\{\text{woman}\} = 1$, that is, it is possible to recreate one of the regressors (the constant 1) by a linear combination of the others. Therefore, the condition is not satisfied.
2. in an election with two candidates, A and B, we regress the share of the vote in favor of candidate A on 1, the share of total expenditure done by candidate A, and the share of total expenditure done by candidate B – **False**, because there are only two candidates, by construction, the sum of the share of expenditure done by A and the share done by B is equal to 100% = 1, which is the constant.

¹⁰In this specific question, we explicitly state that we include “1”, i.e., a constant in the regression. Yet, elsewhere and in general, following usual conventions and the course, all the regressions include a constant absent contrary indication.

Figure 3: Extrait annoté d'une ancienne version du cours d'Econométrie 1.

Exemple n° 2 (suite) : leçon d'une expérience

Traitement	Taille d'éch.	Diff. de taille moyenne/classe	Diff. de score moyen en CP (/100)
Petit	1925	7.0	8,57 (1,97)
Normal, avec aide	2319	0,7	3,44 (2,05)

Source : Krueger (1999), « Experimental Estimates of Education Production Functions », *Quarterly Journal of Economics*.

Table 1 – Différences p/r au groupe « normal, sans aide »

Régressions linéaires

Lemme 1

- $\beta_0 = \text{Cov}(D, Y)/V(D)$ et $\alpha_0 = E(Y) - \beta_0 E(D)$.
- Il existe ε tel que $Y = \alpha_0 + D\beta_0 + \varepsilon$, avec $E(\varepsilon) = E(D\varepsilon) = 0$.

Preuve : 1) les CPO du programme s'écrivent : $E[Y - \alpha_0 - \beta_0 D] = 0$.
 $E[D(Y - \alpha_0 - \beta_0 D)] = 0$.

2. En remplaçant α_0 par $E(Y - \beta_0 D)$ dans (4), on obtient : $E[Y(1 - D/\beta_0)] = \text{Cov}(D, Y - \beta_0 D) = 0$.
 $= E[Y(1 - D/\beta_0)] = E[Y(1 - D/\beta_0) | D = 1] = E[Y(1 - D/\beta_0) | D = 1]$

Régressions linéaires et effets causaux

Proposition 1

Si $\text{Cov}(D, Y(0)) = 0$, $\beta_0 = \delta^T$. ($\beta = 0$) D binaire $\in \{0, 1\}$, D est binaire.

Preuve : on a : $\text{Cov}(D, Y) \stackrel{(1)}{=} \text{Cov}(D, Y(0) + D\Delta) \stackrel{(2)}{=} \text{Cov}(D, D\Delta) \stackrel{(3)}{=} \Delta^T$ où $\Delta := Y(1 - D) - Y(0)$

$\stackrel{(4)}{=} D(Y(1 - D) + (1 - D)\Delta) \stackrel{(5)}{=} D(Y(1 - D) + V(D))$ où $V(D) := E[(D - E(D))^2]$

$\stackrel{(6)}{=} D(Y(1 - D) + E(D)(1 - E(D))) \stackrel{(7)}{=} D(Y(1 - D) + E(Y|D=1))$ où $E(Y|D=1) = E(Y|D=0)$

$\stackrel{(8)}{=} D(Y(1 - D) + E(Y|D=0)) \stackrel{(9)}{=} D(Y(1 - D) + E(Y))$ où $E(Y) = E(Y|D=1) + (1 - E(D))E(Y|D=0)$

$\stackrel{(10)}{=} D(Y(1 - D) + E(Y)) \stackrel{(11)}{=} D(Y) \stackrel{(12)}{=} \text{Cov}(D, Y)$ où Y peut être une r.v. quelconque d'ici de même : $E[\Delta\Delta] = E(D)^2 E[(1 - E(D))\Delta^2] = E(D)^2 E[(1 - E(D))\Delta^2] = E(D)^2$

$\stackrel{(13)}{=} E(D)^2 = E(D)$ pour D binaire $\in \{0, 1\}$ avec proba 50% / 50%

$\stackrel{(14)}{=} E(D) = E(Y)$ pour D binaire $\in \{0, 1\}$ avec proba 50% / 50%

$\stackrel{(15)}{=} E(D) = E(Y) = 1$ pour D binaire $\in \{0, 1\}$

3. we regress $\log(\text{wage})$ on 1, experience, and the square of the experience – **True**, as long as the data-generating process is such that experience is not a constant.¹¹ Remark that experience and the square of experience are correlated, but this is not an issue as long as the correlation is not perfect: exactly 1 or -1 .
4. we regress $\log(\text{wage})$ on 1, the age, the number of years of schooling since the age of 6 years old, and the number of years since the end of schooling – **False**, by construction age = 6 + the number of years of schooling since the age of 6 years old (X_2) + the number of years since the end of schooling (X_3). Therefore, if $X = (X_1, X_2, X_3, X_4)'$, with $X_1 = 1$ the constant and X_4 the age, we have $X_4 = 6X_1 + X_2 + X_3$: the condition is not satisfied. For a close example, see also Question 5 of TD1.

10 Properties of the OLS estimator

As in Chapter 1, let β_0 denote the probability limit of the OLS estimator $\hat{\beta}$ in the linear regression of Y on X , where Y is a real random variable and X a (column) vector of real random variables.

- (a) We assume the relevant moment conditions to define the theoretical linear regression of Y on X (hence the proper definition of β_0) are satisfied. Write those three conditions.

This is directly the conditions relative to the **proper definition of the limit in probability of the OLS estimator** (Proposition 5 of Chapter 1, slide 31).

- (i) **Finite second-order moment for the outcome variable Y :** $\mathbb{E}[|Y|^2] < +\infty$.
- (ii) **Finite second-order moment for the regressors X :** $\mathbb{E}[\|X\|^2] < +\infty$, where $\|X\|$ denotes the usual Euclidean norm, hence taken at the power 2: $\|X\|^2 = \sum_{j=1}^k X_j^2$. Therefore, by linearity of the expectation, the condition $\mathbb{E}[\|X\|^2] < +\infty$ is equivalent to $\forall j \in \{1, \dots, k\}, \mathbb{E}[|X_j|^2] < +\infty$, that is, finite second-order moment for each component X^j (or X_j depending on notation) of X , for each regressor.
- (iii) **Non-perfect collinearity** (in population, i.e., for the data-generating process) **between the regressors**: $\mathbb{E}[XX']$ is invertible, which is condition (Inv th).

- (b) Under those conditions, give the expression of β_0 .

Under those (mild) moment conditions and i.i.d. sampling, the OLS estimator $\hat{\beta}$ converges in probability to

$$\beta_0 = \mathbb{E}[XX']^{-1} \mathbb{E}[XY].$$

- (c) In addition to the previous moment conditions, under which assumptions $x \mapsto x'\beta_0$ is the best linear approximation of the conditional expectation $x \mapsto \mathbb{E}[Y | X = x]$?

This is precisely the second equality of Point 2 of Proposition 5 in Chapter 1:

$$\beta_0 = \arg \min_{b \in \mathbb{R}^{\dim(X)}} \mathbb{E}[(Y - X'b)^2] = \arg \min_{b \in \mathbb{R}^{\dim(X)}} \mathbb{E}[(\mathbb{E}[Y | X] - X'b)^2] = \arg \min_{b \in \mathbb{R}^{\dim(X)}} \sqrt{\mathbb{E}[(\mathbb{E}[Y | X] - X'b)^2]}$$

In other words, β_0 is such that $x \mapsto x'\beta_0$ is the best *linear* approximation of the conditional expectation $x \mapsto \mathbb{E}[Y | X = x]$, “best” in terms of L^2 norm, or Mean Square Error (MSE); that is, with the minimal L^2 distance between the target $\mathbb{E}[Y | X]$ and the linear approximation $X'\beta_0$.

¹¹Behind that abstract formulation, it depends on the population under study: it is the data-generating process of the observations assumed representative of the population under study. If you have a sample of French workers, it makes sense to assume that people have different experiences. But, as a counterexample, imagine you study a population of youth entering the job market such that by construction, experience = 0 for all of them, then the condition is not satisfied (but, of course, in such a study, having experience as an explanatory variable is not interesting).

Reminder: the distance in L^2 norm between two real random variables U and V is $(\mathbb{E}[(U - V)^2])^{1/2}$.

This result does not require additional assumptions, only the three moment conditions mentioned earlier; hence Answer 4 is the correct one.

Warning: the *best* linear approximation does not say that is a *good* approximation! See the explanations at the end of slide 34 and the example in slide 35 of Chapter 1.

1. The components of X are independent – **False**.
2. Y is a Gaussian variable and X is a Gaussian vector – **False**.¹²

Proposition 2 is suggested for the following reason. If (Y, X) is a Gaussian vector, then the conditional expectation and the linear approximation/regression coincide: $\mathbb{E}[Y | X] = X'\beta_0$. Hence, the conditional expectation is linear, and the approximation is perfect. In general, there is no reason that $\mathbb{E}[Y | X]$ be *linear* and, consequently, $X'\beta_0$ is only an approximation (the best linear one) of $\mathbb{E}[Y | X]$. But, there are *exceptions*. In addition to the case of Gaussian vectors, it also holds in the case of a simple linear regression with a binary regressor: $X = (1, D)'$, with $\text{Support}(D) = \{0, 1\}$. In this case, $\mathbb{E}[Y | X] = \mathbb{E}[Y | D] = \alpha_0 + \beta_D D$, with¹³ (using the notation of Chapter 1, slide 34)

$$\beta_D := \frac{\text{Cov}(Y, D)}{\mathbb{V}[D]}, \quad \alpha_0 := \mathbb{E}[Y] - \beta_D \mathbb{E}[D], \quad (4)$$

that is, the conditional expectation coincides with the linear regression. This result actually holds whenever the regression is “*saturated*”: there are as many parameters as possible values for $\mathbb{E}[Y | X]$ (we will see more precisely this notion in Econometrics 2 in a few months). The case of a simple linear regression with a binary regressor is a special case of saturated regressions: two possible values, $(1, 0)$ and $(1, 1)$ for the vector X of regressors, and two parameters/coefficients, α_0 and β_D .

3. β_0 corresponds to the causal effect of X on Y – **False**: it is a crucial point of the course to distinguish between, on the one hand, the linear projection/regression, which can always be done provided the previous moment conditions hold (Proposition 5 of Chapter 1), and, on the other hand, causal effects defined through potential outcomes variables (Chapter 4 and following). A linear regression has no causal interpretation in general: additional assumptions are required to identify (average) causal effects through a linear regression (to be detailed in the following of the course).
4. None, it is always the case provided the previous three moment conditions hold – **True**.

(d) Explain the meaning of “best” in the above-mentioned expression “the best linear approximation of the conditional expectation”.

See above in the solution of (c): the closest approximation in terms of L^2 norm.

11 A famous theorem

Let Y , X^1 , and X^2 be real random variables with the required moment conditions. To obtain the estimated coefficient of X^1 in the multiple linear regression of Y on X^1 and X^2 , we can

This is directly an application (with different notations but in purpose, for you to remember the meaning of the theorem, not only its symbols) of **Frisch-Waugh Theorem** – Chapter 1, Proposition 3,

¹²Remark : U Gaussian and V Gaussian, without additional assumption as regards their potential dependence (their joint distribution), does not imply that the couple (U, V) is a Gaussian vector.

¹³The first equality, $\mathbb{E}[Y | X] = \mathbb{E}[Y | D]$, comes from the fact that the constant is always implicit: given $X = (1, D)'$ is equivalent to given D . Likewise, we write simply $\mathbb{E}[Y]$, not $\mathbb{E}[Y | 1]$, the best (with respect to the L^2 norm) approximation by “nothing”, a constant, which is the expectation.

slide 23 for estimators (empirical version), Chapter 1, Proposition 6, slide 37 for theoretical coefficients (theoretical version).

The result of that theorem is sometimes called “**regression anatomy**” (see *Mostly Harmless Econometrics*, J. Angrist and J.S. Pischke 2008, Section 3.1) because it explains how are computed and obtained coefficients in multiple linear regressions and, consequently, give intuition on the meaning of multiple linear regressions and of “controlling” by some covariates or control variables.

Let quote J. Angrist and J.S. Pischke as it is topical (see 2021 Nobel Prize in Economics granted to J. Angrist, G. Imbens, and D. Card – initially written in October 2021): “This important formula is said to describe the anatomy of a multivariate regression coefficient because it **reveals much more than the matrix formula**.¹⁴ It shows us that each coefficient in a multivariate regression is the bivariate [= simple linear regression] slope coefficient for the corresponding regressor after “partialling out” all the other variables in the model.”

In words, **Frisch-Waugh theorem asserts that** *the coefficient of a given regressor of interest in a multiple linear regression is equal to the slope coefficient in the simple linear regression of the outcome variable on the residual of the regression of that regressor of interest on all the other regressors*. This is true both for coefficients of the theoretical regressions (Proposition 6) and for empirical counterparts, namely with OLS estimators and estimated residuals (Proposition 3).

1. regress Y on X^2 , then regress the residual of that first regression on X^1 – **False**.
2. regress X^2 on X^1 , then regress Y on the residual of that first regression – **False**, it would yield the coefficient of X^2 in the multiple linear regression of Y on X^1 and X^2 instead of the coefficient of X^1 in that multiple linear regression.
3. regress X^1 on X^2 , then regress Y on the residual of that first regression – **True**.
4. simply regress Y on X^1 since we are only interested in the coefficient of X^1 – **False**, it is generally false (nonetheless, we will see exceptions in a future Problem Set).

12 Effets marginaux (interprétation des coefficients)

This question is written in French; if you cannot read French, please contact me: `lucas [dot] girard [at] ensae [dot] fr`.

Cette question s'intéresse à la notion d'effet marginal (marginal effects). Attention : comme précisé en cours, même si on utilise ce mot “effet”, cette notion est **distincte de la notion de causalité et d'effet causal** qui sera formalisée au Chapitre 4 du cours d'Économétrie 1. Conceptuellement, ce serait plutôt **l'effet sur la prédiction** (comment la prédiction linéaire de Y par X varie-t-elle ?) induit par une variation marginale d'un régresseur, les autres régresseurs étant fixés (raisonnement “toutes choses égales par ailleurs”). Gary Chamberlain (cf. Lecture Note 2 – [link](#)) utilise le terme de “**predictive effect**” (par opposition à “causal effect”) : “It measures how the prediction of Y changes as we change the value for one of the predictor variables, holding constant the value of the other predictor variables”.

$X = (X_0, X_1, \dots, X_{k-1})' \in (\mathbb{R}^k)^\Omega$ désigne¹⁵ le vecteur (*colonne*) aléatoire des covariables (aussi appelées : régresseurs, variables explicatives, variables “indépendantes”) et $Y \in \mathbb{R}^\Omega$ est la variable aléatoire réelle de résultat (aussi appelée : variable expliquée, variable “dépendante”).

¹⁴Matrix formula $\hat{\beta} := \widehat{\mathbb{E}}[XX']^{-1}\widehat{\mathbb{E}}[XY] = (n^{-1}\sum_{i=1}^n X_i X_i')^{-1}(n^{-1}\sum_{i=1}^n X_i Y_i)$, and for its limit in probability (theoretical coefficient) under i.i.d. sampling and standard moment condition $\beta_0 := \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$.

¹⁵Ici, pour éviter les confusions avec des puissances qui interviendront, on utilise la notation X_j (contre X^j dans le cours), pour $j \in \{0, 1, \dots, k\}$, pour désigner la j -ème composante de X . Les deux notations sont employées. Pour éviter la confusion avec les indices des observations, on utilise traditionnellement la lettre $i \in \{1, \dots, n\}$ pour numérotter les observations et une autre lettre j (ou k , ℓ) pour les composantes. Ainsi, $X_i \in (\mathbb{R}^k)^\Omega$ est le vecteur aléatoire des covariables pour la i -ème observation. Alors que $X_j \in \mathbb{R}^\Omega$, une variable aléatoire réelle, est la j -ème composante de X , X étant une instance générique du vecteur des covariables ayant la même loi que X_i (*il faut bien comprendre ce point de notation et la possibilité d'omettre l'indice i des observations puisqu'on les suppose i.i.d.*). Enfin, X_{ij} ou X_i^j désigne la j -ème composante de la i -ème observation ; c'est également une variable aléatoire réelle : X_i^j ou $X_{ij} \in \mathbb{R}^\Omega$.

On note $\beta_0 = (\beta_{00}, \beta_{01}, \dots, \beta_{0k-1})' \in \mathbb{R}^k$ le vecteur (non stochastique) des coefficients associés à la régression linéaire (théorique) de Y sur X . Sous les conditions de moments requises (voir Question 10.a), qu'on suppose ici vérifiées, β_0 est la limite en probabilité de l'estimateur MCO, $\hat{\beta}$, de Y sur X , et $\beta_0 = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$ (voir Proposition 5, Chapitre 1).

Alors $X'\beta_0$ est, par construction, la meilleure (au sens de la plus proche pour la norme L^2) approximation de Y par une fonction *linéaire* de X . Remarque : $X'\beta_0 \in \mathbb{R}^\Omega$ est une variable aléatoire réelle. Dit autrement, $X'\beta_0$ est la prédiction linéaire de Y ; la prédiction *théorique* par opposition à la *valeur prédictive* (ou prédiction *empirique*) $\hat{Y} := X'\hat{\beta}$ qui utilise l'estimateur MCO, $\hat{\beta}$, à la place de β_0 (β_0 est inconnu bien sûr, on cherche à l'estimer avec un échantillon de taille finie par $\hat{\beta}$). On peut aussi la noter¹⁶ $\mathbb{L}[Y | X] := X'\beta_0$ ou $\mathbb{E}_{\text{lin}}[Y | X] := X'\beta_0$ (avec la lettre \mathbb{L} ou l'indice “lin” pour linéaire).

Remarque : cette question est entièrement écrite pour les quantités théoriques (voir slide 36 “Effets marginaux théoriques”). Toutefois, il faut bien voir qu'on pourrait tout faire de même avec les contreparties empiriques pour les “effets marginaux”, sous-entendus “empiriques” ou “estimés” (voir diapositives 17 et 18).

On considère ici une variable explicative d'intérêt continue, par exemple la première X_1 (tout marcherait pareil pour une composante quelconque X_j). *L'effet marginal théorique* (respectivement *empirique*) de X_1 (sur Y)¹⁷ est la dérivée de l'application partielle qui à X_1 associe la prédiction linéaire théorique $\mathbb{L}[Y | X] := X'\beta_0$ (respectivement la prédiction empirique ou valeur prédictive $\hat{Y} := X'\hat{\beta}$), les autres variables explicatives éventuelles¹⁸, X_2, \dots, X_{k-1} , étant donc fixées.

Remarque cruciale : *l'effet marginal est donc une fonction*. En général, il faut donc parler de l'effet marginal de X_1 sur Y évalué en un $x \in \text{Support}(X)$ particulier. Cette quantité (qui est un nombre réel, car c'est la dérivée évaluée en un point d'une fonction de \mathbb{R} dans \mathbb{R} , et qui est inconnue ; on devra l'estimer par sa contrepartie empirique) est ainsi définie formellement par :

$$\begin{aligned} \text{effet marginal (théorique) de } X_1 \text{ sur } Y \text{ en } x &:= \frac{\partial X'\beta_0}{\partial X_1} \Big|_{X=x} = \frac{\partial \mathbb{L}[Y | X]}{\partial X_1} \Big|_{X=x} \in \mathbb{R} \\ &= \frac{\partial \mathbb{L}[Y | (X_1, X_{-1})]}{\partial X_1} \Big|_{X=(x_1, x_{-1})} \quad (\text{Eff. Marg. th.}) \end{aligned}$$

Remarques : la première égalité (après la définition “:=”) vient de la définition de la notation $\mathbb{L}[Y | X]$; la seconde égalité vient juste à nouveau d'une notation : $X = (X_1, X_{-1})$ pour le vecteur aléatoire des régresseurs (et idem pour une réalisation particulière, non stochastique, en lettre minuscule, $x = (x_1, x_{-1})$). Cette décomposition est intéressante en pratique pour faire les calculs puisqu'on va dériver par rapport à x_1 seulement (dérivée partielle).

On s'intéresse, pour différents modèles, à l'effet marginal (théorique) de X_1 sur Y ; $X_2 \in \mathbb{R}^\Omega$ est une autre variable aléatoire explicative réelle. Pour chacun des modèles suivants, de (a) à (e),

1. Calculer l'effet marginal de X_1 sur Y évalué en un x quelconque dans le support de X .
2. Est-ce que cet effet marginal dépend de la valeur x des régresseurs ?
3. Donner l'expression de *l'effet marginal (théorique) moyen de X_1 sur Y* . De façon générale, cette quantité (c'est un nombre réel contrairement à l'effet marginal qui est une fonction) est définie comme *l'espérance de l'effet marginal (théorique) de X_1 sur Y prise en X , qui est aléatoire, où l'espérance porte sur les régresseurs X* (l'indice X sous l'espérance explicite cela) :

$$\text{effet marginal (théorique) moyen de } X_1 \text{ sur } Y := \mathbb{E}_X \left[\frac{\partial X'\beta_0}{\partial X_1} \right] = \mathbb{E}_X \left[\frac{\partial \mathbb{L}[Y | X]}{\partial X_1} \right] \in \mathbb{R} \quad (\text{Eff. Marg. Moyen th.})$$

¹⁶Par analogie avec la meilleure approximation par une fonction *quelconque* : l'espérance conditionnelle $\mathbb{E}[Y | X]$.

¹⁷On précise parfois “l'effet marginal de X_1 sur Y ”, et parfois non, en laissant alors implicite le fait qu'il s'agit de l'effet sur la variable expliquée étudiée Y .

¹⁸Notation : on écrit X_{-1} (ou X^{-1} dans le cas des notations des diapositives du Chapitre 1) pour désigner le vecteur X des variables explicatives en excluant le régresseur X_1 .

Quelques remarques et explications complémentaires sur ces définitions.

- (i) On considère une **dérivée** car on s'intéresse à un **effet marginal**. Pour un régresseur X_j binaire, la notion d'un changement *marginal* de X_j n'a pas de sens puisqu'il n'y a que deux valeurs possibles, **0** et **1**. L'effet marginal est donc remplacé par *la différence discrète* suivante :

$$\mathbb{L}[Y | X_{-j} = x_{-j}, X_j = 1] - \mathbb{L}[Y | X_{-j} = x_{-j}, X_j = 0].$$

Plus largement, c'est la même idée de considérer une différence au lieu d'une dérivée pour un régresseur X_j **discret ou catégoriel** (ayant un sens *qualitatif* et non *quantitatif*). Dans ce cas, l'effet marginal (en plus d'être une fonction) est également défini relativement à deux modalités distinctes $(\textcolor{brown}{u}, \textcolor{blue}{v}) \in \text{Support}(X_j)$ possibles de X_j . C'est l'effet (sur la prédiction linéaire $\mathbb{L}[Y | X] = X'\beta_0$) de passer de $X_j = \textcolor{brown}{u}$ à $X_j = \textcolor{blue}{v}$, les autres régresseurs, X_{-j} étant fixés :

$$\mathbb{L}[Y | X_{-j} = x_{-j}, X_j = \textcolor{blue}{v}] - \mathbb{L}[Y | X_{-j} = x_{-j}, X_j = \textcolor{brown}{u}].$$

L'interprétation se fait donc relativement à une modalité de référence, ici **u** .

- (ii) Le fait de considérer une dérivée **partielle** renvoie à l'idée d'un **effet de X_j sur Y “toutes choses égales par ailleurs”** (t.c.e.p.a; *ceteris paribus*) : on regarde l'effet d'une variation de X_j “sur Y ” (plus précisément, sur la meilleure prédiction linéaire en X de Y) en gardant les autres composantes de X fixées. En général, la valeur de cet effet marginal dépend donc évidemment de la valeur x_{-j} fixée des autres composantes puisque l'on considère la dérivée d'une application partielle.
- (iii) On rappelle que **la dérivée est une notion locale**. Par conséquent, a priori, la valeur de l'effet marginal dépend donc également de la valeur x_j à laquelle est évaluée la dérivée partielle.

L'effet marginal est donc en général une fonction qui dépend du point $x \in \text{Support}(X)$ considéré où on l'évalue.

L'exception à la règle est le cas d'un modèle linéaire “simple”¹⁹ où il s'avère que l'effet marginal ne dépend ni de la valeur de x_j , ni de la valeur x_{-j} des autres régresseurs. Il est important de comprendre qu'il s'agit d'un *cas particulier*, de l'exception qui confirme la règle, et que l'effet marginal dépend généralement de la valeur x des régresseurs.

L'effet marginal de X_j sur Y étant en général une fonction, il n'y a pas vraiment de sens à parler de “l’effet marginal au singulier” ; plus précisément, cela désigne une fonction et non un nombre. Toutefois, souvent, on préfère avoir un nombre au lieu d'une fonction pour synthétiser quantitativement l'effet de X_j sur Y , de là les notions d'effet marginal moyen, ou d'effet marginal à la moyenne (ou à un autre point, ou un effet marginal moyen pour un sous-groupe particulier).

L'effet marginal moyen d'une variable explicative continue X_j sur Y est défini en prenant l'espérance (où l'espérance porte sur X)²⁰ de l'effet marginal – voir (Eff. Marg. Moyen th.) ci-dessus.

Remarque : de ce fait, l'effet marginal moyen ne dépend pas seulement de la loi conditionnelle de Y sachant X , mais également de la loi de X , donc de la population d'intérêt considérée ; formellement, c'est une fonction de la loi jointe $P_{(Y,X)}$.

Pour une variable X_j binaire (discrète plus généralement), la définition est identique avec la différence discrète à la place de l'effet marginal :

$$\text{effet marginal moyen de } X_j \text{ (binaire) sur } Y := \mathbb{E}_X \left(\mathbb{L}[Y | X_{-j}, X_j = 1] - \mathbb{L}[Y | X_{-j}, X_j = 0] \right) \quad (\text{Eff. Marg. Moyen th. (discret)})$$

¹⁹Où “ simple” signifie ici sans interactions entre les variables explicatives ni de puissances ou d'autres transformations, les variables explicatives apparaissent seulement en niveau, à la puissance 1 – voir le modèle (a) ci-dessous.

²⁰Convention habituelle : variables aléatoires stochastiques en lettres majuscules et réalisations, nombres non stochastiques en lettres minuscules.

Comme expliqué plus haut, l'effet marginal étant une fonction, on peut aussi simplement l'évaluer en un point particulier donné $x = (x_j, x_{-j})$: c'est l'effet marginal de X_j sur Y au point x . Pour $x = \mathbb{E}[X]$, il s'agit de **l'effet marginal à la moyenne** :

$$\text{effet marginal (théorique) à la moyenne de } X_j \text{ sur } Y := \frac{\partial \mathbb{L}[Y | X]}{\partial X_j} \Big|_{X=\mathbb{E}[X]} \in \mathbb{R}. \\ (\text{Eff. Marg. th. à la moyenne})$$

Comme l'effet marginal moyen, cette quantité ne dépend pas uniquement de la loi conditionnelle $P_{Y|X}$, mais aussi de la distribution marginale de X , via l'espérance $\mathbb{E}[X]$, et donc, au final, de la loi jointe $P_{(Y,X)}$ du couple (Y, X) .

(a) Modèle linéaire (cas simple, c'est-à-dire, avec la terminologie du cours, lorsque les régresseurs ne sont pas fonctionnellement dépendants) Y continue, $X = (1, X_1, X_2)'$ et

$$\mathbb{L}[Y | X] = X' \beta_0 = \beta_{00} + \beta_{01} X_1 + \beta_{02} X_2,$$

ce qu'on peut écrire de façon équivalente²¹

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

Pour tout $x = (1, x_1, x_2)$, $x_1 \in \mathbb{R}$, $x_2 \in \mathbb{R}$, on a

$$\frac{\partial \mathbb{L}[Y | X=x]}{\partial x_1} = \beta_{01}.$$

L'effet marginal, dans ce cas très particulier, ne dépend pas de x ; c'est une fonction constante égale à β_{01} . Par conséquent, l'effet marginal moyen est simplement égal à

$$\mathbb{E}_X \left(\frac{\partial \mathbb{L}[Y | X]}{\partial X_1} \right) = \mathbb{E}_X [\beta_{01}] = \beta_{01}.$$

De même, l'effet marginal à la moyenne, comme l'effet marginal en tout point x , est également égal à β_{01} .

(b) Modèle linéaire (avec des puissances, par exemple un effet quadratique de X_1 ; un cas où les composantes de X sont fonctionnellement dépendantes) Y continue, $X = (1, X_1, X_1^2, X_2)'$, où X_1^2 désigne le carré de la composante X_1 : $X_1^2 = X_1 \times X_1$, et

$$\mathbb{L}[Y | X] = X' \beta_0 = \beta_{00} + \beta_{01} X_1 + \beta_{02} X_1^2 + \beta_{03} X_2.$$

Pour tout $x = (1, x_1, x_2)$, $x_1 \in \mathbb{R}$, $x_2 \in \mathbb{R}$, on a

$$\frac{\partial \mathbb{L}[Y | X=x]}{\partial x_1} = \beta_{01} + 2x_1 \beta_{02}.$$

L'effet marginal dépend donc de x . Plus précisément, dans un modèle linéaire sans interaction entre différents régresseurs (uniquement des puissances ou d'autres fonctions de X_1), l'effet marginal de X_1 ne dépend pas de la valeur x_{-1} des autres régresseurs, mais est bien une fonction de x_1 , et donc de x . Cela permet de prendre en compte un effet quadratique et non simplement linéaire de X_1 sur Y .

²¹**Remarque :** il est important de bien comprendre que ces deux écritures sont équivalentes ; elles écrivent simplement la projection (ou régression) linéaire théorique de Y sur X , ce qu'on peut toujours bien définir sous de simples conditions de moments : Y et X appartiennent à L^2 (c'est-à-dire sont de carré intégrable) et $\mathbb{E}[XX']$ est inversible (c'est-à-dire qu'il n'y a pas de colinéarité parfaite entre les composantes de X) (voir Proposition 5, Chapitre 1). En ce sens, un "modèle linéaire" de la forme $Y = X' \beta_0 + \varepsilon$ avec $\mathbb{E}[X \varepsilon] = 0$ est (quasiment) tautologique : on ne dit rien si ce n'est que les (relativement faibles) conditions de moments permettant de définir la projection linéaire théorique sont respectées. Par la suite, on utilisera l'une ou l'autre de ces deux formulations équivalentes.

L'effet marginal moyen est alors

$$\mathbb{E}_X \left(\frac{\partial \mathbb{L}[Y | X]}{\partial X_1} \right) = \mathbb{E}_X [\beta_{01} + 2X_1\beta_{02}] = \beta_{01} + 2\beta_{02} \mathbb{E}[X_1]$$

où la dernière égalité utilise la linéarité de l'espérance. L'effet marginal moyen dépend donc des caractéristiques de la population considérée, ici de $\mathbb{E}[X_1]$. Dans un tel modèle linéaire, il s'avère que l'effet marginal moyen est aussi égal à l'effet marginal à la moyenne ; mais ce n'est pas le cas en général (voir l'exemple du modèle (e)).

(c) Modèle linéaire (avec des interactions – ici, un terme dit d'interaction sans “main effect” de X_1) ; un autre cas où les composantes de X sont fonctionnellement dépendantes Y continue, $X = (1, X_1 \times X_2, X_2)'$ et

$$Y = \beta_{00} + \beta_{01}X_1X_2 + \beta_{02}X_2 + \varepsilon \text{ avec } \mathbb{E}[X\varepsilon] = 0.$$

Pour tout $x = (1, x_1, x_2)$, $x_1 \in \mathbb{R}$, $x_2 \in \mathbb{R}$, on a

$$\frac{\partial \mathbb{L}[Y | X = x]}{\partial x_1} = \beta_{01}x_2.$$

L'effet marginal dépend donc de x et il dépend en fait de la valeur $x_{-1} = x_2$ ici des autres régresseurs seulement, et non de x_1 . L'effet marginal moyen vaut alors

$$\mathbb{E}_X \left(\frac{\partial \mathbb{L}[Y | X]}{\partial X_1} \right) = \mathbb{E}_X [\beta_{01}x_2] = \beta_{01} \mathbb{E}[X_2]$$

par linéarité de l'espérance. Il est aussi égal à l'effet marginal à la moyenne.

(d) Modèle linéaire (avec des interactions – ici, un terme dit d'interaction et un “main effect” de X_1 ; encore un autre cas où les composantes de X sont fonctionnellement dépendantes) Y continue, $X = (1, X_1, X_2, X_1 \times X_2)'$ et

$$Y = \beta_{00} + \beta_{01}X_1 + \beta_{02}X_2 + \beta_{03}X_1X_2 + \varepsilon \text{ avec } \mathbb{E}[X\varepsilon] = 0.$$

Pour tout $x = (1, x_1, x_2)$, $x_1 \in \mathbb{R}$, $x_2 \in \mathbb{R}$, on a

$$\frac{\partial \mathbb{L}[Y | X = x]}{\partial x_1} = \beta_{01} + \beta_{03}x_2.$$

L'effet marginal dépend de x via la valeur des autres régresseurs $x_{-1} = x_2$. L'effet marginal moyen est alors égal à

$$\mathbb{E}_X \left(\frac{\partial \mathbb{L}[Y | X]}{\partial X_1} \right) = \mathbb{E}_X [\beta_{01} + \beta_{03}x_2] = \beta_{01} + \beta_{03} \mathbb{E}[X_2]$$

par linéarité de l'espérance. Il est à nouveau égal à l'effet marginal à la moyenne.

(e) Modèle linéaire (un autre exemple où les composantes de X sont fonctionnellement dépendantes) Y continue, $X = (1, X_1, X_2, X_1^2, X_1^2 \times X_2)'$ et

$$\mathbb{L}[Y | X] = \beta_{00} + \beta_{01}X_1 + \beta_{02}X_2 + \beta_{03}X_1^2 + \beta_{04}X_1^2X_2.$$

Qu'importe la complexité du modèle (en pratique, il faut réfléchir et se demander quelle forme fonctionnelle pour $\mathbb{L}[Y | X]$ est la plus pertinente), il suffit comme auparavant de suivre les définitions. Ainsi, pour tout $x = (1, x_1, x_2)$, $x_1 \in \mathbb{R}$, $x_2 \in \mathbb{R}$, on a ici

$$\frac{\partial \mathbb{L}[Y | X = x]}{\partial x_1} = \beta_{01} + 2\beta_{03}x_1 + 2\beta_{04}x_1x_2$$

L'effet marginal dépend de x , à la fois via x_1 et via les autres régresseurs $x_{-1} = x_2$. L'effet marginal moyen est alors

$$\mathbb{E}_X \left(\frac{\partial \mathbb{L}[Y | X]}{\partial X_1} \right) = \mathbb{E}_X [\beta_{01} + 2\beta_{03}x_1 + 2\beta_{04}x_1x_2] = \beta_{01} + 2\beta_{03}\mathbb{E}[X_1] + 2\beta_{04}\mathbb{E}[X_1X_2]$$

par linéarité de l'espérance. Il dépend ici d'un moment croisé, $\mathbb{E}[X_1X_2]$, donc de la covariance entre X_1 et X_2 . En général, $\text{Cov}(X_1, X_2) \neq 0$, c'est-à-dire, $\mathbb{E}[X_1X_2] \neq \mathbb{E}[X_1]\mathbb{E}[X_2]$. Par conséquent **ici**, même s'il s'agit d'un modèle linéaire, l'effet marginal moyen n'est pas égal à l'effet marginal à la moyenne, qui est l'effet marginal évalué en $x = \mathbb{E}[X] = (1, \mathbb{E}[X_1], \mathbb{E}[X_2])'$, et qui vaut donc ici :

$$\text{Effet marginal à la moyenne : } \left. \frac{\partial \mathbb{L}[Y | X = x]}{\partial x_j} \right|_{x=\mathbb{E}[X]} = \beta_{01} + 2\beta_{03}\mathbb{E}[X_1] + 2\beta_{04}\mathbb{E}[X_1]\mathbb{E}[X_2].$$

Level-level, log-level, level-log, or log-log linear regressions This point is aside from the initial quiz question, yet it also concerns the interpretation of coefficients.

Besides functional dependence between the components of X , it is possible, **for a given explanatory variable of interest, D a real random variable, to include in the linear regression D itself (its *level*) or the logarithm of D , $\log(D)$ (its *log*)**. Likewise, for an outcome variable $Y \in \mathbb{R}^\Omega$ of interest, we can consider a linear regression of Y on some regressors or of its logarithm $\log(Y)$. In such cases, **the interpretation of the coefficient** (either theoretical or their empirical counterparts, namely OLS estimators) is modified when you want to say something in terms of the initial (in levels) variables D and Y .

Such linear regressions are often called “log-level”, “level-log”, or “log-log” models. **Introducing logarithms allows us to account for non-linearities**. In a basic level-level linear regression of Y on D (and an intercept as usual), the marginal effect of D is constant (see model (a) above): if D increases by one unit (absolute change), we modify our best linear prediction of Y by β_D unit (absolute change). On the contrary, for a log-level model, for instance, that is a linear regression of Y on $\log(D)$, the interpretation mixes relative and absolute changes (see details below).

In that sense, an important point to remember is that linear regression models are not restricted to linear effects! Remark that we already knew that through the introduction of functional-dependent regressors, like D and D^2 to model a quadratic effect; see model (b) above; considering logarithms (for the outcome variable or/and regressors) adds another layer to this (relative) flexibility of linear regressions.

Summary In the course, Chapter 4 will explain that with more details. However, as this could be used before during tutorial sessions, below is a quick review of the interpretations of the different usual linear regressions in levels or logs that can link an outcome variable of interest Y and an explanatory variable of interest D .

Remark 1: the interpretations below are formulated in terms of prediction, which is always possible for any linear regression, or equivalently in terms of marginal effects, which is the same notion: how the (best linear) prediction of Y is modified when there is a marginal change (either an absolute change – level – or a relative change – log) of D (above, X_1 played the role of D).

If the linear regression does identify causal effects, *which is not always the case!* (see later in the course, notably Chapter 4), the interpretation can also be made in terms of causal effects.

Remark 2: 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.

Remark 3: we assume here that D or $\log(D)$ enters in the regression “simply”, namely there is no power or other transformations of D nor interactions with G ; in other words, there is *no* functional dependence between the regressors: ($X = (1, D, G')'$ (level-level or log-level cases) or $X = (1, \log(D), G')'$ (level-log or log-log cases)).

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.

The precision “approximately” refers to the fact that this is only an approximation, valid for a small variation of D (see details in Figures 4 and 5).

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

If D increases by 1 unit (absolute change), we predict that, all else being equal, Y changes by β_D units (absolute change).

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

If D increases by 1 unit (absolute change), we predict that, all else being equal, 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):
$$\Delta Y \approx (\beta_D/100) \times \% \Delta D$$

If D increases by 1% (relative change), we predict that, all else being equal, Y approximately changes by $\beta_D/100$ units (absolute change).

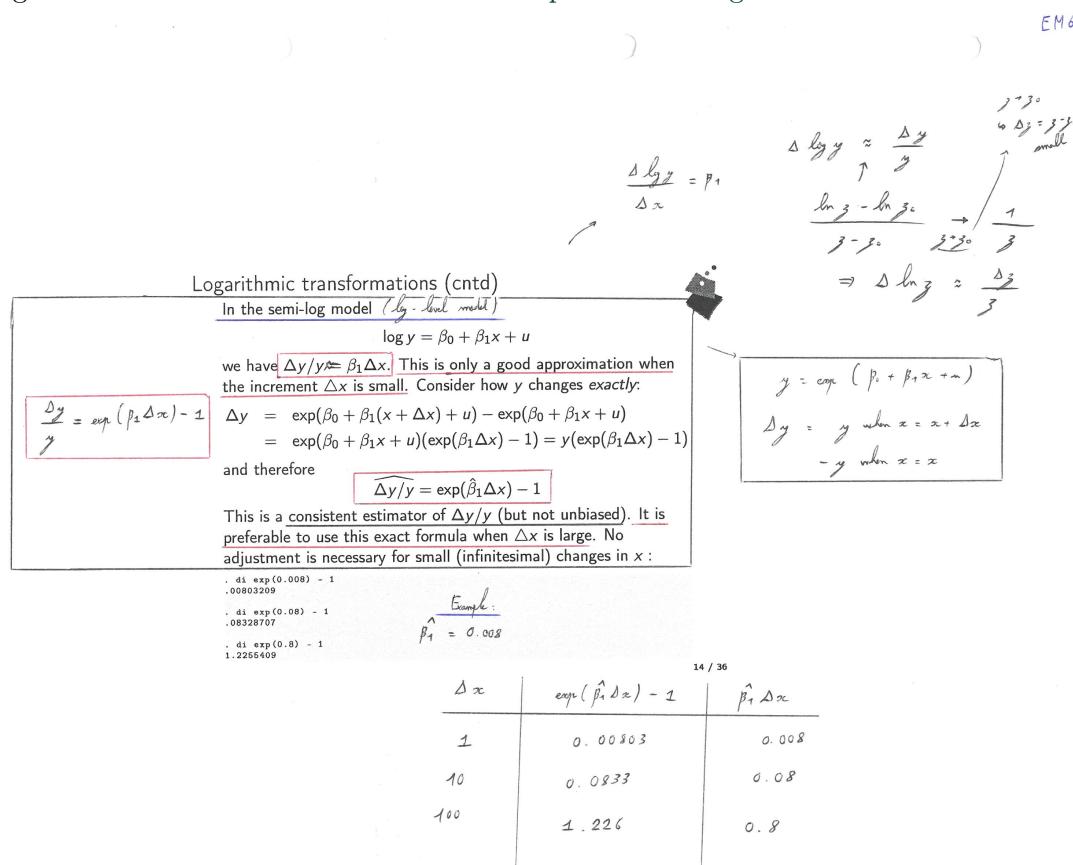
- **log-log:** regression of $\log(Y)$ on $\log(D)$ (and G):
$$\% \Delta Y \approx \beta_D \times \% \Delta D$$

If D increases by 1% (relative change), we predict that, all else being equal, 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.²²

²²In 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 4: Handwritten notes about the interpretation of log-level models.



13 *Population d'intérêt et représentativité

This question is written in French; if you cannot read French, please contact me: [lucas\[dot\]girard\[at\]ensae\[dot\]fr](mailto:lucas[dot]girard[at]ensae[dot]fr).

Figure 6: Un extrait d'un article du *Monde* publié le 1er octobre 2017.

Le gouvernement catalan avait promis d'aller au bout de son projet de référendum sur l'indépendance de la région, et il y est parvenu, dimanche 1^{er} octobre.

Selon Barcelone, le oui a gagné avec 90 % des voix. Quelque 2,26 millions de personnes ont participé au scrutin et 2,02 millions se sont exprimées en faveur de l'indépendance, a assuré le porte-parole du gouvernement catalan, Jordi Turull, dans la soirée. Ces chiffres représentent une participation de près de 42,3 %, la Catalogne comptant 5,34 millions d'électeurs.

Dans les exercices, durant les TD, nous tendrons à mettre de côté certaines questions pourtant cruciales, qu'il faut toujours garder en tête en face d'un jeu de données pour faire de véritables analyses. Cette question, un peu à part, est là pour en rappeler certaines.

Dans cet exemple (Figure 6),

1. Quelle est la population d'intérêt ciblée ?
2. Quelle est la population “effective” couverte par les données ?

Figure 5: Handwritten notes for solutions of TD1 – page 7 about the interpretation of log-level models (see Gary Chamberlain lecture note 2 for more details).

Details sur l'interprétation du log-level model [cf cours Gary Chamberlain, cours 2, section 5]

Assume: $E[\log(Y)|Z] = \beta_0 + \beta_1 Z$ where Y, Z real random variable
 β_0, β_1 real numbers (parameters)

Define $U := \log(Y) - E[\log(Y)|Z]$

so that $E(U|Z) = 0$ given the assumption (#)

Since $\log(Y) = E[\log(Y)|Z] + U$ $\log = \ln$ (base $e \approx 2,718$)
 $\log(Y) = \beta_0 + \beta_1 Z + U$

we have $Y = \exp(\beta_0 + \beta_1 Z + U)$

$\therefore Y = \exp(\beta_0 + \beta_1 Z) \exp(U)$ (equality of random variable)
 $\Rightarrow E(Y|Z) = \exp(\beta_0 + \beta_1 Z) \times E[\exp(U)|Z]$ en prenant l'espérance conditionnelle

In general, $E(U|Z) = 0 \nrightarrow E[\exp(U)|Z]$ is a constant.

Under the additional assumption that it is indeed the case: it is the case for instance,
 $E[\exp(U)|Z] = E[\exp(U)]$, in particular, if U and Z are assumed independent

we have, for any possible values d and c of the explanatory variable Z , $\forall c, d \in \text{Supp}(Z)$

$$\frac{E(Y|Z=d)}{E(Y|Z=c)} = \frac{\exp[(\beta_0 + \beta_1 d) - (\beta_0 + \beta_1 c)]}{\exp[(\beta_0 + \beta_1 c)]} = \frac{E[\exp(U)]}{E[\exp(U)]}$$

$$\frac{E(Y|Z=d)}{E(Y|Z=c)} = \exp(\beta_1(d-c))$$

$$\frac{E(Y|Z=d)}{E(Y|Z=c)} \approx 1 + \beta_1(d-c)$$
 i.e. pour des petits changements de Z passant de $c \approx d$: $|d-c| \approx 0$

d'où: $[-] = \frac{E(Y|Z=d) - E(Y|Z=c)}{E(Y|Z=c)} \approx \frac{1}{E(Y|Z=c)} \times \beta_1(d-c)$ = changement relatif de $E(Y|Z)$ en passant Z de $c \approx d$

$100 \left[\frac{E(Y|Z=d)}{E(Y|Z=c)} - 1 \right] \approx 100 \beta_1 \times (d-c)$ $\Delta \times 100$ sur le coefficient β_1 d'où l'interprétation $\log(Y)$ sur Z

Indice : vous pouvez relire le Chapitre 0 (Introduction), notamment les diapositives relatives aux données de coupe (“cross-sectional data”). Nous aborderons ces problèmes de sélection de façon plus formalisée au second semestre en Économétrie 2 avec les modèles de sélection.

1. La population d’intérêt pourrait être, par exemple²³, l’ensemble des personnes résidant en Catalogne, de nationalité espagnole, et âgées de plus de 18 ans, soit les 5,34 millions d’électeurs évoqués dans l’article.

2. Pour la population “effective” couverte par les données, on peut distinguer deux aspects selon ce qu’on entend précisément par *population*, par opposition à *échantillon* (ce sera peut-être plus clair en introduisant quelques notations plus loin, voir le paragraphe §Formalisation).

D’une part, on peut être certain que la population couverte par les données n’est pas égale à la population d’intérêt 1. au sens où il y a eu 42,3% de participation seulement, soit 2,26 millions d’électeurs contre les 5,34 millions. En ce sens, l’échantillon obtenu par le scrutin ne correspond donc pas à la population d’intérêt ciblée. Mais, en général dans un cadre statistique, ce n’est pas ce qu’on demande à un échantillon : on ne lui demande pas de couvrir toute la population d’intérêt (c’est-à-dire d’être un recensement) mais d’être *représentatif de cette population d’intérêt*.²⁴

D’autre part, et surtout, au vu du résultat de 90% de “oui” parmi les votants et en ayant une certaine connaissance *auxiliaire*²⁵ sur la situation politique en Catalogne à cette époque, on peut fortement douter que l’échantillon soit *représentatif de la population d’intérêt ciblée*.

Il s’agit en effet des personnes résidant en Catalogne, de nationalité espagnole, âgées de plus 18 ans et ayant participé au scrutin. Or, au vu des conditions d’organisation de ce scrutin, on peut penser et argumenter que les personnes ayant participé au scrutin tendent à être davantage en faveur de l’indépendance de la Catalogne.

Formalisation Ce paragraphe cherche à formaliser rapidement les réponses précédentes. Ce n’est pas directement dans le cours d’Économétrie 1 mais c’est intéressant je pense plus largement pour votre formation statistique et économétrique. Vous reverrez par ailleurs cela de façon plus approfondie en Économétrie 2 au second semestre avec les modèles dit de sélection.

Pour un individu i quelconque de la population d’intérêt ciblée (un ou une Espagnole résidant en Catalogne âgée de 18 ans ou plus), notons :

- $Y_i \in \{0, 1\}$ l’indicatrice d’être en faveur de l’indépendance de la Catalogne $Y_i = 1$, 0 sinon ;
- $S_i \in \{0, 1\}$ l’indicatrice d’aller voter au référendum d’indépendance $S_i = 1$, 0 sinon.

Le scrutin nous fournit (Y_1, \dots, Y_n) avec n le nombre de suffrages exprimés ($n = 2,26$ millions ici).

²³Note : la correction ne prétend en rien discuter sur le fond du débat d’indépendance de la Catalogne, en particulier ici sur quels électeurs devraient pouvoir voter sur ce genre de question.

²⁴Dans cet exemple, cette discussion est rendue un peu plus compliquée, désolé pour cela, par le fait qu’il s’agit d’une élection. Typiquement, on souhaiterait qu’une élection soit un recensement de la population et non un sondage (voir par exemple les débats sur le fait d’essayer de rendre le vote obligatoire au moyen d’amendes en cas de non-participation), précisément car c’est un moyen, parmi d’autres, d’assurer la représentativité ; c’est la condition suffisante la plus triviale pour cela : faire en sorte que l’échantillon corresponde à toute la population d’intérêt. Bien sûr, la plupart du temps, c’est impossible ou trop coûteux. Dans un cadre statistique où on cherche à apprendre une caractéristique d’une population d’intérêt, il est tout à fait possible et légitime de le faire en utilisant seulement un échantillon de cette population, un échantillon plus petit que la population d’intérêt, un sondage par opposition à un recensement.

²⁵J’insiste sur ce terme d’information auxiliaire : on ne peut pas suspecter ce problème de sélection de l’échantillon, de non-représentativité uniquement à partir des chiffres de l’article, des résultats du scrutin. Il faut une information, une expertise, une connaissance à propos des données et non dans les données pour suspecter cela. Concrètement, il faut savoir par d’autres sources, sondages, articles, etc. qu’il est très peu probable que la proportion dans l’ensemble de la population catalane en faveur de l’indépendance soit autour de 90% en octobre 2017 et que la situation est vraisemblablement bien plus proche d’un 50-50 voire d’un pourcentage minoritaire en faveur de l’indépendance. Pour de mêmes raisons, les analyses de données “Big Data” et autres qui prétendent ne “faire parler que les données” sont en général des impasses ; il est indispensable d’avoir une certaine connaissance auxiliaire des données, de ce que représentent les variables, comment elles sont définies, collectées, etc. pour réaliser ensuite des analyses statistiques ou économétriques intéressantes.

On va modéliser ces variables comme indépendantes et identiquement distribuées (i.i.d.)²⁶. Les (Y_1, \dots, Y_n) sont ainsi i.i.d. de loi P_Y , la loi marginale inconditionnelle de Y , où, de la même manière que dans le cours, Y désigne une variable générique ayant la même distribution P_Y .

De même, et comme dans le cours où l'on omet parfois l'indice i , (Y, S) désigne un couple générique de variables aléatoires : $(Y, S) \sim P_{(Y,S)}$ où $P_{(Y,S)}$ désigne la loi jointe du couple (Y, S) . On peut noter dès à présent que, par construction, dans notre échantillon, on observe Y seulement si $S = 1$. Autrement dit, on a $S_i = 1$ pour tout $1 \leq i \leq n = 2,26$ millions ici.

Dans cette modélisation²⁷, le paramètre d'intérêt est la proportion de la population en faveur de l'indépendance : $\theta = \mathbb{E}[Y] = \mathbb{P}(Y = 1) = g(P_Y)$, pour exprimer que c'est une fonction de la distribution marginale P_Y de Y , (g est ainsi la fonction qui à une distribution de probabilité associe son espérance).

Pour estimer θ , on pourrait penser comme d'habitude à la moyenne empirique : $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$. Mais ici, **attention** : *on observe Y_i seulement si $S_i = 1$* . Formellement, les données (Y_1, \dots, Y_n) ne sont pas distribuées selon la loi marginale d'intérêt P_Y , mais selon la loi conditionnelle $P_{Y|S=1}$, la loi conditionnelle de Y sachant $S = 1$.

Par la loi des grands nombres (les données sont supposées i.i.d. et Y étant binaire, elle est bornée presque sûrement et admet donc un moment d'ordre 1), on a ainsi :

$$\frac{1}{n} \sum_{i=1}^n Y_i \xrightarrow[n \rightarrow +\infty]{P} g(P_{Y|S=1}) = \mathbb{E}[Y | S = 1].$$

Or, généralement, $\mathbb{E}[Y | S = 1] \neq \mathbb{E}[Y]$ dès lors qu'il existe une dépendance entre Y et S (voir Question 1 de ce quiz). On parle dans ce cas de **sélection endogène** : l'indicatrice S de sélection dans l'échantillon, c'est-à-dire d'observer la variable d'intérêt Y n'est pas indépendante de Y . De ce fait, sans autre hypothèse, on ne peut apprendre à partir de notre échantillon que sur la distribution conditionnelle $P_{Y|S=1}$ et non la distribution d'intérêt P_Y .

Au contraire, si $Y \perp\!\!\!\perp S$, on parle de **sélection exogène** ou encore de **sélection “ignorable”** (“*missing-at-random*”). Dans ce cas, on a en effet $P_{Y|S=1} = P_Y$ et on peut donc ignorer ce problème de **non-réponse** : les non-répondants (ceux qui ne sont pas allés voter au scrutin ici) ont la même distribution que les répondants, autrement dit, les répondants sont représentatifs de l'ensemble de la population, ils répondent en termes de loi de probabilité, de distribution, comme ceux qui ne répondent pas.

Rappel du slide 4, Chapitre 0 (Introduction) : “**Difficulté ici : faire le lien entre une situation réelle et les propriétés des variables aléatoires.**” *Cette phrase vous a peut-être semblé un peu abstraite au premier abord ou anodine, mais elle est très importante et au cœur de l'économétrie.* Une autre manière de le dire : c'est pourquoi vous avez le cours d'Économétrie 1 et pas seulement le cours de Statistique 1.

²⁶Derrière cette modélisation, que nous faisons constamment en statistique et en économétrie pour analyser des données, il faut comprendre la chose importante suivante : pour une personne i donnée, le fait qu'elle soit en faveur ($Y_i = 1$) ou contre ($Y_i = 0$) l'indépendance de la Catalogne n'est pas aléatoire, au sens où si on lui pose la question ou bien dans l'isoloir avant de voter, elle ne jette pas un dé, ne joue pas à pile ou face pour déterminer quel bulletin elle met dans l'enveloppe. *Du point de vue de cette personne i , sa décision n'est pas aléatoire. Elle est modélisée comme aléatoire du point de vue de l'économète ou du statisticien* qui, ne connaissant pas les raisons qui déterminent le choix de la personne, la modélise comme une variable aléatoire (la réalisation d'une variable aléatoire pour un échantillon donné). Lié à cela vous pouvez regarder la notion de *probabilité épistémique ou bayésienne* par opposition au cadre fréquentiste ; il y a de nombreuses références pour cela, par exemple : la **série** du Youtuber Lê Nguyén Hoang (Science 4 All) sur le bayesianisme ou certaines vidéos (**un exemple**) du Youtuber Thibaut Giraud (Monsieur Phi). Pour un sondage, c'est l'aléa instrumental qui détermine qui est interrogé et qui ne l'est pas qui justifie cette modélisation i.i.d.

²⁷On pourrait critiquer cette modélisation, notamment sur deux plans. D'une part, on peut douter de la possibilité de définir ainsi une telle variable Y faisant sens pour tout individu – voir par exemple cet exposé de Pierre Bourdieu de 1972, “*l'opinion publique n'existe pas*” ([lien vers le texte en ligne](#) – [lien vers le texte en version pdf](#)). D'autre part, on pourrait remettre en cause cette idée de construire un choix social, un choix collectif uniquement à partir de l'agrégation de préférences ou d'opinions individuelles données d'avance. Cette approche d'agrégation de préférences individuelles est très répandue dans nos démocraties représentatives, mais on peut y opposer l'idée qu'une démocratie pourrait (devrait même probablement) être plus que cela : il peut y avoir de la valeur dans la délibération, l'échange, la construction (par des procédures) d'une décision collective qui ne soit pas juste une agrégation d'opinions individuelles – voir par exemple les sondages dits “délibératifs” ou le modèle de la convention citoyenne pour le climat.

On a ici un exemple concret : Y et S sont nos variables aléatoires et la situation réelle est le contexte politique de la Catalogne à l'automne 2017 et les conditions d'organisation du scrutin du 1^{er} octobre. De ce fait, on peut en déduire certaines propriétés des variables aléatoires : il est très peu vraisemblable qu'on ait $Y \perp\!\!\!\perp S$, il est raisonnable de penser que $\mathbb{E}[Y | S = 1] > \mathbb{E}[Y] > \mathbb{E}[Y | S = 0]$.

Pour faire le lien entre les objets mathématiques de la formalisation et les réponses en mots du début de la correction :

- La *population d'intérêt ciblée*²⁸ : P_Y , la loi inconditionnelle marginale de Y .
- L'*échantillon* : les réalisations des variables Y_1, \dots, Y_n , où $n = 2,26$ millions dans cet exemple.
- La *population “effective” couverte par les données* est la distribution dans laquelle est tirée l'échantillon : $P_{Y|S=1}$, la loi conditionnelle de Y sachant $S = 1$.
- On dira qu'un échantillon est *représentatif d'une population d'intérêt donnée* s'il est tiré dans la distribution de cette population d'intérêt ; autrement dit, si les variables aléatoires de l'échantillon suivent la loi de la population d'intérêt.
- Ici, il y a un problème de sélection endogène des observations et l'échantillon n'est *pas* représentatif de la population d'intérêt puisqu'on peut raisonnablement penser que Y et S ne sont pas indépendants et donc que $P_Y \neq P_{Y|S=1}$. Dit autrement, la population effectivement couverte par les données (c'est-à-dire celle dont l'échantillon est représentatif) est distincte de la population d'intérêt ciblée.
- Une caractéristique de la population d'intérêt qu'on cherche à inférer à partir de l'échantillon, un *paramètre d'intérêt* : exemple ici $\theta = \mathbb{E}[Y] = g(P_Y)$.
- Un *estimateur* $\hat{\theta}$ du paramètre d'intérêt θ est une fonction des données (Y_1, \dots, Y_n) , dont le but est d'être utile pour apprendre (estimation ou inférence) quelque chose sur θ (ça ne marche pas toujours, ce n'est pas automatique) : exemple ici $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$, mais qui ne marche d'après ce qui précède.

Identification partielle Vous verrez la notion d'identification en Statistique 1 et en Économétrie 2. La notion d'identification *partielle* est hors du champ du cours d'Économétrie 1 ; elle est abordée dans un cours optionnel d'économétrie de troisième année. *Pour autant, il s'agit à mon avis d'une idée à la fois simple (aucune difficulté technique en termes de maths à ce stade notamment) et profonde, intéressante à connaître en tant qu'élève statisticien-économétro-économiste de l'ENSAE.* J'essaye ci-dessous d'introduire de manière informelle cette notion dans le cadre de cet exemple (voir le cours de troisième année pour plus de détails).

L'idée générale de l'identification est de négliger l'incertitude statistique due au fait de n'observer qu'un échantillon *fini* et de voir, dans cette situation idéale dans laquelle on observerait une “infinité” de données, ce qu'on peut apprendre sur le paramètre d'intérêt ciblé θ . En termes de distribution, au lieu d'observer seulement un n -échantillon (Y_1, \dots, Y_n) provenant d'une certaine loi de probabilité P , on considère qu'on observe, que l'on connaît la distribution P dans laquelle est tiré l'échantillon.²⁹

Si on peut écrire le paramètre d'intérêt θ comme une fonction connue de la distribution P , on dit que ce paramètre est *identifié*, parfois *ponctuellement identifié* (“point-identified” en langue anglaise).

²⁸Il y a encore une autre subtilité ici qui concerne la différence entre un cadre en “**population finie**” (toujours le cas à strictement parler en pratique bien sûr, ici 5,34 millions de personnes) et “**en population infinie**” ou “**en distribution**”. Sauf changement de programme, le TD6 de Statistique 1 présente ce cadre de population finie et fait le lien entre une **modélisation en population finie** et la **modélisation standard en population infinie** au moyen de distributions. Disons pour faire vite à ce stade qu'on peut très souvent utiliser le cadre “en population infinie” comme une approximation raisonnable ou une façon pratique de formaliser.

²⁹Derrière cette interprétation, on peut penser au théorème de Glivenko-Cantelli, parfois appelé “théorème fondamental de la statistique” : avec un nombre infini de données dans un modèle d'échantillonnage i.i.d., on peut déterminer exactement la loi des observations.

Dans notre exemple, $\mathbb{E}[Y | S = 1]$ est ponctuellement identifié, de même que toute autre caractéristique de la distribution $P_{Y|S=1}$, par exemple sa variance.

Le point à comprendre dans le cadre d'un problème de sélection est que, dans la perspective de l'identification, on obtient une infinité de données, mais ayant la *même loi* que notre échantillon. Ici, en particulier, cela ne revient pas à dire qu'on aurait une participation au scrutin de 100% car cela signifierait qu'on observe P_Y alors que notre processus générateur de données ("data-generating process", DGP, en anglais) donne un échantillon tiré dans la loi conditionnelle $P_{Y|S=1}$.

Ici, on ne peut pas obtenir $\theta = \mathbb{E}[Y]$ à partir de $P_{Y|S=1}$ dès lors que Y et S ne sont pas indépendants : le paramètre θ n'est donc pas ponctuellement identifié.

Pour autant, sans autre hypothèse, notre scrutin donne néanmoins certaines informations sur θ (voir ci-dessous) ; on dira que le paramètre θ est *partiellement identifié*.

Plus précisément, la description utilisée jusqu'à présent pour notre échantillon ne rend en fait pas compte de toute l'information obtenue : on apprend également des choses sur la distribution de S (entièrement résumée par son espérance $\mathbb{E}[S] = \mathbb{P}(S = 1)$ puisque S est une variable binaire).

En effet, si on classe dans cet ordre arbitraire (l'ordre n'a aucun impact) les observations, on a :

- pour $i = 1$ à $i = 2,02$ millions, $Y_i = 1$ et $S_i = 1$, ceux qui sont allés voter en faveur de l'indépendance ;
- pour $i = 2,02$ millions + 1 à $i = 2,26$ millions, $Y_i = 0$ et $S_i = 1$, ceux qui sont allés voter contre l'indépendance ;
- pour $i = 2,26$ millions + 1 à $i = 5,34$ millions, Y_i inconnu, non observé et $S_i = 0$, ceux qui ne sont pas allés voter.

On apprend en fait des choses sur $P_{Y|S=1}$ et sur P_S . Dans un raisonnement d'identification, on considère que l'on connaît ces deux distributions, comme si on les observait, comme si on observait un échantillon de taille infini.³⁰

Le problème est donc le suivant : en connaissant $\mathbb{E}[Y | S = 1] = 90\%$ et $\mathbb{E}[S] = \mathbb{P}(S = 1) = 42.3\%$, que peut-on dire sur $\mathbb{E}[Y]$?

Par la loi des espérances itérées et puisque $S \in \{0, 1\}$, on a :

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}[\mathbb{E}[Y | S]] \\ &= \mathbb{P}(S = 1)\mathbb{E}[Y | S = 1] + \mathbb{P}(S = 0)\mathbb{E}[Y | S = 0] \\ &= \mathbb{P}(S = 1)\mathbb{E}[Y | S = 1] + (1 - \mathbb{P}(S = 1))\mathbb{E}[Y | S = 0] \\ &= 0.423 \times 0.90 + (1 - 0.423) \times \mathbb{E}[Y | S = 0].\end{aligned}$$

Le terme $\mathbb{E}[Y | S = 0]$ est inconnu, même pire : il n'est pas identifié, ni ponctuellement, ni partiellement puisque que notre expérience (le scrutin) ne nous donne aucune information sur $P_{Y|S=0}$ car, par construction (c'est le problème de sélection justement), on observe Y que lorsque $S = 1$.³¹ Malgré cela, Y étant une variable binaire, forcément, en étant entièrement agnostique, sans faire aucune hypothèse, on sait néanmoins que

$$0 \leq \mathbb{E}[Y | S = 0] \leq 1.$$

Dans notre exemple, cela revient à dire qu'au pire ou au mieux, selon son point de vue, tous ceux qui ne sont pas allés voter sont contre l'indépendance $\mathbb{E}[Y | S = 0] = 0$ ou bien qu'ils sont tous pour l'indépendance $\mathbb{E}[Y | S = 0] = 1$.

³⁰Subtilité additionnelle ici : le cadre étant de population finie et le scrutin étant un recensement (et non un sondage), c'est en fait exactement ce qui se passe ici ! On observe véritablement S_i pour tous les i de notre population finie – voir la note de bas de page précédente sur la distinction entre une modélisation en population finie ou infinie. (Si ces deux notes sont incompréhensibles à ce stade, ce n'est pas grave, vous pourrez y revenir après le TD6 de Statistique 1.) C'est pourquoi je mets une égalité "—" (et non une approximation/estimation \approx) pour $\mathbb{E}[Y | S = 1] = 0.90$ et $\mathbb{E}[S] = 0.423$; les données observées sont bien les valeurs exactes de l'espérance conditionnelle et l'espérance théoriques dans ce cadre d'un scrutin d'une population finie.

³¹On pourrait par un autre moyen (un sondage téléphonique bien réalisé par exemple) chercher à obtenir un échantillon provenant de la loi $P_{Y|S=0}$ et ainsi estimer $\mathbb{E}[Y | S = 0]$.

On en déduit une *région* (ici un simple intervalle) d'identification (“identification set” ou “identified set” en anglais) pour le paramètre d'intérêt $\theta = \mathbb{E}[Y]$:

$$0.423 \times 0.90 + (1 - 0.423) \times 0 \leq \mathbb{E}[Y] \leq 0.423 \times 0.90 + (1 - 0.423) \times 1$$

$$0.3807 \leq \mathbb{E}[Y] \leq 0.9577.$$

On peut donc dire, sans aucune hypothèse, que ce scrutin permet d'affirmer que le pourcentage dans la population d'intérêt ciblée en faveur de l'indépendance de la Catalogne est compris entre 38% et 96%. *Ce n'est pas très informatif certes, mais, étant donné le problème de sélection endogène et la participation assez faible, c'est le plus honnête et juste intellectuellement qu'on puisse faire sans autre hypothèse !* C'est la force de cette approche par identification partielle. Remarquer qu'en bon statisticien, on ne donne jamais d'estimée sans intervalle de confiance³², avoir pour réponse un intervalle (et non juste une valeur) ne doit donc pas être un problème.

On pourrait affiner cet intervalle en acquérant de l'information sur $\mathbb{E}[Y | S = 0]$. Par exemple si un sondage ou d'autres raisons permettent d'encadrer :

$$0.05 \leq \mathbb{E}[Y | S = 0] \leq 0.40,$$

(en mots : dans ceux qui ne vont pas voter, au moins 5% et au plus 40% soutiennent l'indépendance), on obtiendrait

$$0.423 \times 0.90 + (1 - 0.423) \times 0.05 \leq \mathbb{E}[Y] \leq 0.423 \times 0.90 + (1 - 0.423) \times 0.40$$

$$0.40955 \leq \mathbb{E}[Y] \leq 0.6115.$$

Sachant qu'ici on cherche notamment à savoir si $\mathbb{E}[Y]$ est plus petit ou plus grand que 50%, la majorité, on peut faire un dernier calcul intéressant dans cette logique de l'identification partielle³³. Quel pourcentage minimum x (respectivement maximum) en faveur de l'indépendance parmi les personnes n'ayant pas voté faut-il pour être au-delà (respectivement en deçà) de 50% dans l'ensemble de la population d'intérêt ? On résout pour cela :

$$0.423 \times 0.90 + (1 - 0.423) \times x > 0.50 \iff x > \frac{0.50 - 0.423 \times 0.90}{1 - 0.423} \approx 0.207.$$

Si au moins 20.7% des non-répondants soutiennent l'indépendance, alors la majorité de la population d'intérêt est en faveur de l'indépendance. Inversement, si au maximum 20.7% des non-répondants sont en faveur de l'indépendance, on a alors $\mathbb{E}[Y] < 50\%$.

³²D'ailleurs, ici, avec un recensement dans une population finie, on n'a pas à se préoccuper d'inférence. Mais, en général, dans une approche d'identification partielle, il faut également faire de l'inférence sur la région d'identification, sur les bornes de cette région (voir le cours d'économétrie de 3A pour plus de détails).

³³C'est l'économétrien américain Charles F. Manski (né en 1948) qui a introduit et développé cette notion d'identification partielle par des articles puis par son livre *Partial Identification of Probability Distributions*, Springer-Verlag, 2003. Charles F. Manski est toujours actif et ses articles sont souvent intéressants ([site web personnel de Manski](#)).