Quiz 1 – Chapter 1: the fundamentals of linear regressions

(Lucas Girard) – This version: 2 October 2023

Questions

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

Some words about the quiz. 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 $\widehat{\beta}$? 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!
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.
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 $\mathbb{C}\text{ov}(X, \varepsilon)$ is well defined, in the sense that $\mathbb{C}\text{ov}(X, \varepsilon) < +\infty$.
Hint: Cauchy-Schwarz.

(b) Show that

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

that is, in words, ε mean-independent of X implies that ε and X are uncorrelated.
Hint: Law of Iterated Expectations.

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

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

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

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

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

2 OLS estimator with a single binary regressor

The OLS estimator $\widehat{\beta}_D$ of the slope coefficient in the simple linear regression of Y on[2] 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

1. $\widehat{\beta}_D = \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0]$
2. $\widehat{\beta}_D = \overline{Y}_1 - \overline{Y}_0$, where $\overline{Y}_d := \frac{1}{n_d} \sum_{i:D_i=d} Y_i$, and $n_d := \text{Card}(\{i : D_i = d\})$ for $d \in \{0, 1\}$
3. $\widehat{\beta}_D = \sum_{i=1}^n (D_i - \overline{D})(Y_i - \overline{Y}) \,/\, \sum_{i=1}^n (Y_i - \overline{Y})^2$
4. None of the above; if so, write the correct expression below

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 $\widehat{\varepsilon}$):

1. $\widehat{Y} = Y + \widehat{\varepsilon}$
2. $Y = \widehat{Y} + \widehat{\varepsilon}$

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

1. $\overline{\widehat{\varepsilon}} = ?$
2. $\widehat{\mathbb{C}\text{ov}}(D, \widehat{\varepsilon}) = ?$
3. $\widehat{\mathbb{C}\text{ov}}(G, \widehat{\varepsilon}) = ?$

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

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:

1. $R^2 = \widehat{\mathbb{V}}[Y] / \widehat{\mathbb{V}}[\widehat{Y}]$
2. The probability that the regression measures the causal effect of D on Y
3. $R^2 = \widehat{\mathbb{V}}[\widehat{Y}] / \widehat{\mathbb{V}}[Y]$
4. $R^2 = \widehat{\text{Corr}}(Y, \widehat{Y})$
5. None of the above; if so, write the correct expression below

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

[3] Instead, 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, $D \in \mathbb{R}^\Omega$.

[4] In 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.

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:

$$\widehat{\texttt{test\_ce2}} = 67.4 - 0.56\texttt{pc},$$
$$\widehat{\texttt{test\_ce2}} = 68.1 - 0.61\texttt{pc} - 11.4\texttt{red}.$$

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

1. The empirical covariance between `pc` and `red` is negative
2. The empirical covariance between `pc` and `red` is positive
3. We cannot conclude directly here: we should regress `pc` on `red`
4. We cannot conclude here because `pc` and `red` are perfectly colinear

6 Link between simple and multiple regressions (bis)

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),
- `age` is the age in years.

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

$$\widehat{\texttt{lnwage}} = 1.60 + 0.053\,\texttt{eduy},$$
$$\widehat{\texttt{lnwage}} = 0.95 + 0.063\,\texttt{eduy} + 0.015\,\texttt{age}.$$

(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
2. The empirical covariance between `eduy` and `age` is positive
3. The empirical covariance between `eduy` and `age` is necessarily null; otherwise, we could not do the multiple linear regression due to collinearity issues
4. We cannot conclude here: we should regress `eduy` on `age`

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.
`lnwage`	2.24	0.40
`eduy`	12.07	2.69
`age`	36.47	8.51

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

1. True
2. False

If so, compute that value:

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?

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

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 $\mathbb{P}(D = 1) \in (0, 1)$.[5] We denote by $(\widehat{\alpha}_D, \widehat{\beta}_D)$ 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 $\widehat{\beta}_D$

1. converges in probability to $\mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0]$
2. does not necessarily converge in probability: we need to further assume that D and Y are independent
3. does not necessarily converge in probability: we need to further assume that D and Y are uncorrelated
4. does not necessarily converge in probability: we need to further assume that $\mathbb{E}[D^2] < +\infty$

9 About a moment condition

With Chapter 1's notations, the condition $\mathbb{E}[XX']$ invertible is satisfied when[6]

1. we regress log(wage) on $1, \mathbb{1}\{man\}, \mathbb{1}\{woman\}$

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

[6] In 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.*

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
3. we regress log(wage) on 1, experience, and the square of the experience
4. we regress log(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

10 Properties of the OLS estimator

As in Chapter 1, let β_0 denote the probability limit of the OLS estimator $\widehat{\beta}$ 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.

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

(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 $x \mapsto \mathbb{E}[Y \mid X = x]$?

1. The components of X are independent
2. Y is a Gaussian variable and X is a Gaussian vector
3. β_0 corresponds to the causal effect of X on Y
4. None, it is always the case provided the previous three moment conditions hold

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

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

1. regress Y on X^2, then regress the residual of that first regression on X^1
2. regress X^2 on X^1, then regress Y on the residual of that first regression
3. regress X^1 on X^2, then regress Y on the residual of that first regression
4. simply regress Y on X^1 since we are only interested in the coefficient of X^1

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").*

$X = (X_0, X_1, \dots, X_{k-1})' \in (\mathbb{R}^k)^\Omega$ désigne[7] 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").

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, $\widehat{\beta}$, 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 prédite* (ou prédiction *empirique*) $\widehat{Y} := X'\widehat{\beta}$ qui utilise l'estimateur MCO, $\widehat{\beta}$, à la place de β_0 (β_0 est inconnu bien sûr, on cherche à l'estimer avec un échantillon de taille finie par $\widehat{\beta}$). On peut aussi la noter[8] $\mathbb{L}[Y \mid X] := X'\beta_0$ ou $\mathbb{E}_{\text{lin}}[Y \mid X] := X'\beta_0$ (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)[9] est la dérivée de l'application partielle qui à X_1 associe la prédiction linéaire théorique $\mathbb{L}[Y \mid X] := X'\beta_0$ (respectivement la prédiction empirique ou valeur prédite $\widehat{Y} := X'\widehat{\beta}$), les autres variables explicatives éventuelles[10], $X_2, \dots, X_{k-1}$, 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 :

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

Remarques : la première égalité (après la définition ":=") provient de la définition de la notation $\mathbb{L}[Y \mid X]$; 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),

[7] Ici, 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 numéroter 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 $X_{ij} \in \mathbb{R}^\Omega$.

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

[9] On précise parfois "l'effet marginal de X_1 **sur** $\boldsymbol{Y}$", et parfois non, en laissant alors implicite le fait qu'il s'agit de l'effet sur la variable expliquée étudiée Y.

[10] Notation : 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.

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) :

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

(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

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

ce qu'on peut écrire de façon équivalente[11]

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

(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

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

(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

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

(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

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

(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

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

[11] **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.

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`.

Figure 1: 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 1er 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 1),

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

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.