Quiz 3 – Chapter 3: Linear Regressions and Non Causal Predictions

(Lucas Girard) – This version:*8 January 2025. Happy New Year!

Solutions

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

Some words about the quiz. As always henceforth and absent contrary indication, the notation used follows that of the course's slides. $(Y_i, X_i)_{i\geq 1}$ is an i.i.d. sequence of random variables with the same distribution as a generic instance denoted (Y, X). We denote by k the dimension of X, $k := \dim(X)$. Remember that X is a random column vector of size $k \times 1$. We observe a sample $\mathcal{E}_n := (Y_i, X_i)_{i=1,\dots,n}$ of $n \in \mathbb{N}^*$ observations and an out-of-sample vector X_{n+1} of regressors. We are looking for the best (in terms of Mean Square Error, or, equivalently, L^2 norm) prediction of Y_{n+1} using a linear combination of X_{n+1}, that is, using $X'_{n+1}\,\beta(\mathcal{E}_n)$ as a prediction, where $\beta(\mathcal{E}_n)$ is a function of the data $\mathcal{E}_n$. We denote by f^* the conditional expectation of Y given X, that is, $f^*(x) = \mathbb{E}[Y \,|\, X = x]$ for any $x \in \text{Support}(X)$ in the support of X. By construction, $f^*(X)$, which is a real random variable, is the best prediction of Y from X in terms of Mean Square Error (MSE). *Beyond notations, try to be constantly aware of the nature of the objects they denote*: is it a non-stochastic parameter like β_0? Or an estimator, thus a random variable (since it is a function of the stochastic observations), like $\widehat{\beta}$? Likewise, be careful about the dimension of the objects (vectors, matrices, numbers) in computations.
As in the course slides, questions marked with an asterisk are of second-order importance.
Question 1 is a theoretical question to show a fundamental decomposition of the conditional prediction error.
Question 2 proposes to show two classical nonasymptotic results about the OLS estimator. Those two questions are open questions, more advanced compared to the others that are more direct application of the course; yet, they provide suitable training for theoretical-type questions (see third exercises in exams).
Questions 3, 4, and 5 are about the quality of prediction using different (methods to construct) predictors.
Question 6 is about cross-validation.
Question 7 deals with information criteria, comparing the two classical ones (AIC and BIC).
Questions 8 and 9 are about penalized regressions.
Finally, in a sort of motivation of Chapter 4, Question 10 tries to discuss the notion of *stable* or *unstable* environment.
Bonne lecture ! Do not hesitate if you have any questions.

1 A decomposition of the conditional prediction error

For any $\beta(\mathcal{E}_n)$ vector of size $k \times 1$ function of the n-sample $\mathcal{E}_n$, we denote by

$$\text{Err}(\beta(\mathcal{E}_n)) := \mathbb{E}[(Y_{n+1} - X'_{n+1}\beta(\mathcal{E}_n))^2] \qquad \text{(Uncond. Prediction Error)}$$

the prediction error in terms of MSE of using $\beta(\mathcal{E}_n)$ for a linear prediction of Y_{n+1} from X_{n+1}.

For any x in the support of X, we also denote the error conditional on the realization $X_{n+1} = x$ by

$$\text{Err}(\beta(\mathcal{E}_n), x) := \mathbb{E}[\big(Y_{n+1} - X'_{n+1}\beta(\mathcal{E}_n)\big)^2 \,|\, X_{n+1} = x]. \qquad \text{(Cond. Prediction Error)}$$

The main objective of this first question is to prove the following result (1) that decomposes the conditional prediction error, defined in (Cond. Prediction Error), into

- 1st term: an unexplained part that remains unpredictable;
- 2nd term: the prediction error with respect to the *oracle estimator*, that is, the square of the L^2 distance between the conditional expectation and the linear prediction of Y given $X_{n+1} = x$, which is $x'\beta(\mathcal{E}_n)$. Remember that, by construction, the conditional expectation is the best predictor in terms of MSE, but, in general, is unknown.

*Compared to the previous 28 November 2023 version: in 1.(a), solution of the additional question: how to approximate by Monte-Carlo (Cond. Prediction Error) to discuss further what is stochastic in the different objects we define and some explanations on why we do that.

[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.

THEOREM – Decomposition (unexplained part + "oracle error") of $\mathrm{Err}(\beta(\mathcal{E}_n), x)$

$$\text{For any } \beta(\mathcal{E}_n) \text{ and any } x,\ \mathrm{Err}(\beta(\mathcal{E}_n), x) = \underbrace{\mathbb{V}[Y_{n+1} \,|\, X_{n+1} = x]}_{1^{\text{st}} \text{ term}} + \underbrace{\mathbb{E}\Big[\big(f^*(x) - x'\beta(\mathcal{E}_n)\big)^2\Big]}_{2^{\text{nd}} \text{ term}} \qquad (1)$$

(a) (**Warm-up, but also crucial as a first reflex**). Elucidate what is stochastic, that is, what is or are the sources of randomness within the expectation in the definition of $\mathrm{Err}(\beta(\mathcal{E}_n))$. In other words, with respect to which distribution is the expectation in Equation (Uncond. Prediction Error) computed?

In the prediction error (in terms of mean-square error) of using $\beta(\mathcal{E}_n)$ for a linear predictor, $\mathrm{Err}(\beta(\mathcal{E}_n)) := \mathbb{E}[(Y_{n+1} - X'_{n+1}\beta(\mathcal{E}_n))^2]$, within that expectation, *everything* is stochastic:

- (Y_{n+1}, X_{n+1}) are stochastic: it is the out-of-sample observation that we are trying to predict;
- $\beta(\mathcal{E}_n)$ is also stochastic since it is a function of the sample $(Y_i, X_i)_{i=1,\dots,n}$, which is also stochastic: $\beta(\mathcal{E}_n) = g((Y_1, X_1), \dots, (Y_n, X_n))$ for some function g. For instance, if $\beta(\mathcal{E}_n)$ is the OLS estimator in the full/exhaustive regression model of Y on X, we have provided the usual invertibility condition (Inv) (see Chapter 1, slide 16),

$$\beta(\mathcal{E}_n) = g((Y_1, X_1), \dots, (Y_n, X_n)) = \left(\sum_{i=1}^n X_i X_i'\right)^{-1} \left(\sum_{i=1}^n X_i Y_i\right).$$

In terms of distribution, the expectation in (Uncond. Prediction Error) is thus computed with respect to $(\mathrm{P}_{(Y,X)})^{\otimes(n+1)}$, where the notation $\otimes$ denotes independent distributions: if two random variables A and B are independent, we write that the joint distribution $\mathrm{P}_{(A,B)}$ of the couple (A, B) is equal to the "product" of the two marginal distributions: $\mathrm{P}_A \otimes \mathrm{P}_B$. Here, remember that $(Y_i, X_i)_{i\geq 1}$ is an i.i.d. sequence with the same distribution $\mathrm{P}_{(Y,X)}$ as a generic instance (Y, X); hence, formally, $\mathrm{P}_{((Y_i,X_i)_{i=1,\dots,n},(Y_{n+1},X_{n+1}))} = (\mathrm{P}_{(Y,X)})^{\otimes(n+1)}$. In other words, there are $n+1$ random couples with the same distribution as a generic instance (Y, X) in this expectation:

$$\mathbb{E}[(Y_{n+1} - X'_{n+1}\beta(\mathcal{E}_n))^2] = \int \big[y_{n+1} - x'_{n+1} g\big((y_1, x_1), \dots, (y_n, x_n)\big)\big]^2 \, d\,\mathrm{P}_{(Y,X)}^{\otimes(n+1)}\big((y_1, x_1), \dots, (y_n, x_n), (y_{n+1}, x_{n+1})\big).$$

In contrast, consider the following quantity, the out-of-sample prediction error *conditional on a given sample*,

$$\mathbb{E}\big[\big(Y_{n+1} - X'_{n+1}\beta(\mathcal{E}_n)\big)^2 \,|\, Y_1 = y_1, X_1 = x_1, \dots, Y_n = y_n, X_n = x_n\big]. \quad \text{(Cond. } \textit{on sample} \text{ Pred. Error)}$$

In the latter conditional expectation, the only source of randomness comes from the out-of-sample observation (Y_{n+1}, X_{n+1}): we reason conditionally on the sample used to estimate $\beta(\mathcal{E}_n)$.

To be more concrete, imagine we were to approximate the values of (Uncond. Prediction Error) and of (Cond. *on sample* Pred. Error) respectively through Monte-Carlo simulations:

It is a good exercise! The solution is postponed to the next page so that you can (and should!) think about it before going to page 3.

- **For** (Uncond. Prediction Error): for each simulation $s = 1, \dots, S$
  - we draw an n-sample $(Y_i, X_i)_{i=1,\dots,n}$;
  - we compute the realization of $\beta(\mathcal{E}_n)$ on this sample;
  - then, we draw independently the new observation (Y_{n+1}, X_{n+1});
  - and compute the resulting realization of $(Y_{n+1} - X'_{n+1}\beta(\mathcal{E}_n))^2$;

  finally, we average those S realizations to approximate $\mathbb{E}[(Y_{n+1} - X'_{n+1}\beta(\mathcal{E}_n))^2]$.
- **For** (Cond. *on sample* Pred. Error): we fix (at some pre-determined values or by drawing the sample, but we draw it only *once*) an n-sample $Y_1 = y_1, X_1 = x_1, \dots, Y_n = y_n, X_n = x_n$, and we compute the realization of $\beta(\mathcal{E}_n)$ on this sample; then, for each simulation $s = 1, \dots, S$
  - we draw independently the new observation (Y_{n+1}, X_{n+1});
  - and compute the resulting realization of $(Y_{n+1} - X'_{n+1}\beta(\mathcal{E}_n))^2$;

  finally, we average those S realizations to approximate $\mathbb{E}[\big(Y_{n+1} - X'_{n+1}\beta(\mathcal{E}_n)\big)^2 \,|\, Y_1 = y_1, X_1 = x_1, \dots, Y_n = y_n, X_n = x_n]$.

It is important to understand the difference of nature, and, behind, of the sources of randomness between (Uncond. Prediction Error) *and* (Cond. *on sample* Pred. Error): **we should always wonder *what is stochastic in the expectations we consider?*** Remark that it is also the case for probabilities $\mathbb{P}$, variances $\mathbb{V}$, etc. If necessary, we could use a notation to specify the relevant distribution (see, for example, the notation of the form $\mathbb{E}_\theta$ in Statistics 1, or $\mathbb{P}_\mu$, $\mathbb{E}_i$ in Introduction to stochastic processes).

Remark that, for $x \in \text{Support}(X)$, (see (Cond. Prediction Error) recalled below)

$$\text{Err}(\beta(\mathcal{E}_n), x) := \mathbb{E}[\big(Y_{n+1} - X'_{n+1}\beta(\mathcal{E}_n)\big)^2 \,|\, X_{n+1} = x],$$

is yet *another* quantity with another source of randomness.

Exercise: specify the source of randomness, that is, with respect to which distribution is the expectation computed in the definition of $\text{Err}(\beta(\mathcal{E}_n), x)$?

Answer: We can do the same exercise for the quantity (Cond. Prediction Error): for a given $x \in \text{Support}(X)$, the support of X, in

$$\text{Err}(\beta(\mathcal{E}_n), x) := \mathbb{E}[\big(Y_{n+1} - X'_{n+1}\beta(\mathcal{E}_n)\big)^2 \,|\, X_{n+1} = x],$$

Y_{n+1} and $\mathcal{E}_n$ are stochastic/random while X_{n+1} is fixed at a given (arbitrary, but non-stochastic) value x among the set of possible values taken by the regressor X.

Why consider expectations? In practice, remember that we always observe *realizations* of the variables only: the realizations are not stochastic anymore; they are just some numbers. Yet, to analyze the prediction error of a given procedure, we are typically interested in its quality across different possible samples and values of the new regressors: formally, that is why we consider expectations and see some objects as stochastic to study how the procedure behaves on average over different possible realizations of the stochastic objects.[2]

[2] *Version alternative en français de ce paragraphe.* L'idée générale est que, lorsqu'on cherche à évaluer la qualité d'une certaine procédure pour faire des prédictions (par exemple faire des OLS sur l'échantillon d'entraînement puis utiliser les valeurs prédites à partir de cela), on n'est typiquement pas intéressé par ce qu'il se passe pour un échantillon donné et une valeur donnée du régresseur pour l'outcome à prédire à prédire, mais plutôt par ce qu'il se passe « en général » : en moyenne sur les différents échantillons possibles et valeurs possibles du régresseur. C'est pourquoi on définit les mesures de qualité avec ces espérances qui portent sur certains objets considérés comme stochastiques. Dans (Uncond. Prediction Error), on regarde la qualité en moyenne sur les différents échantillons et différentes valeurs possibles du régresseur. Dans (Cond. Prediction Error), on fixe la valeur du régresseur : on regarde la qualité en moyenne sur les différents échantillons de la prédiction.

Here, in (Cond. Prediction Error), we are interested in the quality of the prediction of a given *procedure* – that is, a way to compute β from the training sample $\mathcal{E}_n := (Y_i, X_i)_{i=1,\dots,n}$ – when the value of the new regressor (from which we predict the outcome) is x. We want to study that across the different possible samples $\mathcal{E}_n$ and across the various possible values for the outcome Y_{n+1}: hence the fact of considering those two objects stochastic and using a conditional expectation for the definition of $\mathrm{Err}(\beta(\mathcal{E}_n), x)$.

To better understand what is stochastic and what is not, again, you can imagine how we would approximate the value through Monte-Carlo simulations.

- **For** (Cond. Prediction Error): for each simulation $s = 1, \dots, S$
  - we draw an n-sample $(Y_i, X_i)_{i=1,\dots,n}$;
  - we compute the realization of $\beta(\mathcal{E}_n)$ on this sample;
  - then, we draw independently the outcome of the new observation Y_{n+1} conditional on $X_{n+1} = x$.
    Typically, we can do that by assuming a specific model stating how outcomes are obtained from regressors, say, for an example: $Y = 2 - 3X + X^2 + \nu$ with X univariate and an "error term"/"shock" $\nu \sim \mathcal{N}(0.1 \times |X|, X^2)$.
    With such a model, we follow these intermediary steps to implement the current step:
    * We draw one realization ν from a $\mathcal{N}(0.1 \times |x|, x^2)$;
    * We compute the realization of Y_{n+1} by
    $$Y_{n+1} = 2 - 3 \times x + x^2 + \nu.$$
  - and compute the resulting realization of $(Y_{n+1} - X'_{n+1}\beta(\mathcal{E}_n))^2$;

  finally, we average those S realizations to approximate $\mathbb{E}[\left(Y_{n+1} - X'_{n+1}\beta(\mathcal{E}_n)\right)^2 \mid X_{n+1} = x]$.

To be exhaustive, eventually, we can also consider the quantity

$$\mathbb{E}[\left(Y_{n+1} - X'_{n+1}\beta(\mathcal{E}_n)\right)^2 \mid Y_1 = y_1, X_1 = x_1, \dots, Y_n = y_n, X_n = x_n, X_{n+1} = x], \tag{2}$$

which is the out-of-sample prediction error conditional on a given sample when the value of the regressor from which we predict the outcome is set to some given $x \in \mathrm{Support}(X)$.

- **For** (2): we fix (at some pre-determined values or by drawing the sample, but we draw it only *once*) an n-sample $Y_1 = y_1, X_1 = x_1, \dots, Y_n = y_n, X_n = x_n$, and we compute the realization of $\beta(\mathcal{E}_n)$ on this sample; then, for each simulation $s = 1, \dots, S$
  - we draw independently the outcome of the new observation Y_{n+1} conditional on $X_{n+1} = x$ as in the corresponding step in (Cond. Prediction Error);
  - and compute the resulting realization of $(Y_{n+1} - X'_{n+1}\beta(\mathcal{E}_n))^2$;

  finally, we average those S realizations to approximate $\mathbb{E}[\left(Y_{n+1} - X'_{n+1}\beta(\mathcal{E}_n)\right)^2 \mid Y_1 = y_1, X_1 = x_1, \dots, Y_n = y_n, X_n = x_n, X_{n+1} = x]$.

(b) Prove the result (1). *It is an interesting exercise to train for theoretical-type questions.*

For any $\beta(\mathcal{E}_n) = g((Y_1, X_1), \dots, (Y_n, X_n))$ function of the n-sample $(Y_i, X_i)_{i=1,\dots,n}$ and any x in the support of X, we have

$$\begin{aligned}
\mathrm{Err}(\beta(\mathcal{E}_n), x) &\overset{(a)}{=} \mathbb{E}[\big(Y_{n+1} - x'\beta(\mathcal{E}_n)\big)^2 \mid X_{n+1} = x] \\
&\overset{(b)}{=} \mathbb{E}[\big(\{Y_{n+1} - f^*(x)\} + \{f^*(x) - x'\beta(\mathcal{E}_n)\}\big)^2 \mid X_{n+1} = x] \\
&\overset{(c)}{=} \underbrace{\mathbb{E}[(Y_{n+1} - f^*(x))^2 \mid X_{n+1} = x]}_{=:T_1} + \underbrace{\mathbb{E}[\big(f^*(x) - x'\beta(\mathcal{E}_n)\big)^2 \mid X_{n+1} = x]}_{=:T_2} \\
&\quad + \underbrace{2\,\mathbb{E}[(Y_{n+1} - f^*(x))\,\big(f^*(x) - x'\beta(\mathcal{E}_n)\big) \mid X_{n+1} = x]}_{=:T_3}
\end{aligned}$$

where the equalities come from:

(a) definition of $\mathrm{Err}(\beta(\mathcal{E}_n), x)$ and use of the conditioning event $X_{n+1} = x$;

(b) add and subtract $f^*(x)$;

(c) expand the square and linearity of the conditional expectation.

We consider the three terms T_1, T_2, and T_3 separately.

First term T_1 Regarding the first term T_1, we remark that

$$\begin{aligned}
\mathbb{E}[(Y_{n+1} - f^*(x))^2 \mid X_{n+1} = x] &= \mathbb{E}[(Y_{n+1} - \mathbb{E}[Y \mid X = x])^2 \mid X_{n+1} = x] \\
&= \mathbb{E}[(Y_{n+1} - \mathbb{E}[Y_{n+1} \mid X_{n+1} = x])^2 \mid X_{n+1} = x] \\
&= \mathbb{V}[Y_{n+1} \mid X_{n+1} = x] \ (= \mathbb{V}[Y \mid X = x]),
\end{aligned}$$

where the first equality comes from the definition of f^*, the second the fact that (Y, X) is a generic couple with the same distribution as (Y_{n+1}, X_{n+1}) (this is also why we have the last equality – just notations eventually), and the third is directly the definition of conditional variance (*link to Wikipedia*). Thus T_1 is indeed the first term we want; it corresponds to the unpredictable part of Y, the irreducible error in terms of MSE of predicting Y given only the knowledge of X.

Second term T_2 For the second term, we can suppress the conditioning because:

- $f^*(x) = \mathbb{E}[Y \mid X = x]$ is non-stochastic, it is a real number;
- $\beta(\mathcal{E}_n)$ is a function of $(Y_i, X_i)_{i=1,\dots,n}$ and, by assumption, X_{n+1} is independent of $(Y_i, X_i)_{i=1,\dots,n}$.

Consequently, we have

$$\mathbb{E}[\big(f^*(x) - x'\beta(\mathcal{E}_n)\big)^2 \mid X_{n+1} = x] = \mathbb{E}[\big(f^*(x) - x'\beta(\mathcal{E}_n)\big)^2],$$

which is the desired second term of the result: the difference, the error of using the linear predictor $x \mapsto x'\beta(\mathcal{E}_n)$ instead of the best predictor, which is the conditional expectation function $x \mapsto \mathbb{E}[Y \mid X = x]$.

Third term T_3 It remains to show that the third term, the double product from the square, is equal to 0. It is indeed the case because, remember again that $f^*(x)$ is non-stochastic, and (Y_{n+1}, X_{n+1}) are independent of the sample $\mathcal{E}_n$, hence of $\beta(\mathcal{E}_n)$. Therefore, the two real random variables, $Y_{n+1} - f^*(x)$ on the one hand, and $f^*(x) - x'\beta(\mathcal{E}_n)$ on the other are independent of each other conditionally on $X_{n+1} = x$.

Consequently, the expectation conditional on $X_{n+1} = x$ of their product is equal to the product of the conditional expectations, and we can write

$$\begin{aligned} &2\,\mathbb{E}[(Y_{n+1} - f^*(x))\left(f^*(x) - x'\beta(\mathcal{E}_n)\right) \mid X_{n+1} = x] \\ &\quad = 2\underbrace{\mathbb{E}[Y_{n+1} - f^*(x) \mid X_{n+1} = x]}_{=0}\mathbb{E}[f^*(x) - x'\beta(\mathcal{E}_n) \mid X_{n+1} = x] = 0, \end{aligned}$$

because

$$\begin{aligned} \mathbb{E}[Y_{n+1} - f^*(x) \mid X_{n+1} = x] &= \mathbb{E}[Y_{n+1} - \mathbb{E}[Y_{n+1} \mid X_{n+1} = x] \mid X_{n+1} = x] && \text{(definition of } f^*\text{)} \\ &= \mathbb{E}[Y_{n+1} \mid X_{n+1} = x] - \mathbb{E}[Y_{n+1} \mid X_{n+1} = x] && \text{(linearity of conditional expect.)} \\ &= 0. \end{aligned}$$

(c) Does this result rely on using *linear* predictions $X'_{n+1}\beta(\mathcal{E}_n)$? In other words, does it also holds for *any* prediction $f(X_{n+1}, \mathcal{E}_n)$ based on the data $\mathcal{E}_n$ and X_{n+1}?

In the previous proof, we do not use the fact that the prediction is linear. One key point is the fact that the prediction $x'\beta(\mathcal{E}_n)$ is independent of X_{n+1}. Yet, it is also the case for any prediction, possibly non-linear, $f(x, \mathcal{E}_n)$. As a consequence, the result also holds for any predictor f based on the data $\mathcal{E}_n$:

$$\mathrm{Err}(f, x) = \mathbb{V}[Y_{n+1} \mid X_{n+1} = x] + \mathbb{E}\Big[\left(f^*(x) - f(x, \mathcal{E}_n)\right)^2\Big], \qquad \text{(1 bis)}$$

where $\mathrm{Err}(f, x) := \mathbb{E}[(Y_{n+1} - f(X_{n+1}, \mathcal{E}_n))^2 \mid X_{n+1} = x]$ is the conditional prediction error of any function f used to predict Y_{n+1} from X_{n+1} and the sample $\mathcal{E}_n$.

(1 bis) is thus a generalization to non-linear predictors of the decomposition (1) of the prediction error conditional on $X_{n+1} = x$.

2 *Some nonasymptotic results about the OLS estimator (complements to the proof of Theorem 2 of Chapter 3)

This second question (more advanced or, rather, a bit orthogonal to the asymptotic viewpoint of Chapters 1 and 2 when studying the OLS estimator) aims at proving two classical nonasymptotic results about the OLS estimators.

These two results are used in the proof of Theorem 2 of Chapter 3: "We can then show that" in slides 12 and 13. Thus, this question complements the proof of that theorem.

The proof of Theorem 2 applies those results to the linear regression of Y on X^A, a subset of the entire vector X of regressors. However, here, we will state and prove the results in a generic set-up with standard notations looking at the regression of Y on X.

As in Chapters 1 and 2, we observe a sample $(Y_i, X_i)_{i=1,\dots,n}$ of i.i.d. data, with the same distribution as a generic instance (Y, X), where $Y \in \mathbb{R}^\Omega$ is a real random variable and $X \in (\mathbb{R}^k)^\Omega$ is a random column vector of dimension $k := \dim(X)$. $\widehat{\beta}$ denotes the OLS estimator in the linear regression of Y on X computed on the sample $(Y_i, X_i)_{i=1,\dots,n}$. We assume the standard moment conditions to define the theoretical regression properly: Y and X admit finite second-order moments, $\mathbb{E}[XX']$ is invertible.

Consequently, from Proposition 5 of Chapter 1, we already know that there exist a non-stochastic vector $\beta_0 := \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$ (the limit in probability of $\widehat{\beta}$) and a real random variable ε such that

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

Compared to the *asymptotic* results of Chapters 1 and 2, the two *nonasymptotic* (that is, valid for any sample size) results below require stronger moment conditions, which are presented in the statement of the results. Besides, the two results are *conditional on the realizations of the regressors*, whereas the main results of Chapters 1 and 2 (Proposition 5 of Chapter 1 and Theorem 1 of Chapter 2) hold

unconditionally. To denote this conditioning, we introduce the shortcut notation $\mathcal{X}_n := (X_1, \ldots, X_n)$, as in Chapter 3.

Henceforth, we also assume condition (Inv), slide 16 of Chapter 1:

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \text{ is invertible} \tag{Inv}$$

Indeed, we want to state nonasymptotic results (a.k.a. finite-sample results) on $\widehat{\beta}$: $\widehat{\beta}$ has to be well-defined! Remember that the moment conditions of Proposition 5 of Chapter 1 only ensure that $\widehat{\beta}$ is well defined with probability approaching one (w.p.o) as n goes to infinity.

PROPOSITION – (Conditional) Unbiasedness of the OLS estimator under a linear conditional expectation In addition to the previous moment and invertibility conditions (so-called "technical" conditions), if we assume that (which is a more substantial assumption compared to the previous ones)

$$\exists\, \beta_0 \in \mathbb{R}^{\dim(X)}, \exists\, \varepsilon \in \mathbb{R}^{\Omega} :\ Y = X'\beta_0 + \varepsilon \ \text{ with } \ \mathbb{E}[\varepsilon \,|\, X] = 0,$$

equivalently[3], if we assume a linear conditional expectation:

$$\exists\, \beta_0 \in \mathbb{R}^{\dim(X)} :\ \mathbb{E}[Y \,|\, X] = X'\beta_0,$$

then, the OLS estimator is unbiased conditionally on the regressors, namely

$$\mathbb{E}[\widehat{\beta} \,|\, \mathcal{X}_n] = \beta_0.$$

(a) Prove that proposition. If (Inv) is assumed to hold almost surely, then it is possible to show that the OLS estimator is (unconditionally) unbiased: $\mathbb{E}[\widehat{\beta}] = \beta_0$.

Show that $\mathbb{E}[\varepsilon \,|\, X] = 0$ implies $\mathbb{E}[X\varepsilon] = 0$. The assumption of a linear conditional expectation is thus stronger than the moment conditions of Proposition 5, Chapter 1.

The solutions can be found in Section 16 of Quiz 2, a former Problem Set of the course (questions 1 and 2 of the problem set).

PROPOSITION – Expression of the nonasymptotic conditional variance of the OLS estimator under a linear conditional expectation and strong homoscedasticity If, in addition to a linear conditional expectation and the technical conditions mentioned above, we assume strong homoscedastic error terms, namely

$$\exists\, \sigma_\varepsilon^2 \in \mathbb{R}_+ \text{ (a constant that does not depend on } X) :\ \mathbb{V}[\varepsilon \,|\, X] = \sigma_\varepsilon^2,$$

then

$$\mathbb{V}[\widehat{\beta} \,|\, \mathcal{X}_n] \;=\; \frac{\sigma_\varepsilon^2}{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \;=\; \sigma_\varepsilon^2\left(\sum_{i=1}^{n} X_i X_i'\right)^{-1}.$$

(b) Prove that proposition.

Show that the conditions of the proposition imply the homoscedasticity assumption (Hom) introduced in slide 10 of Chapter 2: $\mathbb{E}[\varepsilon^2 XX'] = \mathbb{E}[\varepsilon^2]\,\mathbb{E}[XX']$.

The solutions can be found in Section 16 of Quiz 2, a former Problem Set of the course (question 4 of the problem set). The last part of question (b) is an application of the Law of Iterated Expectations (a good exercise to check); see also the solutions for Question 13 of Quiz 2 with the discussion about strong homoscedasticity and (weak) homoscedasticity.

[3] See the solution of Quiz 2, Question 13 for further details.

3 Perfect prediction?

Imagine that, exceptionally (because it is *generally unknown*), we know the conditional expectation function $x \in \text{Support}(X) \mapsto f^*(x) = \mathbb{E}[Y \,|\, X = x]$, and we use it to predict Y_{n+1} from X_{n+1} by $f^*(X_{n+1})$.

Then, the resulting prediction error $\mathbb{E}[(Y_{n+1} - f^*(X_{n+1}))^2]$

Even if we knew (in general, f^* is unknown) the best predictor in terms of MSE, namely the conditional expectation function f^*, we would make prediction errors due to the unpredictable part, the first term in the right-hand side in the decomposition result of Equation (1).

Indeed, by the Law of Iterated Expectations for the first equality and the generalized decomposition result (1 bis) (with $f = f^*$) for the second equality, we have

$$\begin{aligned}\mathbb{E}[(Y_{n+1} - f^*(X_{n+1}))^2] &= \mathbb{E}\Big(\mathbb{E}[(Y_{n+1} - f^*(X_{n+1}))^2 \,|\, X_{n+1}] \Big) \\ &= \mathbb{E}\left(\mathbb{V}[Y_{n+1} \,|\, X_{n+1} = x] + \underbrace{\mathbb{E}\Big[(f^*(x) - f^*(x))^2 \Big]}_{=0} \right) = \mathbb{E}\big(\mathbb{V}[Y_{n+1} \,|\, X_{n+1}] \big).\end{aligned}$$

Hence, correct answer 3. Here, note that the training sample $\mathcal{E}_n = (Y_i, X_i)_{i=1,\dots,n}$ is irrelevant since we assume f^* is known. There is no point in using the sample $\mathcal{E}_n$ to approximate/estimate f^*.

1. is equal to 0, the prediction is perfect in this case as we know the "oracle" – **False**, there remains the unpredictable part.
2. tends to 0 as n goes to $+\infty$ – **False**, doubly false, since it is not zero and there is no point in considering n going to infinity here since we do not use the sample $\mathcal{E}_n$.
3. is equal to $\mathbb{E}\big(\mathbb{V}[Y_{n+1} \,|\, X_{n+1}]\big)$ – **True**.
4. tends to $\mathbb{V}[Y_{n+1} \,|\, X_{n+1}]$ as n goes to $+\infty$ – **False**, again we do not use the sample $\mathcal{E}_n$. Besides, remark that the notation "subscript $n+1$" is used to denote an observation that does not belong to $\mathcal{E}_n$ but has no intrinsic relationship with the size n of the sample $\mathcal{E}_n$: it could be replaced by $(Y_{\text{new}}, X_{\text{new}})$ for instance, to avoid the confusion between this notation and considering that the sample size n goes to infinity. Another way to say it: since (Y, X) denotes a generic instance with the same distribution, $\mathbb{V}[Y_{n+1} \,|\, X_{n+1}] = \mathbb{V}[Y_{\text{new}} \,|\, X_{\text{new}}] = \mathbb{V}[Y \,|\, X]$. The prediction error is equal, for any size n of $\mathcal{E}_n$, to $\mathbb{E}\big(\mathbb{V}[Y_{n+1} \,|\, X_{n+1}]\big)$. Hence, it can be said that it tends to $\mathbb{E}\big(\mathbb{V}[Y_{n+1} \,|\, X_{n+1}]\big) = \mathbb{E}\big(\mathbb{V}[Y \,|\, X]\big)$ as n goes to $+\infty$ although such an analysis is irrelevant here.

4 Better prediction with more covariates?

In the context of a stable environment, we predict Y_{n+1} by $X_{n+1}'\widehat{\beta}$ where, as in the course, $\widehat{\beta}$ is the OLS estimator of the linear regression of Y on X obtained from an i.i.d sample $(Y_i, X_i)_{i=1,\dots,n}$. If we include more regressors: for instance, we used first $X = (X^1, \dots, X^k)'$, and then, using $p \in \mathbb{N}^*$ other variables $X^{k+1}, \dots, X^{k+p}$ available from the database, we use $X = (X^1, \dots, X^k, X^{k+1}, \dots, X^{k+p})'$, the quality of the prediction

One of the critical points of Chapter 3 is that, in a prediction task, there is a trade-off between bias and variance. More complex models (with more parameters) are more flexible; hence more likely to be closer than f^*, which improves the quality of the prediction (*less bias*). On the other hand, given a *finite* sample of data (from which the predictor is "learned" or estimated), more complex models are harder to estimate, which can deteriorate the quality of the prediction (*more variance*).

In specific models, it is possible to formalize that general idea. Theorem 1 does so when using the OLS estimator on a subset of regressors. Theorem 2 specifies that result, giving more explicit

expressions for the different terms that make up the prediction error under further assumptions (linear conditional expectation, strongly homoscedastic error terms, and independent regressors). Although both consider $\beta(\mathcal{E}_n) = \widehat{\beta}^A$, the OLS estimator of Y on X^A (the subset of regressors indexed by $A \subseteq \{1, \dots, \dim(X)\}$), remark that Theorem 1 considers the unconditional prediction error (see (Uncond. Prediction Error) above) while Theorem 2 studies the conditional on the training sample prediction error (see (Cond. Prediction Error)).

When the set A of regressors gets larger, which corresponds here to including more regressors, the overall effect on the prediction error is ambiguous: the error can decrease (that is, the quality of the prediction increases) or can increase (the quality of the prediction decreases). This rules out the first three answers. Answer 4 is the correct one. The fact that we can use too many regressors in an out-of-sample prediction task is also illustrated with the example of slides 6 and 7 (The "exhaustive" OLS are not necessarily optimal).

1. cannot worsen; that is, it always weakly increases when adding regressors

   – **False**. It is false for the "prediction error", that is, without contrary indication, the *out-of-sample prediction error*: when trying to predict the value Y_{n+1} of the outcome from an *out-of-sample* regressor X_{n+1}; in contrast with the *in-sample prediction* which refer to the fitted values $\widehat{Y}_i := X_i\widehat{\beta}$ with $i \in \{1, \dots, n\}$ in the (training) sample used to estimate the predictor $\widehat{\beta}$.

   For the in-sample prediction error, that is,

   $$\mathbb{E}[(Y_i - X_i'\widehat{\beta})^2] = \mathbb{E}[(Y_i - \widehat{Y}_i)^2]$$

   when considering the i-th observation, or, on average over the sample[4],

   $$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\widehat{\beta})^2\right]$$

   the quality of the linear prediction from OLS always weakly increases when using more regressors. Indeed, by definition, the OLS estimator $\widehat{\beta}$ on the sample $\mathcal{E}_n := (Y_i, X_i)_{i=1,\dots,n}$ satisfies

   $$\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\widehat{\beta})^2 = \min_{\beta \in \mathbb{R}^{\dim(X)}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2.$$

   We thus obtain (weakly) lower values when using more regressors, that is, with larger $\dim(X)$. With the notation of the question, $\mathbb{R}^k \subset \mathbb{R}^{k+p}$, hence the minimum/infimum over $\mathbb{R}^{k+p}$ is necessarily weakly lower than the one over $\mathbb{R}^k$.

2. always worsens if the added regressors $X^{k+1}, \dots, X^{k+p}$ do not have a causal effect on the outcome variable Y

   False. This is another key point of Chapter 3: the prediction task is orthogonal to causality issues that will be introduced and formalized in Chapter 4 through the notion of potential outcomes.

3. always worsens if the added regressor $X^{k+1}, \dots, X^{k+p}$ are correlated with the other, previously used, regressors $X^1, \dots, X^k$

   False.

4. might worsen if the number $k + p$ of regressors becomes too large

   True.

[4] Note that we have both the theoretical expectation and the empirical counterpart, the average over the n observations. There are thus two means. (i) On a given sample $\mathcal{E}_n := (Y_i, X_i)_{i=1,\dots,n}$, we compute the realization of the OLS estimator $\widehat{\beta}$, use it to compute the fitted values $\widehat{Y}_i := X_i'\widehat{\beta}$, $i = 1, \dots, n$, and finally compute the average (quadratic) error over the sample: $n^{-1}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$. (ii) Then, we consider the average value over the different possible realizations of the sample, that is, for different draws $(Y_i, X_i)_{i=1,\dots,n}$ from $\mathrm{P}_{(Y,X)}^{\otimes n}$; hence the outer expectation.

5 Another method of prediction

In the context of a stable environment, we predict Y_{n+1} by $(X_{n+1}^{\widehat{A}})'\widehat{\beta}^{\widehat{A}}$, where, as in the course, for any vector U and set $A \subseteq \{1,\dots,k\}$, U^A denotes the sub-vector of U with coordinates A, and where $\widehat{A}$ maximizes[5] in $A \subseteq \{1,\dots,k\}$ the R^2 of the linear regression of Y on X^A. Then

Question 4 of Problem Set 2 (about the R-squared) shows that the R^2 mechanically weakly increases in the number of explanatory variables in the model.

Consequently, the set of covariates $\widehat{A}$ that maximises the R^2 of the regression is the exhaustive set $\widehat{A} = \{1,\dots,k\}$. With the notation of the course, it is $\widehat{A}_{\text{naive}} = \{1,\dots,k\}$ (slide 16).

Besides, the solution notes (those in English) already write: "For this reason, when concerned about the prediction of the outcome variable, to decide whether to include a new covariate in the model, it is not a good idea to consider whether the R^2 increases or not since it is mechanically the case $\longrightarrow$ [Transition to Question 5] a try to fix that issue: the adjusted R^2."

Indeed, as seen in Question 4 earlier in that Quiz, the "exhaustive" OLS are generally not the best out-of-sample predictor (even within the class of linear predictors in X). The result is also stated in slide 16 of Chapter 3: "$A^* \neq \{1,\dots,k\}$ in general".

Hence, correct Answer 4.

1. The computation of $\widehat{A}$ is fast even with numerous covariates, say, even if $\dim(X) = k \geq 50$

   – **False**. On the contrary, the computational time is exponential in k (see the first point in slide 25 of Chapter 3).

2. This method corresponds to AIC

   – **False**. The prediction method used does not correspond to AIC because AIC (or BIC) adds a penalization term that is absent here (see the definition of the corresponding minimization problem in slide 24).

3. We obtain $\widehat{A} = \{1,\dots,k\}$, which generally leads to the best (in terms of MSE) prediction of Y_{n+1}

   – **False**.

4. We obtain $\widehat{A} = \{1,\dots,k\}$, which generally does *not* lead to the best prediction of Y_{n+1}

   – **True**.

6 B-fold cross-validation

When used to select a subset A of regressors, B-fold cross-validation

The answer to this question directly comes from **Section 3** of Chapter 3 about "Cross validation", in particular **slide 19**.

1. splits the sample in B parts, $S_1,\dots,S_B$, and compare the estimator across the different subsamples S_b, for $b = 1,\dots,B$ – **False**, we do not consider the estimators but the estimated prediction errors:

$$\widehat{A}_{CV,B} := \underset{\{1\}\subset A\subset\{1,\dots,k\}}{\arg\min} \underbrace{\frac{1}{n}\sum_{b=1}^{B}\sum_{i\in S_b}\left(Y_i - X_i^{A\prime}\widehat{\beta}^A_{-b}\right)^2}_{=:\widehat{\text{Err}}_{CV,B}(A)}$$

2. selects the optimal subset A^* as long as $B = n$ ("leave-one-out cross-validation") – **False**.

[5] We assume here for simplicity that the maximum is unique so that $\widehat{A}$ is well-defined.

3. minimizes the prediction error asymptotically when n goes to $+\infty$ provided that $|S_b| \sim n$ for all $b = 1, \ldots, B$, where $|S_b|$ is the cardinal of the b-th subsample - **False**, it would be true if for all $b = 1, \ldots, B$, $|S_b| = o(n)$ (see the sixth bullet-point of slide 19).

4. is computationally costly since the model is estimated on several subsamples – **True** (see the last bullet-point of slide 19).

7 Information criteria

When using information criteria to select a subset of regressors or, more generally, to choose a model, the BIC (Bayesian Information Criterion) compared to the AIC (Akaike Information Criterion)

The answer to this question directly comes from **Remark 2 of slide 24**. Since the penalization of the BIC is larger than the one of the AIC if and only if $\ln(n) > 2$, that is, $n > e^2 \approx 7.4$, the BIC will choose more parsimonious models (in a weakly sense: the same model or a more parsimonious one) as long as $n \geq 8$, which is the case in virtually all applications; nonetheless, this is why there is the precision "in general" compared to "always" in the propositions.

1. always chooses more parsimonious models – **False**.

2. in general, tends to choose more parsimonious models – **True**.

3. always chooses less parsimonious models – **False**.

4. in general, tends to choose less parsimonious models – **False**.

8 Lasso regression

The Lasso regression, whose estimator is, for any $\lambda > 0$,

$$\widehat{\beta}_{\text{lasso}}(\lambda) := \underset{\beta \in \mathbb{R}^k}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i'\beta)^2 + \lambda \|\beta\|_1,$$

This question is directly related to the **slides 25 and 26 of Chapter 3** about the Lasso.

1. generally yields a sparse estimator, that is, with many components exactly equal to 0 – **True**, see the second bullet-point of slide 26.

2. is computationally demanding as the minimization problem is not convex – **False**, on the contrary, the problem is convex, which is one of the key interests compared to the non-convex optimization problem with the 0 "norm" (see the third bullet-point of slide 25).

3. cannot be solved if $k > n$, where $k = \dim(X)$ and n is the sample size – **False** (see the first point of slide 26).

4. is invariant to a re-scaling of the regressors – **False** (see the last point of slide 25): the norms 1 and 2 depend on the magnitude of β; hence on the units and scaling of the regressors. This is why it is vital to standardize the regressors when implementing penalized regressions.

9 Ridge regression

The Ridge regression, whose estimator is, for any $\lambda > 0$,

$$\widehat{\beta}_{\text{ridge}}(\lambda) := \underset{\beta \in \mathbb{R}^k}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i'\beta)^2 + \lambda \|\beta\|_2^2,$$

This question is directly related to the **slides 28 and 29 of Chapter 3** about the "Ridge" regression.

1. generally yields a sparse estimator, that is, with many components exactly equal to 0

   – **False**, although the penalization shrinks the coefficients of $\widehat{\beta}_{\text{ridge}}(\lambda)$ towards 0 compared to the coefficients of the standard not-penalized OLS estimator, no coefficient is set exactly to 0 generally (see the third bullet-point of slide 29).

2. admits an explicit solution $\widehat{\beta}_{\text{ridge}}(\lambda) = \left(\text{Id}_k + \frac{\lambda^2}{n}\sum_{i=1}^n X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i Y_i\right)$

   – **False**, $\widehat{\beta}_{\text{ridge}}(\lambda)$ does admit an explicit solution, but with the following expression (see the third bullet-point of slide 28):

$$\widehat{\beta}_{\text{ridge}}(\lambda) = \left(\lambda \text{Id}_k + \frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i Y_i\right).$$

   Even without knowing that correct expression, it remains possible to see that the proposition of Answer 2 is false because, when $\lambda = 0$ (no penalization, that is, standard OLS), it does not yield the expression of the OLS estimator $\widehat{\beta} = (\sum_{i=1}^n X_i X_i')^{-1}\sum_{i=1}^n X_i Y_i$.

3. enables to reduce the variance at the cost of increasing the bias by choosing *larger* hyper-parameters λ

   – **True** (see the first and second bullet-points of slide 29). As formalized by **Theorem 1** in slide 8 that decomposes the prediction error into (*1st term*) an unpredictable part + (*2nd term*) a "variance" term + (*3rd term*) a "bias" term, the key trade-off is between bias and variance; in other words, between under-fitting (too much bias compared to variance) and over-fitting (too much variance compared to bias).

   Under the assumption of $\mathbb{E}[Y \mid X] = X'\beta_0$, the OLS is unbiased (see the first proposition of Question 2). Yet, this comes at the cost of possibly large variance. For prediction tasks, it might be preferable to increase the bias and reduce the variance as the Ridge regression does: a larger λ means a larger penalization; hence a solution $\widehat{\beta}_{\text{ridge}}(\lambda)$ of the minimization problem that tends to be more shrunk towards $0_{\mathbb{R}^k}$, that is, with more bias compared to the unconstrained/not-penalized OLS estimator (which typically is never exactly null: $\widehat{\beta}_{\text{OLS}} \neq 0_{\mathbb{R}^k}$), but less variance.

   For an analogy (*concerned with the estimation of a parameter instead of prediction*), you can think about the classical Statistics 1 exercise related to the estimation of the support $[0, \theta]$ of a Uniform distribution: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Unif}([0,\theta])$, $\theta \in \mathbb{R}_+^*$.

   The Method of Moment estimator associated with the first order moment for θ is

$$\widehat{\theta}_{\text{MM}} := \frac{2}{n}\sum_{i=1}^n X_i = 2\overline{X}_n.$$

   $\widehat{\theta}_{\text{MM}}$ is unbiased, consistent, and asymptotically normal: $\sqrt{n}(\widehat{\theta}_{\text{MM}} - \theta) \underset{n\to+\infty}{\overset{d}{\longrightarrow}} \mathcal{N}(0, \theta^2/3)$.

   The Maximum Likelihood estimator for θ is

$$\widehat{\theta}_{\text{ML}} := X_{(n)} = \max\{X_1, \ldots, X_n\}.$$

   It is consistent, biased – it is possible to show that $\mathbb{E}_\theta[\widehat{\theta}_{\text{ML}}] = \frac{n\theta}{n+1} \neq \theta$ –, but it converges in probability to θ at a faster rate: n^{-1} compared to the standard $n^{-1/2}$ of the Central Limit Theorem, as we can show that $n(\theta - \widehat{\theta}_{\text{ML}}) \underset{n\to+\infty}{\overset{d}{\longrightarrow}} \mathcal{E}(1/\theta)$, an Exponential distribution with expectation equal to θ.

   Consequently, in typical cases where n is not too small, it is preferable to use $\widehat{\theta}_{\text{ML}}$: granted it adds some bias but reduces much more the variance compared to $\widehat{\theta}_{\text{MM}}$. Formally, in terms of

Mean Square Error of the two estimators, you can check that:

$$\mathrm{MSE}_\theta(\widehat{\theta}_{\mathrm{MM}}) := \mathbb{E}_\theta[(\widehat{\theta}_{\mathrm{MM}} - \theta)^2] = \frac{\theta^2}{3n} \quad \text{while}$$
$$\mathrm{MSE}_\theta(\widehat{\theta}_{\mathrm{ML}}) := \mathbb{E}_\theta[(\widehat{\theta}_{\mathrm{ML}} - \theta)^2] = \frac{2\theta^2}{(n+2)(n+1)} \underset{n\to+\infty}{\sim} \frac{2\theta^2}{n^2}.$$

Remember that the Mean Square Error is of the dimension of the variance: a square. To have the rates of the convergence in distributions, we rather look at the Root Mean Square Error (RMSE), that is, the proper L_2 distance (instead of the square of the distance) between the estimator and the target:

$$\mathrm{RMSE}_\theta(\widehat{\theta}_{\mathrm{MM}}) := \left(\mathbb{E}_\theta[(\widehat{\theta}_{\mathrm{MM}} - \theta)^2]\right)^{-1/2} = \frac{\theta}{\sqrt{3}\sqrt{n}} \quad \text{while}$$
$$\mathrm{RMSE}_\theta(\widehat{\theta}_{\mathrm{ML}}) := \left(\mathbb{E}_\theta[(\widehat{\theta}_{\mathrm{ML}} - \theta)^2]\right)^{-1/2} \underset{n\to+\infty}{\sim} \frac{\sqrt{2}\theta}{n}.$$

4. enables to reduce the variance at the cost of increasing the bias by choosing *lower* hyper-parameters λ

   – **False**, this is the reverse dependence for the impact of the hyper-parameter λ: the bias grows with λ and the variance decreases with λ.

10 *Unstable environment

As motivated in Chapter 3, non-causal predictions are related to prediction in a *stable* environment.

(a) Formally (that is, in terms of mathematical properties of some random variables), how would you define the contrary case: a prediction task in an *unstable* environment?

The formalization of a *stable* environment for a prediction task is the fact that the sample observations have the *same distribution* as the out-of-sample observation we are trying to predict:

$$\forall i \in \{1, \dots, n\}, (Y_i, X_i) \sim (Y, X) \sim (Y_{n+1}, X_{n+1}),$$

or, equivalently, expressed in terms of distributions,

$$\forall i \in \{1, \dots, n\}, \mathrm{P}_{(Y_i,X_i)} = \mathrm{P}_{(Y,X)} = \mathrm{P}_{(Y_{n+1},X_{n+1})}.$$

On the contrary, a prediction task in an *unstable* environment would be formalized by the fact that we have an i.i.d. training sample $(Y_i, X_i)_{i=1,\dots,n}$ (with, in particular, the same distribution as the first observation (X_1, Y_1)) but the distribution of the out-of-sample observation we want to predict is *different* from the distribution of the training observation we observe, that is

$$\forall i \in \{1, \dots, n\}, \mathrm{P}_{(Y_i,X_i)} = \mathrm{P}_{(Y_1,X_1)}, \text{ but } \mathrm{P}_{(Y_{n+1},X_{n+1})} \neq \mathrm{P}_{(Y_1,X_1)}.$$

(b) Explain what the difficulty is in this case and why considering causal relationships is essential compared to the setting of stable environments.

In this case, the issue is that, since the distributions are different, we cannot be sure that the relations we can learn/estimate between X and Y on the training sample, where (Y, X) have the joint distribution $\mathrm{P}_{(X_1,Y_1)} \overset{\text{noted}}{=:} Q$, will hold for the prediction since $(Y_{n+1}, X_{n+1}) \sim \mathrm{P}_{(Y_{n+1},X_{n+1})} \overset{\text{noted}}{=:} T \neq Q$ (note: the letter T used to denote the distribution of the observations we want to predict stands for "Target" here, not "Training").

Maybe the two distributions Q and T are not too far, in a sense to be made precise, so that we can hope that the relations we observe between X and Y on the sample $\mathcal{E}_n = (Y_i, X_i)_{i=1,\dots,n}$ can nonetheless be informative and useful to predict Y_{n+1} out of X_{n+1} (see the theme and literature about "transfer learning"). But maybe not!

In that setting of unstable environments, learning relationships between X and Y through the analysis of $\mathcal{E}_n$ is not always useful to predict Y_{n+1} out of X_{n+1}: those relationships might be linked to the particular environment; formally, they hold within the joint distribution Q of (Y_1, X_1) but might not exist anymore in another environment (in the distribution T of (Y_{n+1}, X_{n+1})).

In that sense and formalization, we could say that the links between Y and X that exist only within the joint distribution Q are *pure correlation* as opposed to *causal relationships* between X and Y, which are somewhat more substantial and assumed to exist both in the distributions Q and T. The causal relationships are useful and required to make accurate predictions as opposed to the correlations that only exist in the distribution Q but not in the target distribution T of the prediction task.

Of course, a criticism is that, in social sciences, "causal" relationships are also rooted in some specific environment (institutional, historical, cultural, etc.). Yet, the idea behind this is that the "causal" relationships remain, in some sense, more stable.

(Here may come in a future version a toy example composed of ice-creams, drownings, and an unstable environment with the recruitment of lifeguards. For the moment, you can imagine it, which is probably even a better exercise.)

11 *Some references (to watch if you want)

You will study these issues of predictions more in-depth in your second semester course, "Theoretical Foundations of Machine Learning" (+ various specific courses in the third year).

You can also look at the solutions (in English) of TD2 about the R-squared, notably the last question where there is a toy simulation of over-fitting compared to under-fitting.

There are numerous resources online about machine learning and (non-causal) predictions. Within the environment of French scientific "vulgarisateurs",

- Lê Nguyên Hoang (Youtube channel "Science4All", in French) has a whole series (54 episodes as of now) about artificial intelligence and machine learning (*link*);
- David Louapre (Youtube channel "Science Étonnante") made several outstanding (as always – excellent incentive, among others, to learn French!) videos about artificial intelligence and machine learning:
  - A general introduction to deep learning (*link*),
  - Can artificial intelligence be creative? It's a magic video about AlphaGo defeating Lee Sedol at the Go game (*link*),
  - How does IA understand our language? a video specific to Natural Language Processing (NLP) (*link*),
  - An application to biology: "Protein folding and DeepMind's AlphaFold A.I." (*link*).