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

(Lucas Girard) – This version: 7 November 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. $(Y_i, X_i)_{i\geq 1}$ 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, $f^*(x) = \mathbb{E}[Y \mid X = x]$ 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

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

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.

[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?

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

(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}?

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 $\mathcal{X}_n := (X_1, \dots, X_n)$, as in Chapter 3.

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

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

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[2], 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.

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: $\mathbb{E}[\varepsilon^2 XX'] = \mathbb{E}[\varepsilon^2]\,\mathbb{E}[XX']$.

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 $\mathbb{E}[(Y_{n+1} - f^*(X_{n+1}))^2]$

1. is equal to 0 (the prediction is perfect in this case as we know the oracle)
2. tends to 0 as n goes to $+\infty$
3. is equal to $\mathbb{E}\big(\mathbb{V}[Y_{n+1} \,|\, X_{n+1}]\big)$
4. tends to $\mathbb{V}[Y_{n+1} \,|\, X_{n+1}]$ as n goes to $+\infty$

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

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

1. cannot worsen; that is, it always weakly increases when adding regressors
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
3. always worsens if the added regressor $X^{k+1}, \dots, X^{k+p}$ are correlated with the other, previously used, regressors $X^1, \dots, X^k$
4. might worsen if the number $k + p$ of regressors becomes too large

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[3] in $A \subseteq \{1, \dots, k\}$ the R^2 of the linear regression of Y on X^A. Then

1. The computation of $\widehat{A}$ is fast even with numerous covariates, say, even if $\dim(X) = k \geq 50$
2. This method corresponds to AIC
3. We obtain $\widehat{A} = \{1, \dots, k\}$, which generally leads to the best (in terms of MSE) prediction of Y_{n+1}
4. We obtain $\widehat{A} = \{1, \dots, k\}$, which generally does *not* lead to the best prediction of Y_{n+1}

6 B-fold cross-validation

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

1. splits the sample in $B \geq 2$ parts, $S_1, \dots, S_B$, and compare the estimator across the different subsamples S_b, for $b = 1, \dots, B$
2. selects the optimal subset A^* as long as $B = n$ ("leave-one-out cross-validation")
3. minimizes the prediction error asymptotically when n goes to $+\infty$ provided that $|S_b| \sim n$ for all $b = 1, \dots, B$, where $|S_b|$ is the cardinal of the b-th subsample
4. is computationally costly since the model is estimated on several subsamples

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)

1. always chooses more parsimonious models
2. in general, tends to choose more parsimonious models
3. always chooses less parsimonious models
4. in general, tends to choose less parsimonious models

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

8 Lasso regression

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

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

1. generally yields a sparse estimator, that is, with many components exactly equal to 0
2. is computationally demanding as the minimization problem is not convex
3. cannot be solved if $k > n$, where $k = \dim(X)$ and n is the sample size
4. is invariant to a re-scaling of the regressors

9 Ridge regression

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

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

1. generally yields a sparse estimator, that is, with many components exactly equal to 0
2. admits an explicit solution $\widehat{\beta}_{\text{ridge}}(\lambda) = \left(\text{Id}_k + \frac{\lambda^2}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i Y_i \right)$
3. enables to reduce the variance at the cost of increasing the bias by choosing *larger* hyper-parameters λ
4. enables to reduce the variance at the cost of increasing the bias by choosing *lower* hyper-parameters 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?

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