Econometrics 1

Chapter 3 : Regressions and Non Causal Predictions

Xavier D'Haultfœuille and Elia Lapenta

Plan

The problem of non-causal prediction

- We look for the best prediction of Y_{n+1} by using the input X_{n+1} and an iid sample $(Y_i, X_i)_{i=1,\ldots,n}$
- Prediction in a « stable environment » :

$$(Y_{n+1}, X_{n+1}) \sim (Y_1, X_1) \sim (Y, X) \,.$$

- Cases where Y is observed after X and Y is at the basis of a decision :
  - $Y =$ amount reimbursed by a borrower : important for deciding whether to give a loan or not ;
  - $Y =$ future price of a stock as a function of economic variables ;
  - $Y =$ quantity of glucose in the blood for a diabetic (as a function of behavioral variables or genetic data)
- This is a general problem, but here we will focus on the prediction using linear models

Plan

- We consider a sequence $(Y_i, X_i)_{i\geq 1}$ of i.i.d. random variables. We observe $\mathcal{E}_n = (Y_i, X_i)_{i=1\dots n}$ and X_{n+1} but not Y_{n+1}.
- We look for the best prediction of Y_{n+1} by using a linear combination of X_{n+1} :
$$\arg\min_{\beta(\mathcal{E}_n)} E\left[\left(Y_{n+1} - X'_{n+1}\beta(\mathcal{E}_n)\right)^2\right], \quad (1)$$
where $\beta(\mathcal{E}_n)$ is a function of $\mathcal{E}_n$ (so it is random).
- Example : $\beta(\mathcal{E}_n) =$ OLS coefficient of Y on X in the sample $\mathcal{E}_n$.

The « exhaustive » OLS are not necessarily optimal

- Let us assume that $Y = 1 + \sum_{j=2}^{k} X^j/j + \varepsilon$, with $X^1 = 1$ and $(X^2, ..., X^k, \varepsilon) \sim \mathcal{N}(0, I_k)$.
- We seek to predict Y_{n+1} by $X_{n+1}^{\to j}\widehat{\beta}^{\to j}$, with $X^{\to j} = (X^1, ..., X^j)$ and $\widehat{\beta}^{\to j}$ OLS estimator of the regression of Y on $(X^1, ..., X^j)$.
- Estimation error as a function of j (here for $n = 30$ and $k = 10$) :

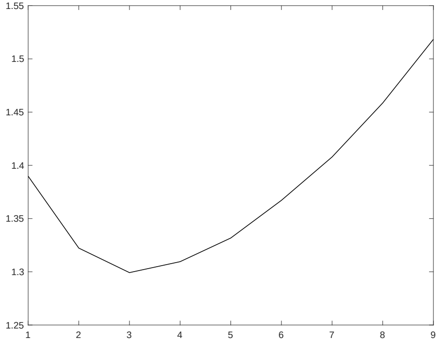

Figure 1 – $E\left[\left(Y_{n+1} - X_{n+1}^{\to j}\widehat{\beta}^{\to j}\right)^2\right]$ as a function of j.

The « exhaustive » OLS are not necessarily optimal

- Intuition :
  - adding explanatory variables initially allows to better explain Y.

  $\Rightarrow$ $j \mapsto E\left[\left(Y_{n+1} - X_{n+1}^{\to j}\widehat{\beta}^{\to j}\right)^2\right]$ initially decreases ;
  - but adding too many variables leads to « overfitting » : we adapt too much to the sample data that is random.

  $\Rightarrow$ This produces an imprecise estimator for the prediction function : $j \mapsto E\left[\left(Y_{n+1} - X_{n+1}^{\to j}\widehat{\beta}^{\to j}\right)^2\right]$ increasing for $j \geq j_0$.
- Questions :
  - Can we formalize and study mathematically such a trade off ?
  - How can we "optimally" select the explanatory variables to solve (1) ?

A first decomposition

- For any $A \subset \{1, ..., k\}$, $1 \in A$, and $x \in \mathbb{R}^k$, let $x^A = (x^j)'_{j \in A}$. We denote by $\widehat{\beta}^A$ the OLS estimator of Y on X^A computed on the sample $\mathcal{E}_n$.

Theorem 1

Let $f^*(x) = E(Y|X=x)$ and $Err(A) = E\left[(Y_{n+1} - X^{A\prime}_{n+1}\widehat{\beta}^A)^2\right]$. We have

$$Err(A) = E\left[(Y_{n+1} - f^*(X_{n+1}))^2\right] + E\left[\left(f^*(X_{n+1}) - X^{A\prime}_{n+1}E[\widehat{\beta}^A]\right)^2\right] + E\left[\left(X^{A\prime}_{n+1}(\widehat{\beta}^A - E[\widehat{\beta}^A])\right)^2\right]. \quad (2)$$

- Best prediction : $f^*(X_{n+1})$. But f^* is usually unknown !
- 2nd term : bias term, error from approximating $f^*(X_{n+1})$ with $X^{A\prime}_{n+1}E[\widehat{\beta}^A]$.
- 3rd term : « variance » of $X^{A\prime}_{n+1}\widehat{\beta}^A$. In general $\to 0$ when $n \to \infty$.
- In general, 2nd term $\downarrow$ when $A \uparrow$ (i.e., when A is a larger set). What about the 3rd term ?

$$\text{We have } \mathrm{Err}(A) = E\left[\left(Y_{n+1} - f^*(X_{n+1}) + f^*(X_{n+1}) - X_{n+1}^{A}{}'\widehat{\beta}^A\right)^2\right]$$
$$= E\left[(Y_{n+1} - f^*(X_{n+1}))^2\right] + E\left[\left(f^*(X_{n+1}) - X_{n+1}^{A}{}'\widehat{\beta}^A\right)^2\right]$$
$$+ 2E\left[(Y_{n+1} - f^*(X_{n+1}))\left(f^*(X_{n+1}) - X_{n+1}^{A}{}'\widehat{\beta}^A\right)\right].$$

- Moreover, the 3rd term T_3 satisfies

$$T_3 = 2E\left[(Y_{n+1} - f^*(X_{n+1}))\left(f^*(X_{n+1}) - X_{n+1}^{A}{}'\underbrace{E[\widehat{\beta}^A | X_{n+1}, Y_{n+1}]}_{=E[\widehat{\beta}^A]}\right)\right]$$
$$= 0 \ \text{ because } E\left[(Y_{n+1} - f^*(X_{n+1}))\, g(X_{n+1})\right] = 0 \text{ for any function } g.$$

- Thus,

$$\mathrm{Err}(A) = E\left[(Y_{n+1} - f^*(X_{n+1}))^2\right] + E\left[\left(f^*(X_{n+1}) - X_{n+1}^{A}{}'\widehat{\beta}^A\right)^2\right].$$

- Similarly,

$$
\begin{aligned}
& E\left[\left(f^*(X_{n+1}) - X_{n+1}^{A}{}'\widehat{\beta}^A\right)^2\right] \\
=& E\left[\left(f^*(X_{n+1}) - X_{n+1}^{A}{}'E[\widehat{\beta}^A] + X_{n+1}^{A}{}'\left(E[\widehat{\beta}^A] - \widehat{\beta}^A\right)\right)^2\right] \\
=& E\left[\left(f^*(X_{n+1}) - X_{n+1}^{A}{}'E[\widehat{\beta}^A]\right)^2\right] + E\left[\left(X_{n+1}^{A}{}'\left(E[\widehat{\beta}^A] - \widehat{\beta}^A\right)\right)^2\right] \\
& + 2E\left[\left(f^*(X_{n+1}) - X_{n+1}^{A}{}'E[\widehat{\beta}^A]\right) X_{n+1}^{A}{}'\left(E[\widehat{\beta}^A] - \widehat{\beta}^A\right)\right].
\end{aligned}
$$

- Here again, the 3rd term T_3' satisfies, by the Law of Iterated Expectations,

$$
\begin{aligned}
T_3' =& E\left[\left(f^*(X_{n+1}) - X_{n+1}^{A}{}'E[\widehat{\beta}^A]\right) X_{n+1}^{A}{}'\left(E[\widehat{\beta}^A] - E[\widehat{\beta}^A | X_{n+1}]\right)\right] \\
=& 0 \; \square
\end{aligned}
$$

- In order to have a simple expression of the 2nd and 3rd term, let us assume that

$$Y = X'\beta_0 + \varepsilon,\ E(\varepsilon|X) = 0,\ V(\varepsilon|X) = \sigma^2.$$

- Let's assume also that $(X^j)_{j=2,\dots,k}$ are mutually independent and $E(X^j) = 0,\ V(X^j) = 1$ for $j > 1$.

Theorem 2

Under the previous conditions and by denoting $\mathcal{X}_n^A = (X_1^A, \dots, X_n^A)$,

$$E\left[(Y_{n+1} - X_{n+1}^A{}'\widehat{\beta}^A)^2|\mathcal{X}_n^A\right] = \underbrace{\sigma^2}_{\text{1st term}} + \underbrace{\|\beta_0^{c_A}\|^2}_{\text{2nd term}} + \underbrace{\left(\sigma^2 + \|\beta_0^{c_A}\|^2\right)\frac{|A|}{n}}_{\text{3rd term}} + o_P\left(\frac{1}{n}\right),$$

with c_A *complementary set of* A, $|A| = \text{card}(A)$, *and* $o_P(1/n)$ *is a random variable* R_n *such that* $nR_n \xrightarrow{P} 0$.

$\Rightarrow$ The 3rd term has an ambiguous effect : when $A \uparrow$, $\sigma^2 + \|\beta_0^{c_A}\|^2 \downarrow$ but $|A| \uparrow$.

- The term $|A|$ is linked to overfitting.

Proof of Theorem 2*

- First, let us notice that the decomposition in (2) remains valid conditionally on $\mathcal{X}_n^A$:

$$\begin{aligned}\mathrm{Err}(A|\mathcal{X}_n^A) =& E\left[(Y_{n+1} - f^*(X_{n+1}))^2 \,|\mathcal{X}_n^A\right] + E\left[\left(f^*(X_{n+1}) - X_{n+1}^A{}' E[\widehat{\beta}^A|\mathcal{X}_n^A]\right)^2 |\mathcal{X}_n^A\right] \\ &+ E\left[\left(X_{n+1}^A{}'(\widehat{\beta}^A - E[\widehat{\beta}^A|\mathcal{X}_n^A])\right)^2 |\mathcal{X}_n^A\right]. \qquad (3)\end{aligned}$$

- Under the conditions of Theorem 2, $f^*(x) = x'\beta_0$, so the first term T_1 on the right hand side of (3) verifies

$$T_1 = E\left[\left(Y_{n+1} - X'_{n+1}\beta_0\right)^2 |\mathcal{X}_n^A\right] = E\left[\varepsilon_{n+1}^2|\mathcal{X}_n^A\right] = \sigma^2.$$

- Let us notice that

$$Y = X^{A\prime}\beta^A + \underbrace{X^{^cA\prime}\beta^{^cA} + \varepsilon}_{=\varepsilon^A}$$

- Moreover, since $(X^j)_{j=2,\ldots,k}$ are mutually independent, $E[\varepsilon^A|X^A] = E[\varepsilon^A] = 0$. We can then show that

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

- So, the 2nd term T_2 on the right hand side of (3) satisfies

$$\begin{aligned} T_2 =& E\left[\left(X_{n+1}^{cA\,\prime}\beta_0^{cA}\right)^2\right] \\ =& \beta_0^{cA\prime} E\left[X_{n+1}^{cA} X_{n+1}^{cA\,\prime}\right]\beta_0^{cA} \\ =& \|\beta_0^{cA}\|^2. \end{aligned}$$

- Finally, by the mutual independence of $(X^j)_{j=2,\ldots,k}$,

$$V[\varepsilon^A|X^A] = V[\varepsilon^A] = \sigma^2 + \|\beta_0^{cA}\|^2.$$

- We can then show that

$$V[\widehat{\beta}^A|\mathcal{X}_n^A] = \frac{\sigma^2 + \|\beta_0^{cA}\|^2}{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^A X_i^{A\prime}\right)^{-1}.$$

- Accordingly, for the 3rd term T_3 on the right hand side of (3) we have

$$
\begin{aligned}
T_3 &= E\left[X_{n+1}^{A}{}' V\left(\widehat{\beta}^A | \mathcal{X}_n^A, X_{n+1}^A\right) X_{n+1}^A | \mathcal{X}_n^A \right] \\
&= \frac{\sigma^2 + \|\beta_0^{cA}\|^2}{n} E\left[X_{n+1}^{A}{}' \left(\frac{1}{n}\sum_{i=1}^{n} X_i^A X_i^{A\prime} \right)^{-1} X_{n+1}^A \middle| \mathcal{X}_n^A \right] \\
&= \frac{\sigma^2 + \|\beta_0^{cA}\|^2}{n} \text{trace}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i^A X_i^{A\prime} \right)^{-1} E\left[X_{n+1}^A X_{n+1}^{A}{}' \middle| \mathcal{X}_n^A \right] \right] \\
&= \frac{\sigma^2 + \|\beta_0^{cA}\|^2}{n} \text{trace}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i^A X_i^{A\prime} \right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i^A X_i^{A\prime} + o_P(1) \right) \right] \\
&= \frac{\sigma^2 + \|\beta_0^{cA}\|^2}{n} \text{trace}\left[\text{Id}_{|A|} + o_P(1) \right] \\
&= \left(\sigma^2 + \|\beta_0^{cA}\|^2\right) \frac{|A|}{n} + o_P\left(\frac{1}{n}\right) \quad \square
\end{aligned}
$$

Plan

Problem with the naïve approach

- To obtain $A^* = \arg\min_{\{1\}\subset A\subset\{1,\dots,k\}} \mathrm{Err}(A)$, we can consider

$$\widehat{\mathrm{Err}}_{\text{naïf}}(A) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^{A\prime}\widehat{\beta}^A)^2,$$

$$\widehat{A}_{\text{naïf}} = \arg\min_{\{1\}\subset A\subset\{1,\dots,k\}} \widehat{\mathrm{Err}}_{\text{naïf}}(A).$$

- Problem : as R^2 grows with A , $\widehat{\mathrm{Err}}_{\text{naïf}}(A) \downarrow$ with A

$\Rightarrow$ We always have (if $k \leq n$) $\widehat{A}_{\text{naïve}} = \{1, ..., k\}$ while $A^* \neq \{1, ..., k\}$ in general.

- Origin of the problem : correlation between $\widehat{\beta}^A$ and (X_i^A, Y_i), while the $\widehat{\beta}^A$ in $\mathrm{Err}(A)$ is independent from (X_{n+1}, Y_{n+1}).

Uncrossed Validation

- With the *uncrossed* validation we ensure to recover such an independence.
- The main principle :
  - We split our sample in two : $S_1 \cup S_2 = \{1, ..., n\}$, $S_1 \cap S_2 = \emptyset$.
  - We estimate β^A only on $S_1 \Rightarrow \widehat{\beta}^A_{S_1}$ (S_1=training/estimation sample) ;
  - Then, we estimate $\mathrm{Err}(A)$ with $\widehat{\mathrm{Err}}_{S_2}(A) = \sum_{i \in S_2} (Y_i - X_i^{A\prime} \widehat{\beta}^A_{S_1})^2 / |S_2|$ (S_2 = validation sample).
  - We compute $\widehat{A} = \arg\min_{\{1\} \subset A \subset \{1,...,k\}} \widehat{\mathrm{Err}}_{S_2}(A)$.

Two-fold Cross Validation

- The previous approach has the drawback of introducing an asymmetry between the observations, according to whether they are in S_1 or S_2.
- We can recover the symmetry by exchanging the roles of S_1 and S_2, and then by aggregating the two errors :

$$\widehat{\mathrm{Err}}_{CV}(A) = \frac{1}{n}\left[\sum_{i\in S_2}(Y_i - X_i^{A\prime}\widehat{\beta}^A_{S_1})^2 + \sum_{i\in S_1}(Y_i - X_i^{A\prime}\widehat{\beta}^A_{S_2})^2\right].$$

- Thus, we compute $\widehat{A}_{CV} = \arg\min_{\{1\}\subset A\subset\{1,\dots,k\}} \widehat{\mathrm{Err}}_{CV}(A)$.
- This corresponds to the 2-fold Cross Validation.

- Generalization of the previous principle.
- Let $(S_1, ..., S_B)$ be a partition of $\{1, ..., n\}$.
- For $b = 1, ..., B$, we compute $\widehat{\beta}^A_{-b}$ on $\cup_{b' \neq b} S_{b'}$.
- We then minimize

$$\widehat{\text{Err}}_{CV,B}(A) = \frac{1}{n} \sum_{b=1}^{B} \sum_{i \in S_b} (Y_i - X_i^{A\prime} \widehat{\beta}^A_{-b})^2.$$

- Extreme case : cross validation of *one* against *all* (« leave-one out cross-validation ») : $B = n$ and $S_b = \{b\}$.
- We can show that asymptotically we minimize the prediction error if for all b, $|S_b|/n \to 0$ (as in the previous case)...
- ...But computationally costly !

Example

- Let us consider again the example

$$Y = \sum_{j=1}^{k} X^j / j + \varepsilon, \quad k = 10^2,$$

$$X^1 = 1, \quad (X^2, ..., X^k, \varepsilon) \sim \mathcal{N}(0, I_k).$$

- Here we only choose among the subsets of A having the form $\{1, \dots, j\}$.
- 5-fold cross validation.
- Minimal theoretical error$=V(\varepsilon)=$1 (normalized to 1).

Table 1 – Prediction error from C.V.

n	30	90	450
Prediction error	1.50	1.32	1.15

Example

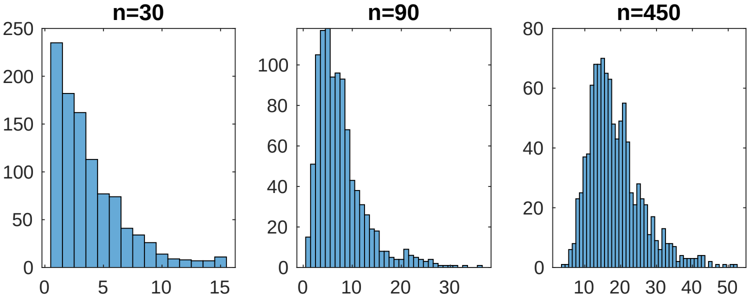

Figure 2 – Distribution of the number of regressors selected by CV

$\Rightarrow$ The number of selected regressors slowly increases with n.

Plan

Introduction

Prediction : general ideas

Cross Validation

Penalized Regressions

- Let us recall the initial problem :

$$\arg\min_{\beta(\mathcal{E}_n)} E\left[\left(Y_{n+1} - X'_{n+1}\beta(\mathcal{E}_n)\right)^2\right]. \tag{4}$$

- As we previously noticed, the naïve empirical counterpart

$$\arg\min_{\beta\in\mathbb{R}^k} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X'_i\beta\right)^2$$

is not satisfying, as it does not « penalize » for the « complexity » of β.

$\Rightarrow$ Modify the program (4) by introducing a penalty :

$$\arg\min_{\beta\in\mathbb{R}^k} \frac{1}{n}\sum_{i=1}^{n}(Y_i - X'_i\beta)^2 + f(\beta),$$

with $f(.)$ that « increases with the complexity » of β.

- Here, we are interested in $f(\beta) = \lambda\|\beta\|_p$ with $\lambda > 0$ and $p \in \{0, 1, 2\}$:

$$\|\beta\|_0 = \sum_{j=1}^{k} \mathbb{1}\left\{\beta_j = 0\right\}, \ \|\beta\|_1 = \sum_{j=1}^{k} |\beta_j|, \ \|\beta\|_2 = \left(\sum_{j=1}^{k} \beta_j^2\right)^{1/2}.$$

Penalization in « norm » 0

- In this case, we solve :

$$\arg \min_{\{1\} \subset A \subset \{1,\ldots,k\}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{X}_i^{A\prime} \widehat{\beta}^A)^2}_{\widehat{\sigma}_A^2} + \lambda |A|$$

- We obtain different estimators by different choices of λ. The two most popular choices are :
  - $\lambda = 2\widehat{\sigma}_A^2 / n$. This is equivalent to minimizing the Akaike Information Criterion (AIC) ;
  - $\lambda = \widehat{\sigma}_A^2 \ln(n)/n$. This is equivalent to minimizing the Bayesian Information Criterion (BIC).
- Remark 1 : The information criteria are developed for models estimated by maximum likelihood, but they can also be adapted to the present context.
- Remark 2 : since, in general, $\ln(n) > 2$, the BIC tends to choose more parsimonious models.
- The AIC chooses a model with minimal prediction error (the BIC has other theoretical advantages).

- Problem with the previous approaches : computational time exponential in k, as there are $2^k - 1$ possible models in A, and so $2^k - 1$ regressions to compute.
- If $k = 10^2$, approximately 1.3×10^{30} regressions !
- So, instead of an ℓ_0 penalization let us consider an ℓ_1 penalization :

$$\widehat{\beta}_{\text{lasso}}(\lambda) = \arg\min_{\beta \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i'\beta)^2 + \lambda \|\beta\|_1. \tag{5}$$

$\Rightarrow$ The program is now convex in β, and it can be quickly solved even if k is large.

- As compared to the OLS, the scale of X^j changes the prediction $\Rightarrow$ we prior standardize each component of X.

Advantages of Lasso

- It is possible to solve (5) even if $k > n$ (large dimensional problem).
- The solution will be « sparse » : many components of $\widehat{\beta}_{\text{lasso}}(\lambda)$ will be equal to 0.
- We will therefore have an automatic selection of the components of X.
- If only few X's have a significant effect on Y (« sparsity » condition), the Lasso will be almost optimal asymptotically.
- Formally, if $E(Y|X) = X'\beta_0$ with $||\beta_0||_0 = s_0$, then under some conditions on X we will have (for a certain constant C) :

$$E[(Y_{n+1} - X'_{n+1}\widehat{\beta}_{lasso})^2] \leq \sigma^2 + C\lambda^2 s_0.$$

- So, if $\lambda \to 0$ as $n \to \infty$, the Lasso estimator will tend towards the optimal prediction.
- Popular choices of λ : cross-validation.
(N.B. : We can simply compute $\widehat{\beta}_{\text{lasso}}(\lambda)$ for all λ)

- We consider again the model $Y = \sum_{j=1}^{k} X^j/j + \varepsilon, \quad k = 10^2,$
$$X^1 = 1,\ (X^2, \ldots, X^k, \varepsilon) \sim \mathcal{N}(0, I_k).$$
- Minimal theoretical error $= 1$. Error from Lasso (λ is chosen by C.V.) :

Table 2 – Prediction error from Lasso

n	30	90	450
Prediction error	2.63	2.33	2.16

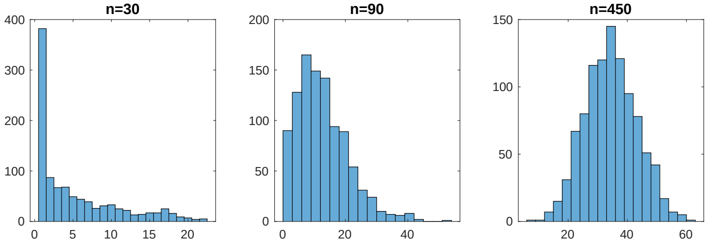

Figure 3 – Distribution of the number of regressors selected by Lasso.

Penalization in norm 2 : « Ridge » Regression

- We finally consider a penalization ℓ_2 :

$$\widehat{\beta}_{\text{ridge}}(\lambda) = \arg\min_{\beta \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i{}'\beta)^2 + \lambda \|\beta\|_2^2. \tag{6}$$

- As previously done, we prior standardize each component of X.
- This problem admits an explicit solution :

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

- Note : as well as for the Lasso, we can compute $\widehat{\beta}_{ri}(\lambda)$ even if $k > n$.

Advantages of the « Ridge » Regression

- Compared to the OLS, the « ridge » regression allows for a variance reduction, at the cost of introducing a bias.
- Let us assume that $E(Y|X) = X'\beta_0$ and $V(Y|X) = \sigma^2$, then :
  - the bias grows with λ ;
  - the variance decreases with λ.
- Compared to Lasso, no coefficient is set to 0 (but compared to the OLS, the coefficients are all shrunk towards 0).
- The estimator is consistent if $\lambda \to 0$ when $n \to \infty$. It can have a satisfying behavior even without sparsity conditions.

Example

- We consider again the model $Y = \sum_{j=1}^{k} X^j/j + \varepsilon, \quad k = 10^2,$
$$X^1 = 1,\ (X^2, ..., X^k, \varepsilon) \sim \mathcal{N}(0, I_k).$$
- Error from the Ridge regression (λ is chosen by C.V.) :

Table 3 – Prediction error from the Ridge regression

n	30	90	450
Prediction Error	2.57	2.53	2.25

- The results obtained in this example are comparable to those in the Lasso example.

Summary

- Non causal prediction : predict Y by X from a sample having the same law as (X, Y).

$\Rightarrow$ No interest in knowing if the coefficients of X represent a causal effect or not.

- Trade off between fit of the model ($\Rightarrow$ the model fits well f^*) vs stability (having too many parameters increases the variance of the estimators).
- Cross Validation : separation between training and validation sample, exchange their roles.
- Penalized Regression :
  - Information criteria (« norm » 0 penalization).
  - Lasso (norm 1 penalization).
  - Ridge (norm 2 penalization).