Statistical learning for regression: beyond ordinary least squares

Mathieu Rosenbaum

École Polytechnique

Fall 2023

Plan

Introduction

Linear regression model

- The linear regression model is the most classical one enabling us to forecast a variable Y from explanatory variables X.
- We set $Y = X\theta + \xi$.
- Y is $(n, 1)$ observed, X is (n, p), deterministic, observed, θ is $(p, 1)$ to be estimated and ξ is a non-observed $(n, 1)$ random noise vector such that $\mathbb{E}[\xi] = 0$ and $\text{Var}[\xi] = \sigma^2 I_n$.
- For example, Y represent the returns of some companies and X some accounting ratios.
- We usually estimate θ by OLS.

Plan

OLS and Gauss Markov theorem

Definition and properties

- We assume $\text{Rank}(X) = p$.
- The OLS estimator is defined by $\widehat{\theta} = (X'X)^{-1}X'Y$.
- $\widehat{\theta}$ is an unbiased estimator of θ and $\text{Var}[\widehat{\theta}] = \sigma^2(X'X)^{-1}$.
- Gauss Markov theorem : $\widehat{\theta}$ is the linear unbiased estimator of θ with minimal variance.
- Explanations and proofs on tablet.

OLS and Gauss Markov theorem

Gaussian case

- We assume in addition that ξ is a Gaussian vector.
- $\widehat{\theta} \rightsquigarrow \mathcal{N}(\theta, \sigma^2 (X'X)^{-1})$.
- We define $\widehat{\xi} = Y - X\widehat{\theta}$ and $\widehat{\sigma^2} = (\widehat{\xi}'\widehat{\xi})/(n-d)$.
- $\widehat{\sigma^2}$ is an unbiased estimator of σ^2.
- $\frac{\widehat{\theta}-\theta}{\sqrt{\widehat{\sigma^2}(X'X)^{-1}_{jj}}} \rightsquigarrow \mathcal{T}(n-d)$.
- This enables us to do student tests for significance of a coefficient.
- Explanations and proofs on tablet.

Plan

Pitfalls of the OLS

Despite the previous results, it is sometimes better not to use OLS.

Limitations of the OLS

- They favor bias over variance (bias$=0$). Thus the variance of the coefficients can be large, which is a problem when we want to do forecasting.
- OLS do not give any zero coefficient. No variable selection. Hence we obtain a lot of coefficients in the model, which makes it hard to interpret.
- $(X'X)$ has to be invertible and in particular we need $n \geq p$, which is not the case in many examples.

Pitfalls of the OLS

To correct the second limitation we can use the following empirical approach.

Empirical variables selection

- We estimate the model with all explanatory variables and compute the $p-$values for Student test of all the coefficients. We also compute $\|Y - X\widehat{\theta}_p\|$.
- We suppress the coefficient with the largest $p-$value and re-estimate the model without the variable associated to this coefficient. We compute again $p-$values and $\|Y - X\widehat{\theta}_{p-1}\|$.
- We continue the procedure and stop when $\|Y - X\widehat{\theta}_{p-k}\|$ starts increasing significantly.
- Other approaches : Mallows C_p, AIC, BIC criterion....

Plan

Ridge regression

Introduction and definition

- Enables us to correct the first and third limitations.
- If $p > n$, we cannot use OLS : $\min_{\theta} \|Y - X\theta\|^2 = 0$ and infinity of solutions.
- More fundamentally, we may want to control the estimated θ_j to avoid them being too big (even if $p \leq n$).
- We penalise the usual OLS distance and minimise $F(\theta) = \|Y - X_\theta\|^2 + \lambda\|\theta\|^2$.
- λ is a penalty parameter to be chosen by he user. $\lambda \to 0$: $\widehat{\theta}_{ridge} = \widehat{\theta}_{ols}$; $\lambda \to +\infty$: $\widehat{\theta}_{ridge} = 0$.

Ridge regression

Explicit computations

- We obtain $\widehat{\theta}_{ridge} = (X'X + \lambda I_p)^{-1} X'Y$.
- $\mathbb{E}[\widehat{\theta}_{ridge}] = (X'X + \lambda I_p)^{-1} X'X\theta$.
- $\widehat{\theta}_{ridge}$ is biased.
- $\mathrm{Var}[\widehat{\theta}_{ridge}] = \sigma^2 (X'X + \lambda I_p)^{-1} X'X (X'X + \lambda I_p)^{-1}$.
- $\widehat{\theta}_{ridge}$ can be superior than $\widehat{\theta}_{ols}$ in term of quadratic loss.
- Explanation and proofs on tablet.

Ridge regression

How to choose λ ? Cross-validation

- The optimal λ depends on θ which is unknown.
- We use cross-validation.
- We split data in K subsets of equal size (typically $K = 5$ or $K = 10$).
- For many possible λ, for each k, we compute $\widehat{\theta}_{-k}(\lambda)$ over all the data except those of subset k.
- We compute the error of this estimator in the sense of empirical prediction on the subset k :

$$(CV_{error})_k^{(\lambda)} \frac{1}{\text{Card subset } k} \sum_{Y_j, X_{j\cdot} \text{ in subset } k} (Y_j - X_{j\cdot}\widehat{\theta}_{-k}(\lambda))^2.$$

- Total cross validation error : $\frac{1}{K}\sum_{k=1}^{K}(CV_{error})_k^{(\lambda)}$.
- We pick λ minimising this quantity.

Plan

Lasso estimator

Introduction and definition

- Least absolute shrinkage and selection operator (Tibshirani 1996).
- We replace the l_2 norm of the ridge estimator by an l_1 norm.
- $\widehat{\theta}_{lasso} = \underset{\theta}{\text{argmin}} \|Y - X\theta\|^2 + \lambda \sum_{j=1}^{p} |\theta|_j$.
- This is equivalent to $\underset{\theta}{\text{argmin}} \|Y - X\theta\|^2$ under the constraint $\sum_{j=1}^{p} |\theta|_j \leq t$.
- This is equivalent to $\underset{\theta}{\text{argmin}} \sum_{j=1}^{p} |\theta|_j$ under the constraint $\|Y - X\theta\|^2 \leq t'$.

Lasso estimator

Interest of the Lasso

- We are often looking for sparse solutions.
- The natural idea is to consider $\underset{\theta}{\text{argmin}} \|Y - X\theta\|^2$ under the constraint $\|\theta\|_{l_0} \leq t$, where the l_0 norm denotes the number of non-zero coefficients.
- Not doable algorithmically : NP-hard problem.
- Why taking the l_1 norm instead ?
- It makes the problem feasible : $q = 1$ is the smallest q making the constraint region $\sum_j |\theta_j|^q$ convex. Non convex problem ($q < 1$) are hard to deal with. With $q = 1$ very fast algorithm such as LARS.
- The l_1 norm promotes sparsity : If we minimise an l_1 norm over points satisfying an l_2 constraint, we are likely to get sparse solutions.
- Explanations on tablet.