Econometrics 1

TD 5: non-causal predictions

(Chapter 3)

This exercise aims at getting the best prediction of a protein level, which is a biomarker for prostate cancer. We use Stamey et al. (1989)[1]'s data, following the steps of Hastie et al. (2009)[2]. The outcome of interest is the logarithm of the prostate specific antigen levels (`lpsa`, denoted by Y from now on). The vector of covariates X includes the logarithm of the cancer volume (`lcavol`), the logarithm of the prostate weight (`lweight`), the patient's age (`age`), the logarithm of benign prostatic hyperplasia levels (`lbph`), an indicator for seminal vesicle invasion (`svi`), the logarithm of capsular penetration levels (`lcp`), Gleason score (`gleason`) and the share of Gleason scores that are equal to to 4 or 5 (`pgg45`). Finally, the variable `train`, which was generated by Hastie et al., separates the training sample (`train=T`) from the validation sample (`train=F`). We suppose that $(X_i, Y_i)_{i=1...n}$ is an i.i.d random sample and we denote by $\mathcal{T}$ (of size $n_{\mathcal{T}}$) the training sample.

1. Regress `lpsa` on the whole set of covariates using the training sample. Compute the prediction error on the training sample and on the validation one. Compare these results with the ones you get from the model that only includes a constant as vector of covariates.

2. Let $\widetilde{Y}_i$ be a random variable that is independent from Y_i and such that $\widetilde{Y}_i|X_i$ follows the same distribution as $Y_i|X_i$. Let :

$$\begin{aligned} \text{Err}_{\text{train}} &:= \frac{1}{n_{\mathcal{T}}} \sum_{i\in\mathcal{T}} (Y_i - X_i'\widehat{\beta})^2, \\ \text{Err}_{\text{in}} &:= \frac{1}{n_{\mathcal{T}}} \sum_{i\in\mathcal{T}} E[(\widetilde{Y}_i - X_i'\widehat{\beta})^2 | (X_i, Y_i)_{i\in\mathcal{T}}]. \end{aligned}$$

   Explain these different notions and notations. We also define

$$\text{op} := \text{Err}_{\text{in}} - \text{Err}_{\text{train}}$$

   Show that

$$E[\text{op} \,|\, (X_i)_{i\in\mathcal{T}}] = \frac{2}{n_{\mathcal{T}}} \sum_{i\in\mathcal{T}} \text{cov}(Y_i, \widehat{Y}_i | (X_i)_{i\in\mathcal{T}}).$$

[1] cf. Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E. and Yang, N. (1989). "Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II radical prostatectomy treated patients", Journal of Urology (16), 1076–1083.

[2] Hastie, Tibsirani et Friedman (2009), "Elements of Statistical Learning", Springer Series in Statistics.

Based on this result, explain why we talk about optimism for "op" and comment your results for Question 1.

3. Install the `leaps` package in R by executing `install.packages("leaps").` You can find more information about the package by using `?leaps`.

   (a) Use `leaps` to determine the optimal sub-models of all sizes.

   (b) Identify the best model based on the error in the validation sample.

   Explanation of `leaps`: *The `leaps` package in R allows for the selection of the most predictive variables for a target variable. The selection is based on information criteria such as BIC, AIC, and adjusted R^2, which help analyze the added value of each new variable.*

   To obtain the best sub-models with up to p variables, use the `regsubsets` function from `leaps`, specifying the number of variables you wish to include. For each sub-model, information criteria can be calculated to identify the most informative variables.

   For more information about the package: https://cran.r-project.org/web/packages/leaps/leaps.pdf

4. What is the optimal model according to AIC? According to BIC? Does the difference between the two make sense?

5. We want to perform Lasso or ridge regression. Why is it desirable to standardize the variables beforehand? Perform this standardization.

6. Install the `glmnet` package in R by executing the following command: `install.packages`. Then use this package to estimate a LASSO model regressing Y on X, and produce a graph displaying the estimated LASSO coefficients as a function of $\ln(\lambda)$, where λ is the penalty parameter.

   (a) Perform a LASSO regression of Y on X using `glmnet`.

   (b) Plot a graph showing the evolution of the estimated coefficients as a function of $\ln(\lambda)$.

   (c) Comment on the results.

   To learn more about the `glmnet` package: check the following links:

   - Documentation for the glmnet package
   - Introduction to LASSO with `glmnet`

7. What is the optimal penalty parameter λ obtained through cross-validation using the `glmnet` package with the `cv.glmnet` option in R? Comment on the differences between the estimated LASSO coefficients and those obtained from an OLS regression (post-LASSO) using only the variables selected by the LASSO model.

   (a) Use cross-validation to determine the optimal value of λ using `cv.glmnet`.

   (b) Comment on the differences between the "Lasso" and "Post-OLS" columns.

8. What prediction error on the validation sample does this λ correspond to? Does it minimize the prediction error on this sample, and why?

9. Perform ridge regression by choosing the penalty parameter through cross-validation. Compare the results to those of the Lasso from question 8, and the OLS of the full model.