Series 9

1. This exercise is taken from Section 6.2.2 at pages 224 – 226 of the ISL book (G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning - with Applications in R*, 2017).

We want to compare the lasso and ridge solutions on the specific case where the number of observations n is equal to the number of variables p ($n = p$) and the covariate matrix $\mathbf{X}$ is the identity matrix.

a) Derive the OLS estimator for the regression parameter β.

b) Derive the ridge estimator for the regression parameter β with penality parameter λ.

c) Derive the lasso estimator for the regression parameter β with penality parameter λ.

d) Compare the lasso and ridge estimators obtained from the last two points and discuss their respective shrinkage properties.

2. Consider the two hypotheses

$$\begin{aligned} H_0 &: \quad y = \beta_0 + \epsilon, \\ H_1 &: \quad y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{10} x_{10} + \epsilon. \end{aligned}$$

where the x_i are considered fixed and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$.

a) Let p_1 denote the p-value of the t-test associated to the variable x_1 in the full model. Prove that p_1 has a uniform distribution on $[0, 1]$ under H_0.

b) Generate 1000 datasets with $n = 200$ observations under H_0 with $\beta_0 = 2$ and $\sigma = 1$ as well as the response y, in the following way:

1. For each variable x_i draw n i.i.d. realizations from $\mathcal{N}(0, 1)$ and keep these fixed for all datasets.
2. Draw n iid realizations of the noise ϵ from $\mathcal{N}(0, 1)$ and construct corresponding realizations of $y_1, \ldots, y_n$. Repeat this 1000 times.
3. Record the realization of the p-value p_1 for each dataset and draw an histogram. Comment on your results.
4. For the same 1000 datasets, first perform best-subset model selection using the Mallow's Cp criterion to choose a sub-model. If the variable x_1 is included in this sub-model, record the p-value associated to its t-test. Draw an histogram and comment your results, what is different from the last point?
5. Peform again the same procedure but with sample splitting this time, i.e., split each dataset in two parts, select a best sub-model on one half of the data points (again with best-subset selection and Mallow's Cp criterion) and record the p-value associated to x_1 (if x_1 is included in the model) from the fit on the other half of the points. Draw a histogram and comment on your results.

3. Do the conceptual exercise 1 at page 297 of the ISL book (G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning - with Applications in R*, 2017).

4. Do the practical exercise 9 at pages 299 – 300 of the ISL book (G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning - with Applications in R*, 2017).

Preliminary discussion: Friday, May 10.

Deadline: Friday, May 17.