Project Task 1 - BAN404

Jonas Andersson

From February 20, 2019, 09.00 until February 27, 2019, 14.00

Introduction

The report should be no longer than 15 pages. You are allowed to submit R-code in addition; this is encouraged if the code is well documented. To the extent that you are making your own choices in the project, you should choose methods from the first 7 chapters of James et al. (2015).

The project

Start by installing the R-package wooldridge, Shea (2018) and load it. The dataset that you shall analyze is called ceosal2 and contain information about the compensation of 177 CEO's. The dataset also contain 14 variables that might help in predicting CEO compensation. The R-help to the data contains the variable names. Your task is to predict the output variable salary.

a. Describe relevant features of the input and output variables with descriptive statistics.

b. Use standard linear regression (on all or a subset of the predictors), one of the three subset selection methods (best subset, forward stepwise, backward stepwise), lasso and ridge regression to predict salary. Include brief explanations on how the other methods you use are developed from standard linear regression.

c. Evaluate the predictions with at least one of the methods

- the validation set approach
- leave-one-out cross-validation
- K-fold cross-validation

d. Choose an additional prediction method and compare with the other methods.

e. Summarize your results in a conclusion.

References

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2015. *An Introduction to Statistical Learning.* Springer.

Shea, Justin M. 2018. *Wooldridge: 111 Data Sets from "Introductory Econometrics: A Modern Approach, 6e" by Jeffrey M. Wooldridge.* https://CRAN.R-project.org/package=wooldridge.