Modern Data Mining, HW 3

Group Member 1
Group Member 2
Group Member 3

Due: 11:59Pm, 2/27, 2022

Contents

1 Overview

Multiple regression is one of the most popular methods used in statistics as well as in machine learning. We use linear models as a working model for its simplicity and interpretability. It is important that we use domain knowledge as much as we could to determine the form of the response as well as the function format for the factors. Then, when we have many possible features to be included in the working model it is inevitable that we need to choose a best possible model with a sensible criterion. `Cp`, `BIC` and regularizations such as LASSO are introduced. Be aware that if a model selection is done formally or informally, the inferences obtained with the final `lm()` fit may not be valid. Some adjustment will be needed. This last step is beyond the scope of this class. Check the current research line that Linda and collaborators are working on.

This homework consists of two parts: the first one is an excercise (you will feel it being a toy example after the covid case study) to get familiar with model selection skills such as, `Cp` and `BIC`. The main job is a rather involved case study about devastating covid19 pandemic. Please read through the case study first. It is time that group members work together to run a real project. This project is for sure a great one listed in your CV.

For covid case study, the major time and effort would be needed in EDA portion.

1.1 Objectives

- Model building process
- Methods
  - Model selection
    - All subsets
    - Forward/Backward
  - Regularization
    - LASSO (L1 penalty)
    - Ridge (L2 penalty)
    - Elastic net
- Understand the criteria
  - `Cp`
  - Testing Errors
  - `BIC`
  - `K fold Cross Validation`
  - `LASSO`
- Packages
  - `lm()`, `Anova`
  - `regsubsets()`
  - `glmnet()` & `cv.glmnet()`

2 Review materials

- Study lecture: Model selection
- Study lecture: Regularization
- Study lecture: Multiple regression

Review the code and concepts covered during lectures: multiple regression, model selection and penalized regression through elastic net.

3 Case study 1: `ISLR::Auto` data

This will be the last part of the Auto data from ISLR. The original data contains 408 observations about cars. It has some similarity as the Cars data that we use in our lectures. To get the data, first install the package `ISLR`. The data set `Auto` should be loaded automatically. We use this case to go through methods learned so far.

Final modelling question: We want to explore the effects of each feature as best as possible.

1) Preparing variables:

a) You may explore the possibility of variable transformations. We normally do not suggest to transform x for the purpose of interpretation. You may consider to transform y to either correct the violation of the linear model assumptions or if you feel a transformation of y makes more sense from some theory. In this case we suggest you to look into `GPM=1/MPG`. Compare residual plots of MPG or GPM as responses and see which one might yield a more satisfactory patterns.

In addition, can you provide some background knowledge to support the notion: it makes more sense to model `GPM`?

b) You may also explore by adding interactions and higher order terms. The model(s) should be as *parsimonious* (simple) as possible, unless the gain in accuracy is significant from your point of view.

c) Use Mallow's C_p or BIC to select the model.

2) Describe the final model and its accuracy. Include diagnostic plots with particular focus on the model residuals.

- Summarize the effects found.
- Predict the `mpg` of a car that is: built in 1983, in the US, red, 180 inches long, 8 cylinders, 350 displacement, 260 as horsepower, and weighs 4,000 pounds. Give a 95% CI.
- Any suggestions as to how to improve the quality of the study?

4 Case study 2: COVID

See covid_case_study.Rmd.