ESE 542: Statistics for Data Science

Chapter 6 Project

In addition to answering the **bolded** questions on Coursera, also attach your notebook, both as `.ipynb` and `.html`.

In the following exercise, we will perform model selection to find the best model for two datasets. Perform the following analyses by starting a new notebook.

1 Part A

First, we will run multiple linear regression on the Auto dataset and use subset selection to find the best model. This dataset contains the following nine columns from 392 cars:

- mpg: continuous
- cylinders: multi-valued discrete
- displacement: continuous
- horsepower: continuous
- weight: continuous
- acceleration: continuous
- model year: multi-valued discrete
- origin: multi-valued discrete
- car name: string

1. Produce a scatter plot matrix which includes all the variables in the dataset. Comment on your observations. Are there any variables in particular which seem to be strongly correlated?
2. Compute a matrix of correlations between the variables using the `pandas corr()` function.
3. Using `StatsModels`, perform linear regression with ‘mpg’ as the response variable and all other variables except ‘name’ as predictors. Print the results of your regression analysis. Comment on the output and **answer the following on Coursera**:

(a) **What is the relationship between the predictors and the response?**

(b) **Which predictors appear to have a statistically significant relationship with the response variable at a 95% confidence level?**

(c) **What does the coefficient for the 'year' variable suggest?**

4. Select the optimal model by manually performing forward stepwise selection. The goal of this exercise is to show the sheer number of models needed for forward stepwise selection. To do this, first split the dataset into a training set and a test set, with a `test_size`of 20% and `random_state`= 42. It is important to only use the training set to train the model. You may use the `processSubset`function from the recitation, but you should at least run 3 iterations of forward stepwise selection manually, as we wish to see each step of forward stepwise selection. First, run linear regression with one variable. Select the best model using training RSS as the performance metric. Using that first variable, continue adding variables, one at a time, until your linear model includes all of the variables. Afterwards, calculate the test RSS of all your models and select the one that minimizes test RSS. *Hint*: You can use the result of linear regression from `Stats Models` to calculate RSS by looking at the sum of the squared residuals. **Answer on Coursera: Which variables are included in your model after performing forward stepwise selection?**

5. Using the full dataset, fit a linear regression model with interaction effects between 'displacement', 'weight', 'year', and 'origin'. Do any interactions appear to be statistically significant? *Hint*: In addition to the full model with all seven predictors, your model should include six more interaction terms. **Answer on Coursera: Which of the interaction terms are statistically significant?**

2 Part B

Next, we will use the College dataset to predict the number of applications received using the other variables in the College dataset. We will then use regularization to study their effects on our model.

The College dataset contains the following variables from 777 different universities and colleges in the US:

- Private : Public/private indicator
- Apps : Number of applications received
- Accept : Number of applicants accepted
- Enroll : Number of new students enrolled
- Top10perc : New students from top 10

- Top25perc : New students from top 25
- F.Undergrad : Number of full-time undergraduates
- P.Undergrad : Number of part-time undergraduates
- Outstate : Out-of-state tuition
- Room.Board : Room and board costs
- Books : Estimated book costs
- Personal : Estimated personal spending
- PhD : Percent of faculty with Ph.D.'s
- Terminal : Percent of faculty with terminal degree
- S.F.Ratio : Student/faculty ratio
- perc.alumni : Percent of alumni who donate
- Expend : Instructional expenditure per student

1. Split the dataset into a training set and a test set, with a `test_size` of 20% and `random_state` = 1. Set the dataframe index to be the 'Names' column.

2. Fit a linear model using `Stats Models` on the training set, and report the test MSE obtained. Name this variable `test_MSE` and **enter your answer on Coursera**.

3. Fit a ridge regression with a λ parameter of 0. **Answer on Coursera: What do you notice about the test MSE in this case?**

4. Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained. *Hint*: Look at the recitation guides for how to implement cross-validation with `RidgeCV`. `RidgeCV` essentially performs hyper-parameter optimization (more on this in the next recitation) by testing all possible parameters through cross validation. For its parameters, specify `KFold cross validation` with ten folds, `scoring` with mean squared error, `normalization` set to true, and 50 equally space λ values ranging from 10^2 to 10^3. Name the selected value of λ as `ridge_select` and calculate the corresponding test MSE as `test_MSE_ridge`. **Enter your answers on Coursera.**

5. Compare the ridge regression coefficients when using $\lambda = 0$ and the value for λ given by `RidgeCV`. Comment on your observations.

6. Fit a lasso model on the training set, with λ chosen by cross-validation. Specify `KFold cross validation` with ten folds, `normalization` set to true, and 50 equally space λ values ranging from 10^2 to 10^3. Name the selected value of λ as `lasso_select` and calculate the corresponding test MSE as `test_MSE_lasso`. Also report the number of non-zero coefficient

estimates by looking at the output of `pd.Series(lasso.coef_,index=x.columns)`. **Enter your answers on Coursera.**

7. Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these three approaches?