IEOR 242: Applications in Data Analysis, Fall 2021

Homework Assignment #1

September 10, 2021

Problem 1: (15 points)

James Chapter 3, Exercise 4 (Polynomial vs. cubic regression).

Problem 2: (10 points)

James Chapter 3, Exercise 5 (Fitted values in simple linear regression without an intercept).

Problem 2: Forecasting Nissan Rogue Sales (Adapted from Bertsimas 22.1) (75 points)

Nearly all companies seek to accurately predict future sales of their product(s). If the company can accurately predict sales before producing the product, then they can better match production with customer demand, thus reducing unnecessary inventory costs while being able to satisfy demand for their product.

In this exercise, you are asked to predict the monthly sales in the United States of the Nissan Rogue automobile. Nissan is a brand of Japanese automobiles which is now the sixth-largest automobile manufacturer in the world, and was the number one car manufacturer in the United States in 2014. The Rogue is a car model of Nissan that was first produced in 2008. It is Nissan's best selling car in the United States. We will use linear regression to predict monthly sales of the Rogue using economic indicators of the United States as well as (normalized) Google search query volumes. The data for this problem is contained in the file `Rogue-242-Fall2021.csv`. Each observation in the file is for a single month, from January 2008 through July 2021. The variables are described in Table 1.

a) (25 points) Start by splitting the data into a training set and two testing sets, testing set A and testing set B. The training set should contain all observations from 2008 through 2015. Testing set A should have all observations from January 2016 through December 2019, and testing set B should have all observations from January 2020 through July 2021.

Consider just the four independent variables `Unemployment`, `RogueQueries`, `CPIEnergy`, and `CPIAll`. Using your regression skills, select a subset of these four variables and construct a regression model to predict monthly Rogue sales (`RogueSales`). Try to choose which of the four variables to use in your model in order to build a high-quality linear regression model.

Table 1: Variables in the dataset `Rogue-242-Fall2021.csv`.

Variable	Description
`MonthNumeric`	The observation month given as a numerical value (1 = January, 2 = February, 3 = March, etc.).
`MonthFactor`	The observation month given as the name of the month
`Year`	The observation year.
`RogueSales`	The number of units of the Nissan Rogue sold in the United States in the given month and year.
`Unemployment`	The estimated unemployment rate (given as a percentage) in the United States in the given month and year.
`RogueQueries`	A (normalized) approximation of the number of Google searches for "Nissan Rogue" in the United States in the given month and year.
`CPIAll`	The consumer price index (CPI) for all products for the given month and year. This is a measure of the magnitude of the prices paid by consumer households for goods and services.
`CPIEnergy`	The monthly consumer price index (CPI) for the energy sector of the US economy for the given month and year.

Use the training set to build your model, and do not add any additional variables beyond the four indicated independent variables. Write a brief explanation (no more than one page, preferably less) – targeted to a statistically literate manager – describing how you decided on the variables to use in the model and the quality of the linear regression model's predictions, as evaluated using the training set (there is no need to consider the test set for this part of the problem). Be sure to address the following in your explanation:

i) What is the linear regression equation produced by your model, and how should one interpret the coefficients for the independent variables? Consider interpretability issues when writing down the equation (e.g., do not just copy and paste the output from Python).

ii) How did you select the variables to include in your linear regression model?

iii) Do the signs of the model's coefficients make sense? Are you reasonably sure that the

signs are correct?

iv) How well does the model predict training set observations? Can you justify the model's performance on the training data with a quantifiable metric?

b) (15 points) Let us now try to further improve the linear regression model by modeling seasonality. In predicting demand and sales, seasonality is often very important since demand for most products tends to be periodic in time. For example, demand for heavy jackets and coats tends to be higher in the winter, while demand for sunscreen tends to be higher in the summer.

Construct a new linear regression model using the `MonthFactor` variable as an independent variable, in addition to all four of the variables you used at the start of part *(a)*. **There is no need to do variable selection for this part of the problem.** As before, construct your model based on the training data.

Answer the following questions about this modeling exercise.

i) Describe your new model. What is the regression equation? (Do not simply copy and paste output from Python.) How should one interpret the coefficients of each of the `MonthFactor` dummy variables?

ii) What is the training set R^2 for the new model? Which variables are significant?

iii) Do you think adding the independent variable `MonthFactor` improves the quality of the model? Why or why not?

iv) Can you think of a different way that you might use the given data to model seasonality? Do you think your new way would improve on the best model you have constructed so far? (By the way, later in the course we will have a lecture dedicated to basic time series modeling, and we will explore a number of ways to construct models using datasets with an associated time component.)

c) (15 points) Build a final model using a subset of the independent variables used in parts *(a)* and *(b)*, providing a brief justification for the variables selected. What is the training set R^2? What are the OSR^2 values for testing set A and testing set B? How do these three numbers compare to each other? Please provide a plausible explanation for any significant differences you do or do not observe.

d) (10 points) Let us now consider adding an additional feature/variable to your final model from part *(c)*. Based on your knowledge and intuition, think of a monthly variable that you hypothesize might be related to Nissan sales. Provide a one or two sentence explanation for your choice. Search online for a data source for your chosen variable (if you are not able to find data, then you need to pick a different variable), and append your collected data as a new column in the original data file. (It is OK to use variables similar to what we used above, i.e., a different economic indicator or Google trends data for a different search term, but feel free to get as creative as you like.)

Now, build a new regression model with your additional chosen feature in addition to the features that you selected in part *(c)*. Does the new feature add any predictive value? Justify your answer based on the results of your analysis.

e) (10 points) In regression analysis, a loss function ℓ takes as input a predicted value $\hat{y}$ and an observed value y, and it returns a value $\ell(\hat{y}, y)$ which is interpreted as the loss/error/cost associated with predicting $\hat{y}$ when the actual value of the dependent variable is y. So far, we have only considered the squared loss $\ell(\hat{y}, y) := (y - \hat{y})^2$, which is the most standard loss function in regression. However, the squared loss is not always the most appropriate or the most effective in every situation.

Consider the following (greatly simplified) scenario regarding how Nissan makes monthly production decisions. Firstly, the management at Nissan has decided to use the predictions of your regression model to directly set monthly inventory levels. That is, if your model predicts that next month's Rogue sales will be $\hat{y}$, then Nissan will have available exactly $\hat{y}$ Rogue units to be sold next month. (You may ignore integer constraint issues and assume that Nissan can produce fractional units.) Whenever a unit is not sold in a given month, then Nissan can use that unit to offset part of next month's production. For example, if Nissan has five Rogue units available in January but only sells three of them, then they can carryover two units to February. For simplicity, you may assume that the number of units carried over from month to month is always less than or equal to the target inventory levels given by the predictions of your model. Finally, there is a cost of \$500 associated with carrying over a unit from one month to the next. Suppose that Nissan earns a profit of \$3000 for each Rogue unit that it sells. Propose a loss function ℓ that accurately models this situation and explain your reasoning.