CUHK Business School

Problem Set 2

DSME 6756: Business Intelligence Techniques and Applications (Winter 2021)

Due at 9:30AM, Thursday, December 23, 2021

Instructions

Please read the Jupyter Notebook for Session 2 `2-Linear-Regression.ipynb` from the beginning to **Section 5.2. Understanding Linear Regression Results**. Submit a Jupyter Notebook of your solutions with code on Blackboard. The total achievable points are 5 for this problem set. Please name your Jupyter Notebook as
`YourLastName_YourFirstName_PS2.ipynb` (e.g., `Zhang_Renyu_PS2.ipynb`)

1. Playing with the WHO Data Set (ctd.) (0.5 point)

Please read the data set `WHO.csv` into Python and Demonstrate the relationship between income level vs. life expectancy through visualization.

2. User Retention (1.5 points)

The data set `Retention.csv` contains the active user information of an App for 3 days. It has three variables:

- *user_id*: A unique identifier for each user.
- *play_duration*: The amount of time (in minutes) the user uses the App.
- *day*: The day of the record.

Note that only users who are active (i.e., log into the App) will be recorded in this data set. Please load the data set into Python and answer the following questions:

(a) (1 point) The retention rate of day i is defined as the proportion of active users of day i who will remain active in day $i+1$. Calculate the retention rate of day 1 and day 2, respectively.

(b) (0.5 point) Define the users whose play_duration exceed 6 minutes in day i as very active users. The other users are defined as marginally active users. Compare the retention rates of very active users and marginally active users in day 1 and day 2.

3. Forecasting Auto Sales (3 points)

In this problem, we will try to predict monthly sales of an Auto Brand.

The file `Auto.csv` contains data for the problem. Each observation is a month, from January 2010 to February 2014. For each month, we have the following variables:

- $Month$ = the month of the year for the observation (1 = January, 2 = February, 3 = March, ...).
- $Year$ = the year of the observation.
- $AutoSales$ = the number of units of the Auto sold in the United States in the given month.
- $Unemployment$ = the estimated unemployment percentage in the United States in the given month.
- $Queries$ = a (normalized) approximation of the number of Google searches for "Auto" in the given month.
- CPI_energy = the monthly consumer price index (CPI) for energy for the given month.
- CPI_all = the consumer price index (CPI) for all products for the given month; this is a measure of the magnitude of the prices paid by consumer households for goods and services (e.g., food, clothing, electricity, etc.).

Load the data set into Python and split the data set into training and testing sets as follows: Place all observations for 2012 and earlier in the training set, and all observations for 2013 and 2014 into the testing set.

(a) (0.4 point) Build a linear regression model to predict monthly Auto sales using Unemployment, CPI_all, CPI_energy and Queries as the independent variables. Use all of the training set data to do this. Please try to interpret your estimation results.

(b) (0.4 point) We would now like to improve the model by incorporating seasonality. Seasonality refers to the fact that demand is often cyclical/periodic in time. For example, demand for warm outerwear (like jackets and coats) is higher in fall/autumn and winter than in spring and summer. In our problem, since our data includes the month of the year in which the units were sold, it is feasible for us to incorporate monthly seasonality. From a modeling point of view, it may be reasonable that the month plays an effect in how many Auto units are sold. To incorporate the seasonal effect due to the month, build a new linear regression model that predicts monthly Auto sales using Month as well as Unemployment, CPI_all, CPI_energy and Queries. Do not modify the training and testing data frames before building the model. Based on the model estimation results, how do you evaluate the new model compared with the original one?

(c) (0.4 point) In the new model, given two monthly periods that are otherwise identical in Unemployment, CPI_all, CPI_energy and Queries, what is the absolute difference in predicted Auto sales given that one period is in January and one is in March? Consider again the new model, given two monthly periods that are otherwise identical in Unemployment, CPI_all, CPI_energy and Queries, what is the absolute difference in predicted Auto sales given that one period is in January and one is in May? Is there anything you feel uncomfortable about this finding?

(d) (0.4 point) Alternatively, we consider Month as a factor variable, instead of a numeric variable. Then, we can use the binary variable technique introduced in Session 2's lecture to build a linear regression model. Why do you think we should use the factor variable instead of the numeric variable to represent month?

(e) (0.4 point) Re-run the regression with the Month variable modeled as a factor variable. From the new regression results, what seasonality pattern have you observed?

(f) (0.5 point) Another peculiar observation about the regression results (with month as a factor variable) is that the signs of the Queries variable and the CPI_energy variable. Why their signs are counter-intuitive? Please try to give an explanation for such phenomenon and find a way to address this issue. You may need to remove some independent variables and re-build the linear regression model.

(g) (0.5 point) Use out-of-sample test to evaluate all your models built to estimate the sales of Auto. Report the out-of-sample R^2 of each model and discuss which model you would like recommend to this Auto Brand for their sales forecasting.