**” Black Friday Practice Problem”**

Rui Shi 182007853

Abstract

In order to predict the purchase amounts of different customers in Black Friday according to historic data, I try to build a regression model. First, I need to do data preprocessing, and then run model optimization with different methods for model selection and model diagnostics after building a simple regression model. Finally, I tried PCA regression to eliminate multicollinearity.

**KEYWORDS: Dummy Variable, Linear Regression, Model** **Diagnostics, Model Selection, Multicollinearity, PCA Regression,**

1, Introduction

A retail company “ABC Private Limited” wants to understand the customer purchase behavior (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high volumes products from last month.

The data set also contains customer demographics (age, gender, marital status, city\_type, stay\_in\_current\_city), product details (product\_id and product category) and Total purchase\_amount from last month.

Now, they want to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.

Data

*Variable* *Definition*

User\_ID User ID

Product\_ID Product ID

Gender Sex of User

Age Age

Occupation Occupation (Masked)

City\_Category Category of the City (A,B,C)

Stay\_In\_Current\_City\_Years Number of years stay in current city

Marital\_Status Marital Status (0,1)

Product\_Category\_1 Product Category1 (Masked)

Product\_Category\_2 Product Category2 (Masked)

Product\_Category\_3 Product Category3 (Masked)

Purchase Purchase Amount (Target Variable)

**Tips: A product belongs to category 1 may also belongs category 2 and 3.**

2, Data Preprocessing

Since all the 9 predictors are discrete, most of them are different categories. We need to construct dummy variables to process the original data.

2.1, Dummy Variable

* **Definition:­­­**

A dummy variable is one that takes the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome.

* **Take “City\_Category” as example.**

There are three different cities “A”, ”B”, ”C”. I split “City\_Category” into two new variables “City\_CategoryA”, “City\_CategoryB”.

If the observation belongs to city A, then assign 1 to “City\_CategoryA” and 0 to “City\_CategoryB”.

If the observation belongs to city B, vice versa.

If the observation belongs to city C, assign 0 to both of them.

* **The reason why we can’t set up the third variables “City\_CategoryC”.**

To avoid multicollinearity: “City\_CategoryA” + “City\_CategoryB” + “City\_CategoryC” =1.

* **The variables “Gender”, “Occupation”,” Product\_Category\_1”,**

**“Product\_Category\_2”, “Product\_Category\_3” represent different categories should also be split to dummy variables.**

2.2, Scores

* **For the variables “Age”, “Stay\_In\_Current\_City\_Years”, it may be**

**more reasonable to represent them in numbers but not dummy variables.**

**So, I assigned scores to them.**

For instance, assign 1-7 to “Age” from the youngest intervals to the oldest intervals.

3, Model Optimizatiom

3.1, Model Diagnostics Ⅰ

Firstly, I fitted the model with all the variables and plotted the standardized residuals against fitted values (figure 1), and QQ plot (figure 2). I found from the plots that mean structure and normality are violated. It seems that . Therefore, I tried two methods to make transformation upon Y.

1, **Make log transformation upon Y** and draw the two plots again. (figure 3&4)

2, **Use Box-Cox method** to obtain the best lambda (figure 5) and draw the plots for the third time. (figure 6&7)

Comparing the two groups of plots upon two transformations, **I consider log transformation of Y a better way** because it solves the problem of mean structure violation somehow.

3.2, Model Selection

I tried four methods—**step-wise backward selection, stepwise forward, backward and hybrid selection.(use AIC as criterion)**—to obtain four different models with different predictors. Then I compared the adjusted R2 and did **K-fold Cross validation**. The best model which holds the biggest adjusted R2 and smallest prediction error is obtained from stepwise forward selection, therefore I chose the best model to move on.

3.3, Model Diagnostics Ⅱ

1, **Check** **multicollinearity.**

Serious multicollinearity since the average VIF almost reaches 5 and the maximum VIF is as big as 48.

2, **Check influential observations**

No influential point.

4, PCA Regression (to eliminate multicollinearity)

(1), Before doing PCA, I used the following formula for **standardization of observations in different scales**.

(2), After that I did PCA and got the screeplot. (figure 8)

(3), Make transformation on data upon the principle components and Fit regression models with the first 57 principle components whose accumulative variance proportion has just reach 0.9.

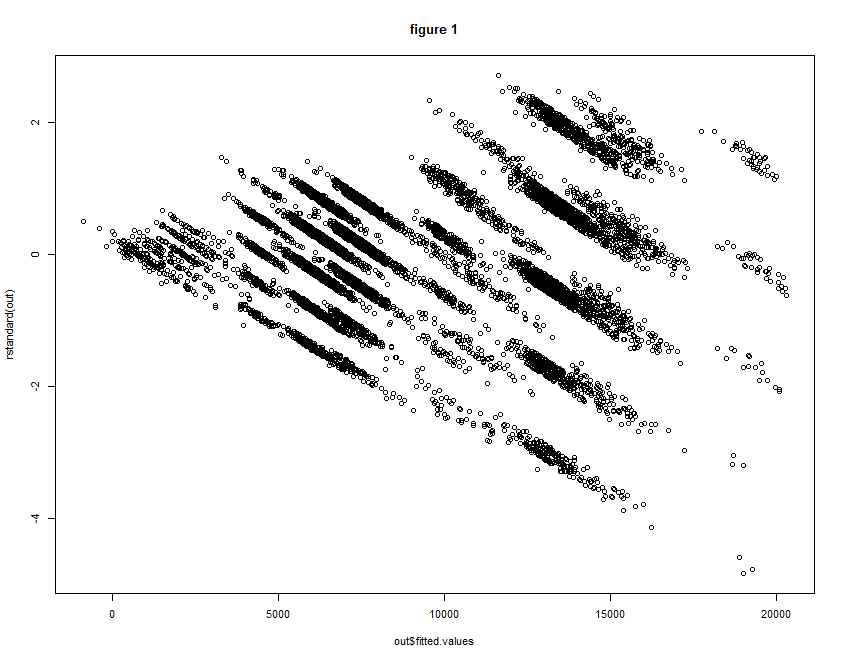
(4), Plot residuals against fitted values for diagnostics. From the plot (figure 9), we can see that I still need to make log transformation upon y.

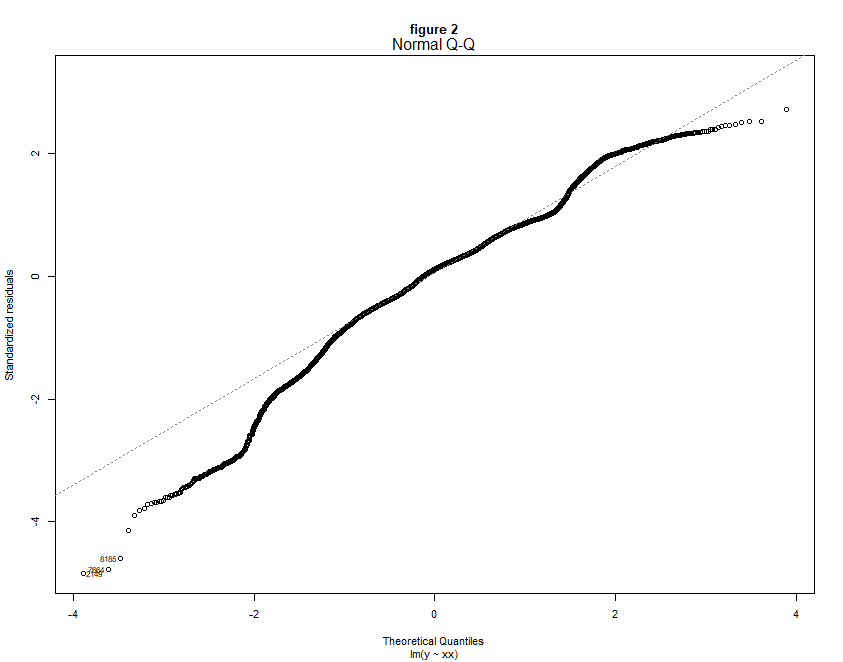
5, Do stepwise backward selection. (use AIC as criterion) and finally I got the PCA regression model.

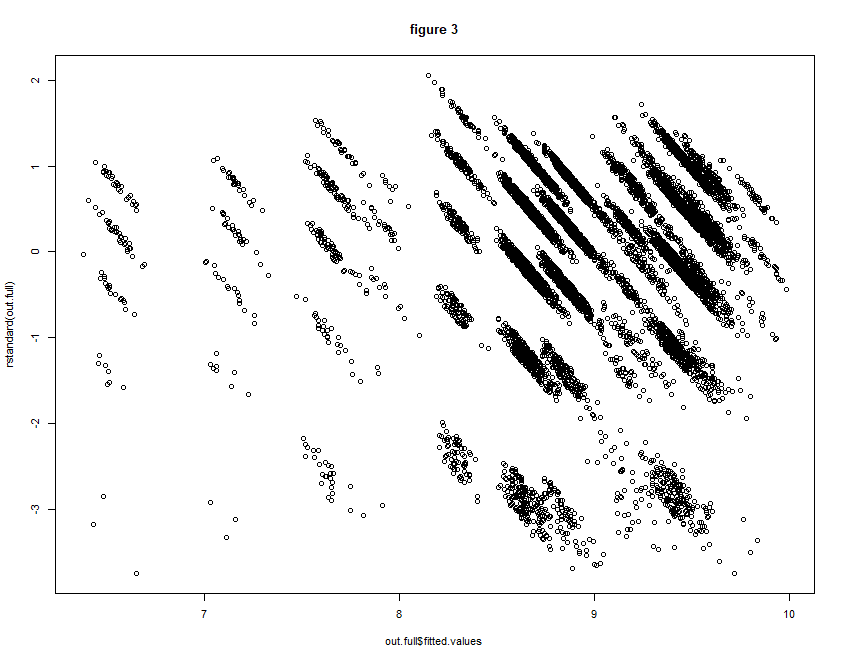
5, Summary

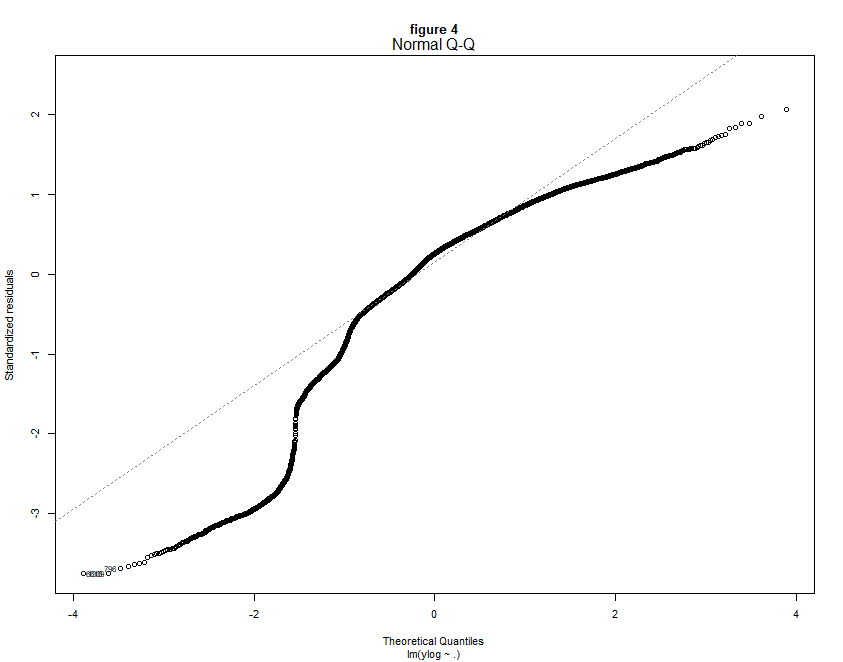
It’s totally a new attempt to fit linear regression model upon the dataset whose predictors are all discrete. Constructing dummy variables offers us a great solution for including all these discrete variables which are not numerical into the model. After data preprocessing. I tried almost every mothed I have learned in the class including model diagnostics and stepwise model selection as well as K-fold Cross validation. Finally, I tried PCA regression to eliminate multicollinearity.

Appendix:









**figure 5**

