” Black Friday Practice Problem”

Rui Shi 182007853

Problem Statement

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.

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.

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" which represent different categories should also be split to dummy variables.**

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.

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 $\sigma^2 \propto [E(Y)]^2$. 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.

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 R^2 and did **K-fold Cross validation**. The best model which holds the biggest adjusted R^2 and smallest prediction error is obtained from stepwise forward selection, therefore I chose the best model to move on.

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.

PCA Regression (to eliminate multicollinearity)

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

$$X_j^* = \frac{X_j - \mu_j}{\sqrt{\sigma_{jj}}}$$

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.

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 1

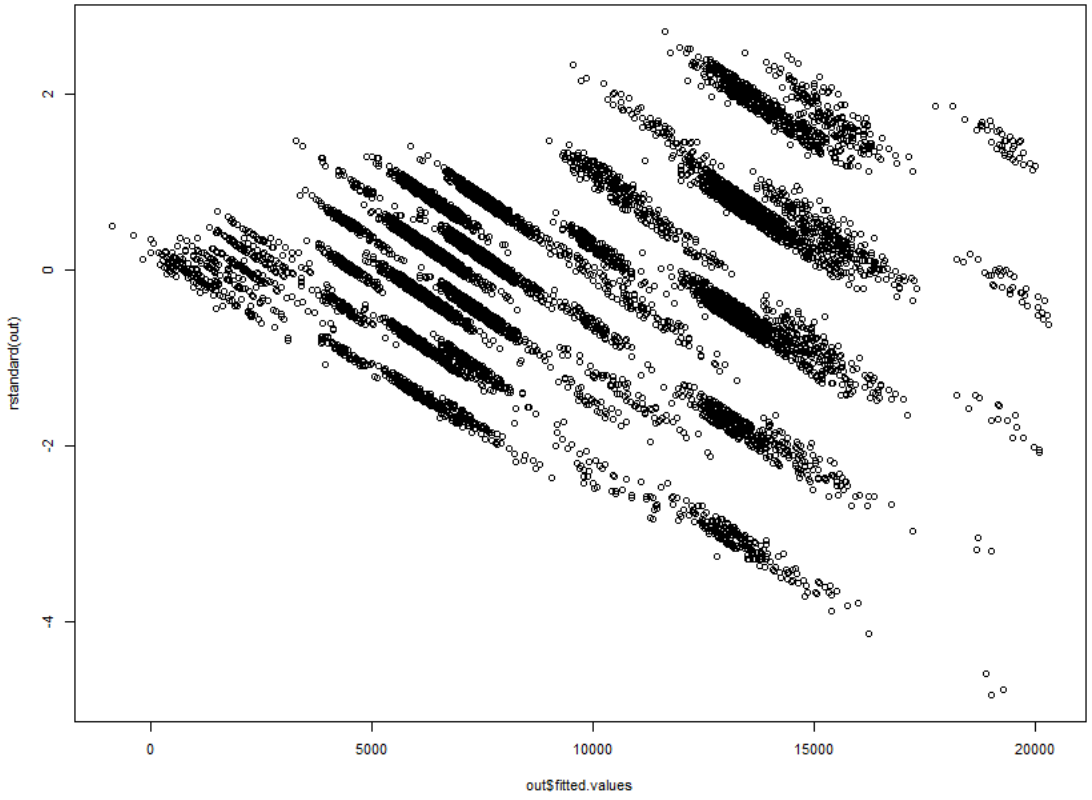


figure 2

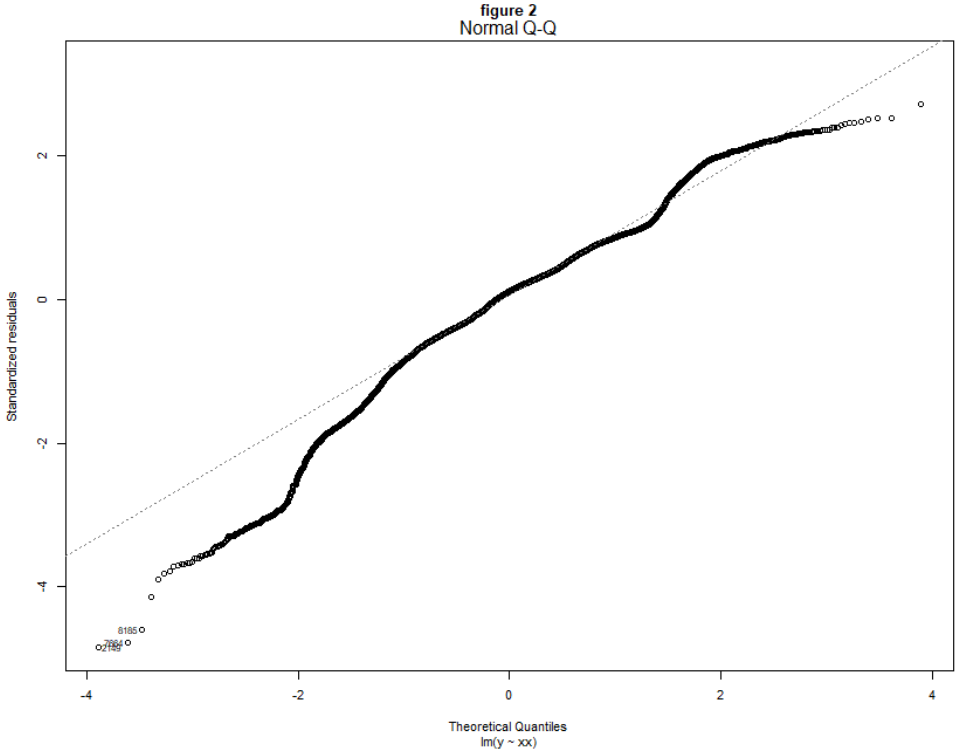


figure 3

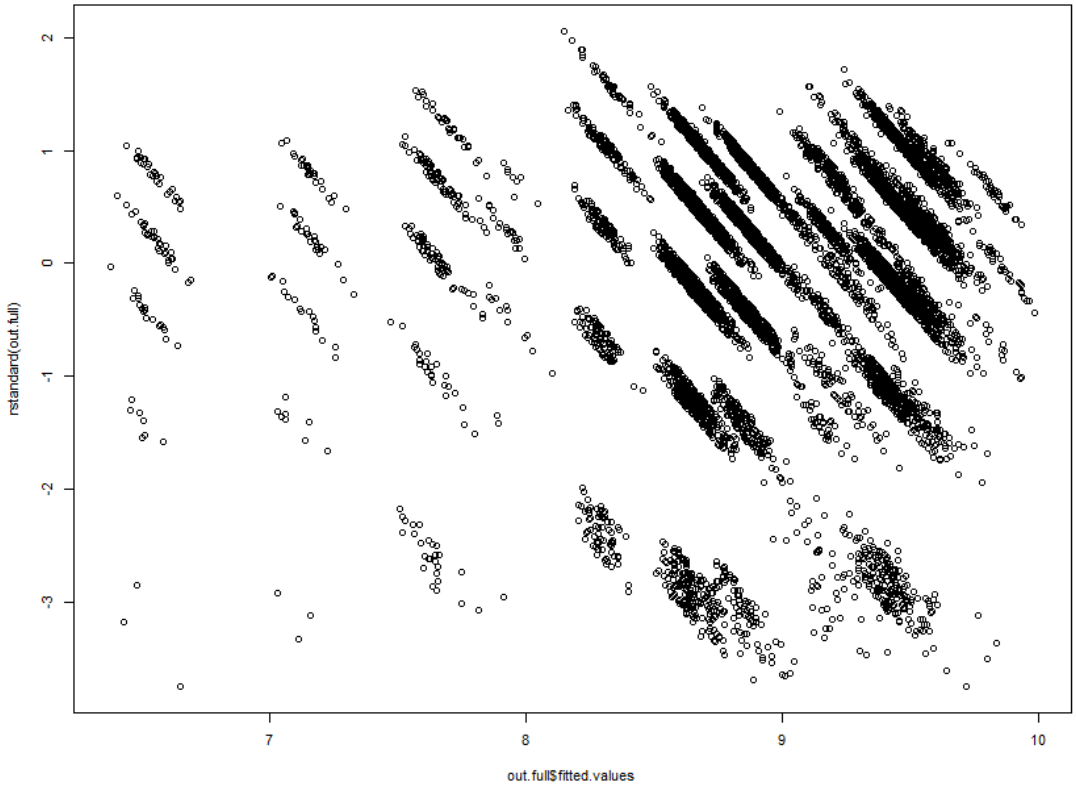


figure 4

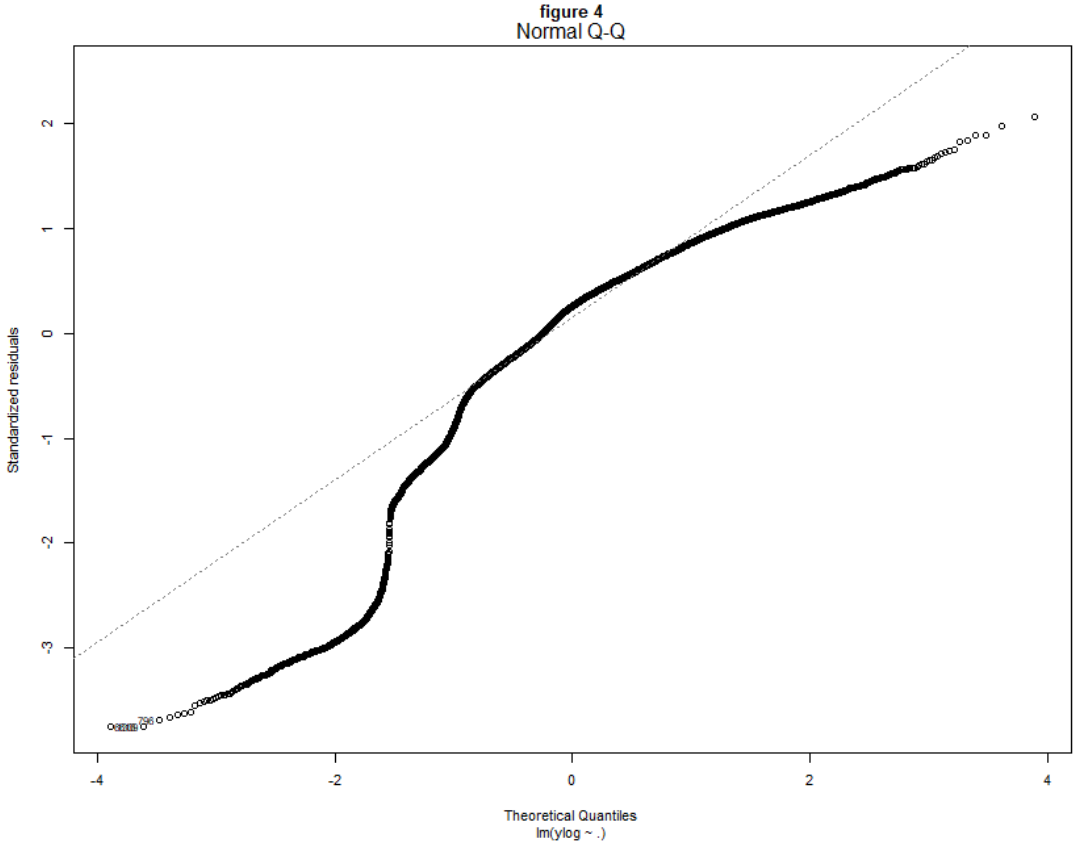


figure 5

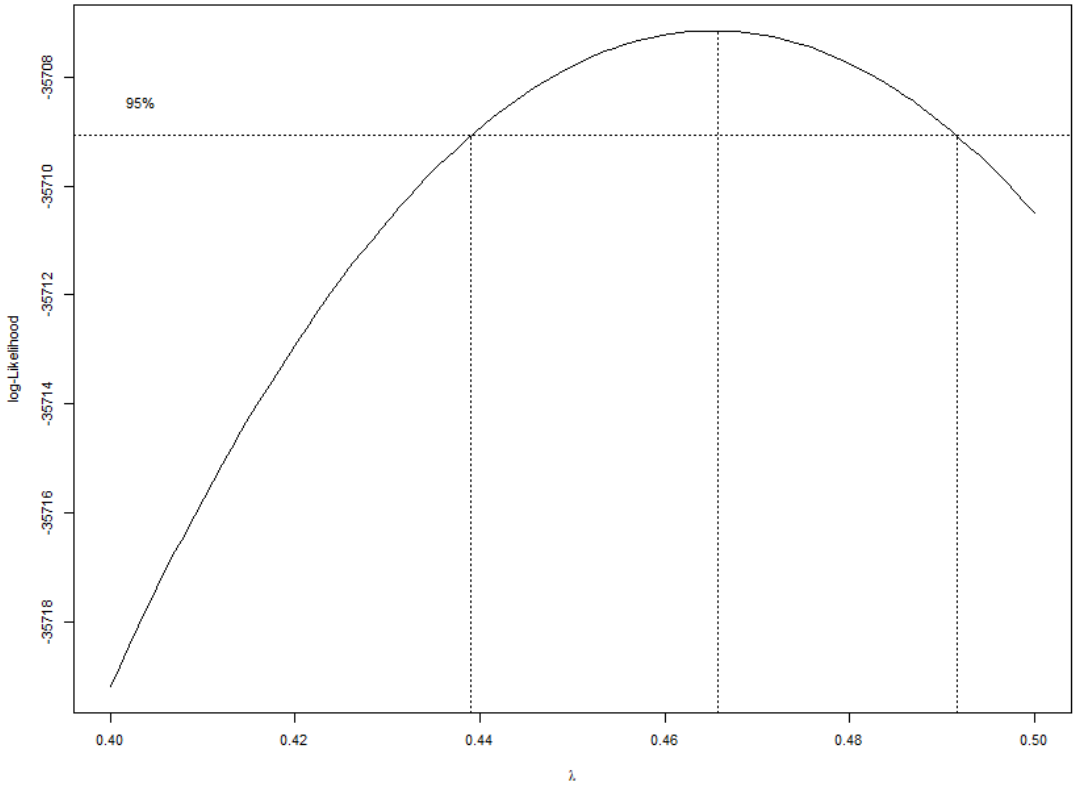


figure 6

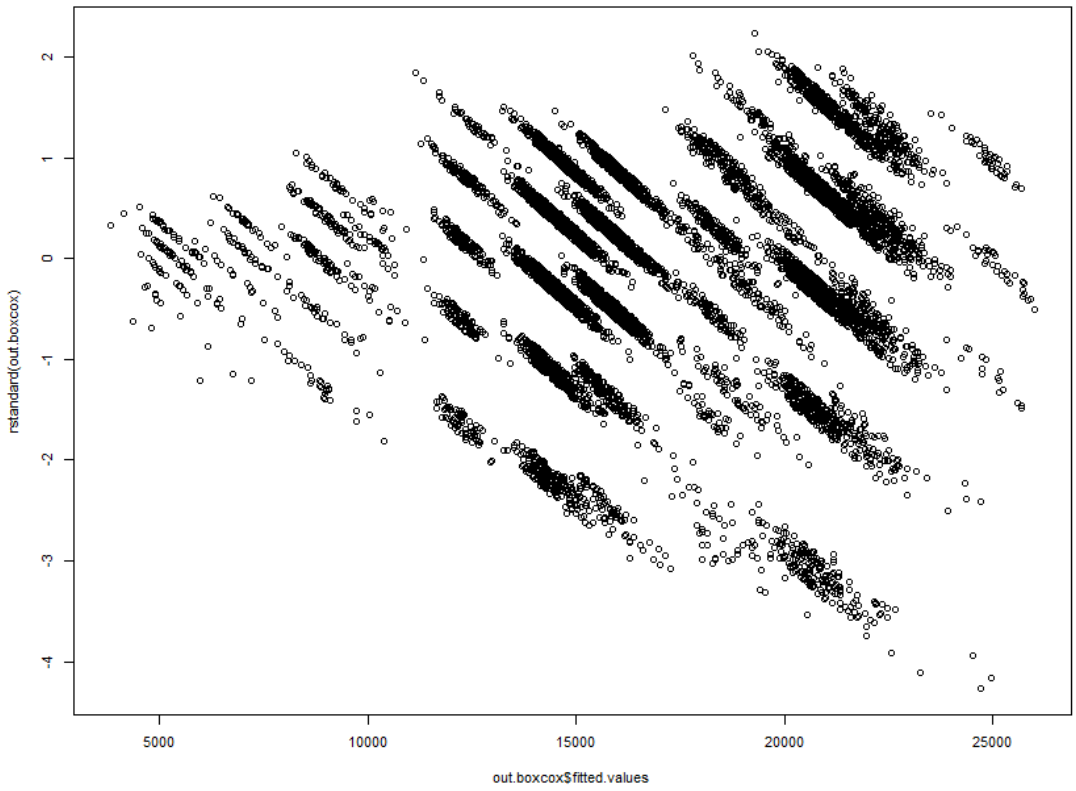


figure 7

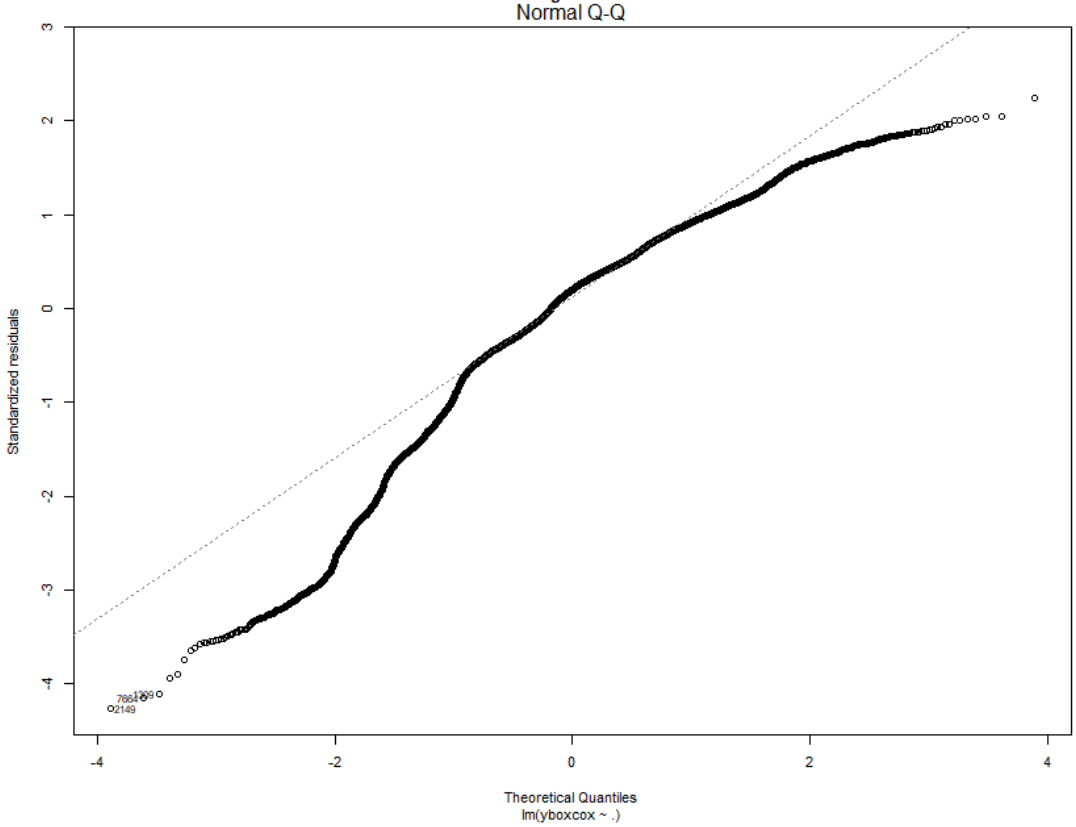


figure 8

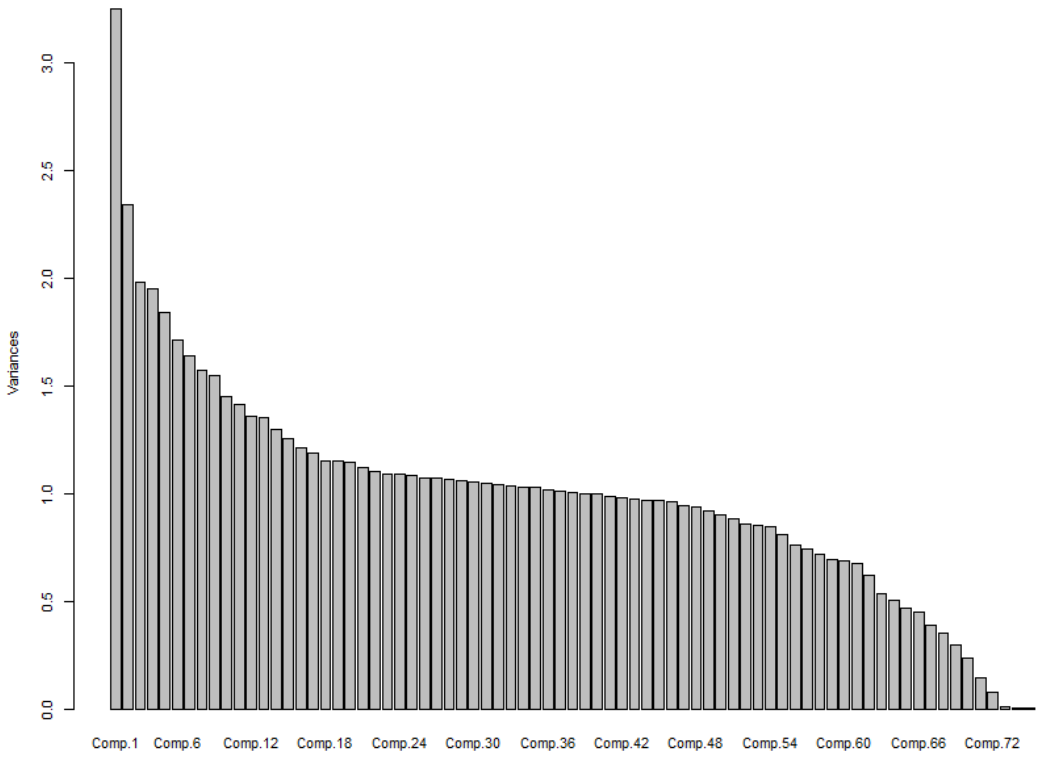
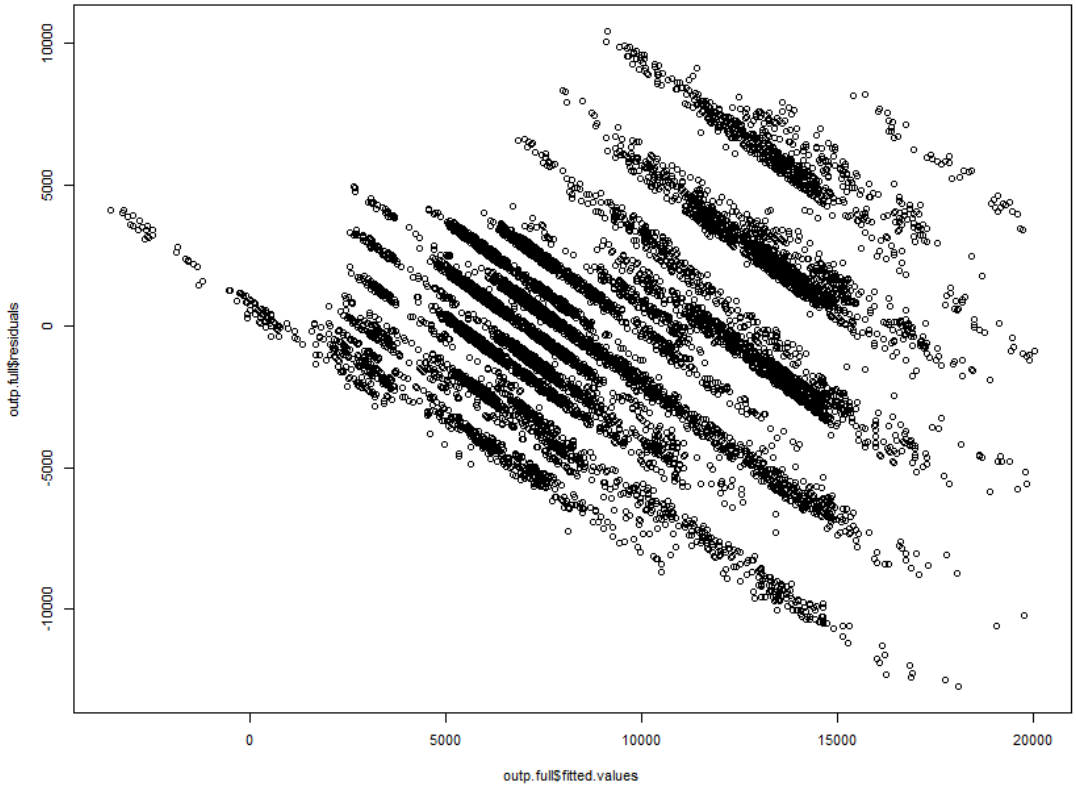


figure 9


outp.full$residuals
10000
5000
0
-5000
-10000
0
5000
10000
15000
20000
outp.full$fitted.values