Course Assignmemt 2 2023-24

Advanced Data Analysis with R

Nikolaos Papadopoulos, Postgraduate student in Statistics, AUEB

2024-02-11

Contents

1. Introduction

1.1. Abstract

The following report is the outcome of a statistical data analysis, performed for the students' evaluation of the course “Advanced Data Analysis with R” in the department of Statistics of the Athens University of Economics and Business, where a total of 5000 clients' data were given by “Motodynamiki”. The main objective is to describe the characteristics of the clients that have purchased on-desk increments, in order for the company to create a more targeted campaign.

1.2. Data Presentation and Correction

1.2.1. Columns' Manipulation

The columns “**OnDesk.Insurance.Net.Revenue**” and “**OnDesk.Non.Insurance.Net.Revenue**” are renamed as “ODI” and “ODNI” in short. Also, a new variable called “**OnDesk.Net.Revenue**” is added as the sum of the “ODI” and “ODNI” variables and will be referred to as “ODNR”. Since we contain the ODNR, we can also remove the rest two in order to avoid multicollinearity in the next steps of the analysis.

Another important element that should be added is the amount of Increments that each client bought. We are not truly interested in this phase on what products the customers bought, rather on how many money they payed “On Desk” and so we can simplify the data by adding a column.

Some unimportant for the analysis variables are removed, which are the “**Res.no**”, “**Agr.no**”, “**Driver.ID**”, “**Check.out.Station**” variables and correspond to the identification of the drivers. For the first part of the analysis, the variables “**Check.out.Date**”, “**Check.out.Time**”, “**Booking.Date**” and “**Booking.Time**” will be initially removed as well for the first part of the analysis and later on, their importance will be judged to determine if it is reasonable for them to exist in the analysis.

1.2.1. Rows' Manipulation

The reason that segment exists in this assignment is because there are some inserts that are probably untrustworthy, for example, in the age variable, there are observations where the age is 122, which doesn't seem normal. Fun fact, all of them are in Greece. So, in the age variable, we will focus only on those whose age is less than 90.

1.3. Dealing with NAs

In order to perform a good analysis it is imperative to have the biggest possible data set, and avoid throwing out whole rows due to only 1 NA when the rest columns include important information. One of the continuous variables that contain lots of NAs is the “Driver.Age” variable. Many methods can be applied to “guess” the missing values, but firstly, let’s see the correlation plot of the numerical variables and focus on the information obtained by the age variable. Eventually, it seems that the age variable is not correlated to any of the other numerical variables that are selected, so instead of performing conditional sampling, it is reasonable to just create random values and plug them instead of the NAs.

2. Descriptive Data Analysis

2.1. ODNR variable

In contents of the descriptive part of the analysis, some important summary statistics are included both for the explanatory variable (ODNR) and the rest variables depending on their nature, as well as the distributions that they follow since we can obtain important information for the rest of the analysis.

Primarily, the interest circles around the ODNR variable and more specifically on the non zero values, where the descriptive statistics are the following;

These plots represent the ODNR and non-zero ODNR distribution respectively. The left plot indicates a huge peak on 0 and a right-skewed tail then. On the right plot that the 0 values are absent, one may observe that there is still a huge peak closer to zero than any other value and again, a right-skewed tail. However, the data in the right plot seem to be spread with much better properties than the left side plot.

Another interesting presentation of the ODNR variable is if the variables is split in 3 spaces and plot the histograms for each interval. One nice representation is if we split the data as follows;

Now, we get a better image on how many observations exist in each interval (roughly) and how they are distributed inside of it. As expected, the higher the interval, the less observations and a more right-skewed tail.

2.2. Rest variables

Of course the descriptive statistics of the data set are not surrounded exclusively by the explanatory variable. Below a quick representation of the rest statistics is cited;

The above table contains plenty useful information about the nature of each variable that could be mentioned. Later on, maybe they will be handy in the analysis as well.

3. Pairwise Comparisons and correlations

3.1. Qualitative variables

In this section, in order to find possible correlation between the qualitative variables and ODNR, many methods can be applies from factor analysis, but the briefest method shall be the ANOVA for modelling comparisons between the Null models and the ones that contain each time one of the qualitative variables.

Note: As before, the Incremental variables are not going to be counted in this analysis.

The p-values of the model that each one of them creates when we perform the regression with ODNR and compare to the Null model, are all less than 0.05, so we assume that there is a relation between these variables with the ODNR individually.

3.2. Quantitative variables

3.2.1. Correlation plot

For this part, the analysis is going to be split in 2 parts, where the one will concern the numerical variables, and then the categorical variables, always comparing to our explanatory variable, ODNR.

To start with, a correlation plot between all numeric variables will take place. We are interested in all the numerical variables' correlations because we know what dependencies exist in our data set, so that we avoid possible problems with the assumptions of our models.

The focus is on the row of the correlation between the ODNR and the rest variables, and as one may see, there is a very low correlation, almost insignificant. Usually values less than 0.3 are considered as no correlation, between 0.3 and 0.3 as existent correlation, and from 0.6 to 1 as strong correlation. Same applies for negative values for negative correlations.

3.2.2. Simple Linear Regressions

The purpose of this paragraph is to verify the above results. Every quantitative variable is going to be set into a regression model with the ODNR.

The variables that are significant from this analysis, come to be the “Driver Age”, “Pre.paid.Amount”, “C.O.Mileage”, “Internet.Non.Insurance.Net.Revenue”, “Rental_Cost_Res” and “TI”. This is a sign for the next part of the analysis.

4. Regression Analysis

The last part of the analysis will be the regression models that will show which are the main variables that affect the explanatory ODNR.

4.1. Multiple Regression Model

Firstly, we are interested to see the properties of the saturated model of the variables that are kept in the analysis.

The best model primarily based on AIC and on R-square adjusted seems to be the saturated model of the variables we have kept which is the following:

$$log(ODNR) = \beta_0 + \beta_1 RentalCostRes + \beta_2 TI + \beta_3 Days + \beta_4 Group + \tag{1}$$

$$\beta_5 Rate.title + \beta_6 C.O.Mileage \beta_7 Internet.Insurance.Net.Revenue + \beta_8 Sales.Channel.2 + \epsilon \tag{2}$$

However, due to the fact that the -almost zero- differences of AIC in the last 4 models, we are going to keep a simpler model to continue the analysis.

The final model is the following

$$log(ODNR) = \beta_0 + \beta_1 RentalCostRes + \beta_2 TI + \tag{3}$$

$$\beta_3 Days + \beta_4 Group + \beta_5 Rate.title + \epsilon \tag{4}$$

where TI is the total increments bought.

5. Interpretation

The assumptions of the model are not satisfied, so the intercepts of the variables can be misleading.

However, the interpretation for the quantitative data is the following:

Suppose we want to see how the variable “TI” affects the model. We consider all the other variables equal to 0, and the categorical ones to be at the base category which is set automatically. The effect of an additional TI, will be equal to the intercept of TI.

Now, the interpretation for the qualitative data is different:

Suppose we want to see how the variable “Group” affects the model. We consider all the other variables equal to 0, and the categorical ones to be at the base category which is set automatically except the one that corresponds to the group, which will be the category that we test. Each category has a different number, so it affects differently the model.