Mustafa Abdullayev

USA Cars Assignment 3

AY6010 – Probability Theory and Introductory Statistics

Term: Fall 2019

Instructor: Tom Breur

10/27/2019

Introduction

This is Microsoft Word Report accompanying Microsoft Excel Workbook. In my workbook, I analyzed USA cars dataset. Data was provided by class instructor, Tom Breur, and derived from USA government database. It consisted of information about percentage of household without vehicle and average number of vehicles in 317 USA cities for years 2015 and 2016. Data was cleaned, organized and ready for analysis. With this dataset, I performed various chi-square tests in order to test the independency of car ownership between 2015 and 2016. Moreover, I performed regression analysis. I tested conditions for regression analysis to see either conditions are met or not. Also, I utilized powerful Excel built-in functions and graphs to dive deeper to observe hidden patterns and visually communicate my findings to audience.

Part 1 – Chi-Square Test for Dependency

In order to see if 2015 and 2016 car ownership values are dependent or not, I used Chi-Square test for Independency. Before performing my analysis, I standardized this dataset. Standardization is especially important in regression analysis and it is also very useful for us. This process allows us to put all of our values in the same ground. I converted our values to their z-scores by subtracting mean values and then dividing it with standard deviation. Moreover, I gathered these values to bins according to their z-values. My hypotheses are as follow:

1. H0 : The rows are independent of the columns
2. H1 : The rows are not independent of columns

After calculating chi-square, I found that my p-value is almost 0.97. This is extremely big, and it suggest that we are not able to reject our null hypothesis. This is expected result since I do not see any reason to believe that, 2016 car ownership is dependent on 2015 car ownership.

Part 2 - Regression Analysis

In this part, I performed Regression analysis with Excel. First, I randomly selected 200 data points form population. Then I copied these values (in order to stop refreshing values since I used randbetween() function) as a number. I calculated R squared to 0.96. This means that 94% of variation in 2016 car ownership can be explained by variation in 2015 car ownership.

Afterward, I checked if assumptions of my analysis are indeed satisfied. In the second chart, I observed that, with the increasing predicted values, variance in residuals is also increasing. This is contradictory fact with the homoscedasticity of values in regression analysis. Also, in the first chart I observed that there is not any meaningful pattern in the distribution of residuals. This suggest that these residuals are independent.

Finally, I performed Chi-Square goodness of fit test in order to see if residuals are normally distributed.

Looking at the chart we see that values are skewed.

1. H0 : The observed values are well fit by the distribution of the expected results
2. H1: The observed values are not well fit by the distribution of the expected values

I pooled these residuals into 15 bins in order to perform chi-square test. According to my test, I calculated p-value to be extremely small (p < 0.001). Thus, this imply that I have to reject the null hypothesis. That means our values are not normally distributed and this is the violation of the regression analysis assumptions.

Conclusion

To conclude, I used Excel and its libraries to analyze a USA car ownership data. Data had information about 2015 and 2016 car ownership statistics for USA states and cities. Key finding was that, car ownership rates in 2015 and 2016 are highly correlated. Also, using chi-square test yielded that these values are not dependent on each other. Finally, I conducted residual analysis in order to check the assumptions of regression analysis.