

**ASSIGNMENT FRONT SHEET**

**Course Name: ALY6010 71904 Prob Theory and Intro Stats**

**Professor Name: Dr. Morteza Ziyadi**

**Student Name: Dong Quoc Tuong (Lukas)**

**Student Class: Fall 2019 CPS Term: A. 2019**

|  |
| --- |
| **Module 6: Regression and Correlation Analysis and Analysis of Variance (ANOVA)**  **Completion Date: October 27st Word Count: 936 Due Time:12:00am** |

**Statement of Authorship**

*I confirm that this work is my own. Additionally, I confirm that no part of this coursework, except where clearly quoted and referenced, has been copied from material belonging to any other person e.g. from a book, handout, another student. I am aware that it is a breach of Northeastern University’s regulations to copy the work of another without clear acknowledgement and that attempting to do so renders me liable to disciplinary procedures. To this effect, I have uploaded my work onto Turnitin and have ensured that I have made any relevant corrections to my work prior to submission.*

**Tick here** to confirm that your paper version is identical to the version submitted through Turnitin

1. **Introduction**

The report illustrates how we can perform prediction just like Module 4 and 5 based on the Chi Squares. Then from that on, we will conduct regression and correlation analysis to learn about the associations between the numerical and categorical data that were given to us, justify and explain our conclusion about the hypothesis we propose. We will also examine the difference between Test of “Goodness to fit” and Test of “Independency”. The location quotients of NYC and LA serves as the foundational database for part 1 while a sample of 324 randomly selected professions from the total selected LQ will be used for part 2.

1. **Analysis**

**Part 1**



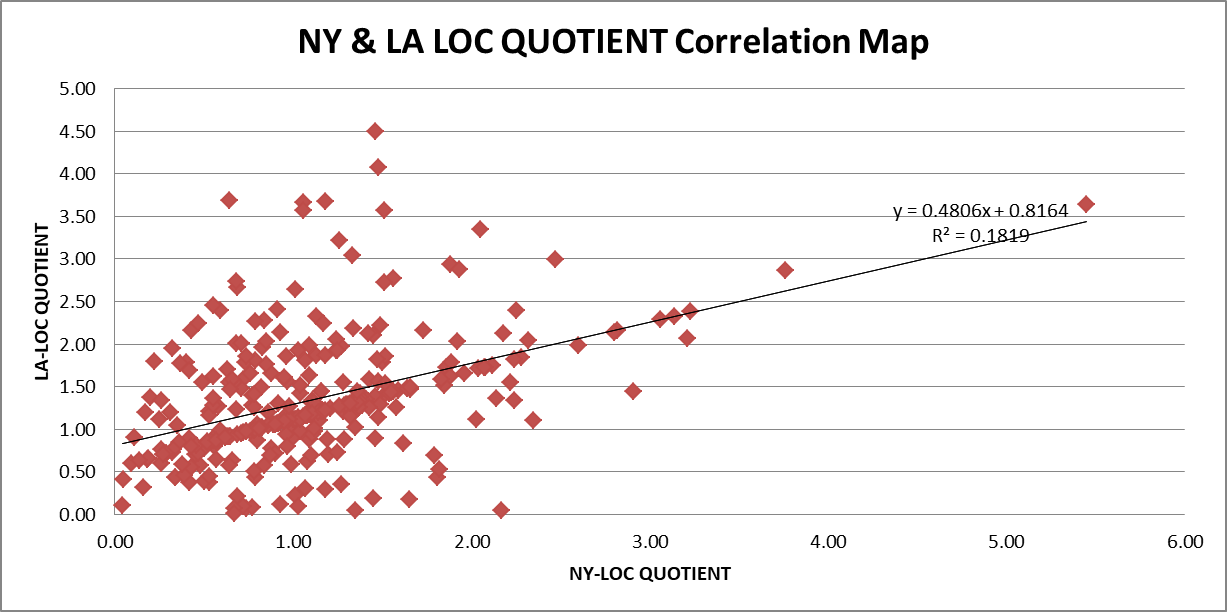


We first draw a sample size of 200 variables from the database for both NYC and LA. Then we find the Mean and Standard deviation for both of them. Sample of LA has a greater mean and standard deviation compared to the NYC’s. Next, we use the standardize formula to put them in the standardized values base on the mean and standard deviation. Since we are trying to disprove that LOC QUOTIENTS and locations are independent variables.(Fahim Zibran, 2007) We put it as our Null hypothesis and the one that we are trying to find is the alternative hypothesis. Next we will count the frequencies of the standardized sample and divide the number into 7 different bins, from -05<= to >4. These are observed values or real values that we can take from the sample. For the expected value of each category in the bin, we will take the total of the bin, multiply it with the total of variables that we took for each groups and divide it to the total of 400 variables we have. Because we take a sample of 200 for each of them, each bin will have similar outcomes for both LA and NYC. Test statistic X2 is calculated by squaring the difference between the Observed and Expected values and then dividing it for the Expected value. X2= 15.9 and X=3.99. We have 7 bins and rows so the degree of freedom is 6. Finally we employ the 1-Chisq.dist to find the P value. Since the P value is less than the Alpha value (0.01<0.05), we can reject the null hypothesis and say that LOC QUOTTIENTS and locations are two dependent factors. We have enough.

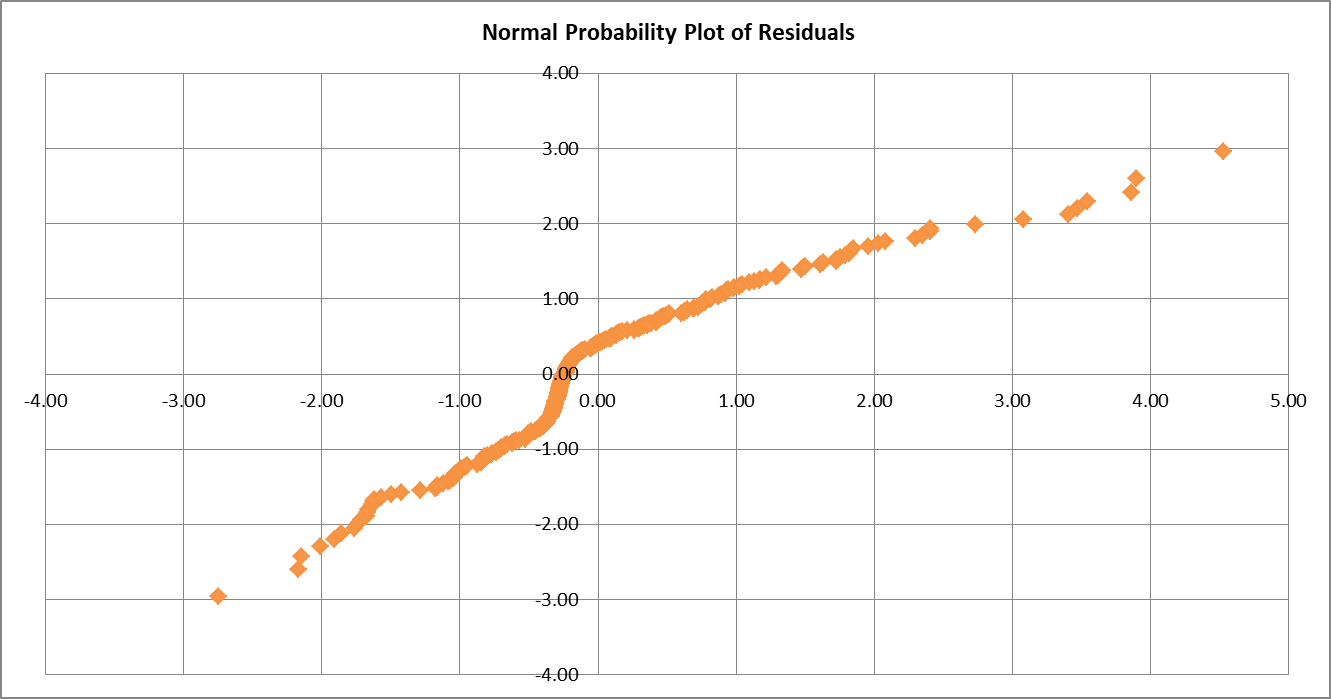


**Part 2**

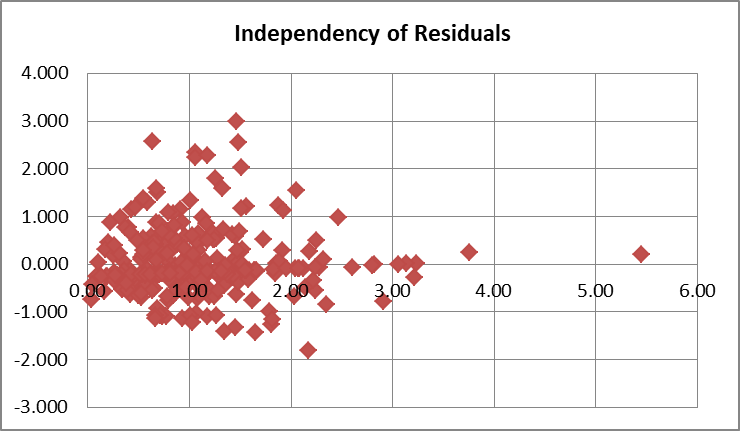


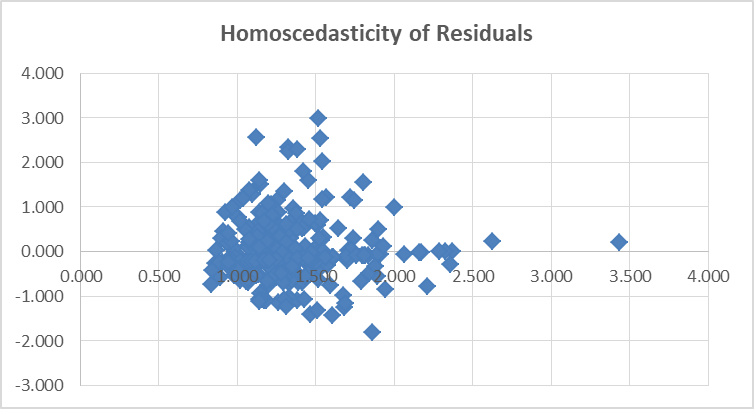
****

We use the Slope, intercept, and correl to calculate the slope, intercept and correlation, respectively. Keeping in mind that NY sample is put in the x-axis and LA is put in the y-axis. Then we have the find the determination R2 by squaring the correlation figure. However, we can just simply draw the correlation map from the sample size. The picture above has shown that the equation for regression line is y=0.4806x +0.8164. A value of 0.4806 for the slope means that, increasing NY LOC QUOTIENT by 1 equals to increasing LA LOC QUOTIENT by 1.3. A value of 0.426 for R indicates a minor positive correlation between these two figures. A value of 0.182 for R2 that only approximately 20% of the variance in the LA LOC QUOTIENT is explained by variations in NYC LOC QUOTE through thanks to the regression model.

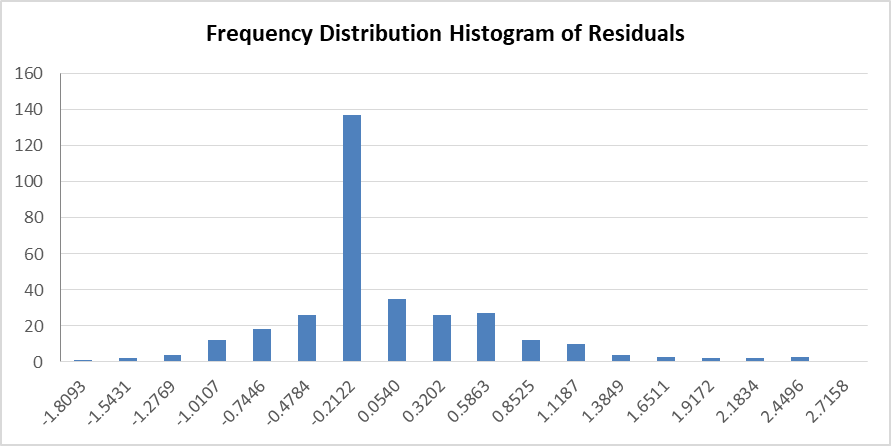


We find the expected value of the LA LOC QUOTIENT by multiplying the slope result to the existing sample’s variable and then plus with the intercept. We basically executing the formula that we just found above. The residual value is the difference between the observed and predicted value of LA LOC QUOTIENT and then we will find the residual mean, standard deviation, minimum, maximum and the number of residual. After that, we will standardize the residual, put them in order and rank them from 1 to the highest position. Take the rank minus 0.5 and then divide everything to the sample size. The Z value is computed by the norm.s.inv formula. We will then proceed to do the chi-square goodness of fit test to see if the observed value fits a certain distribution or not. (Gooch, 2011)So the data can be categorized as normal probability distribution or the binominal distribution. The Independency of Residuals figure shows us that most of the residuals do not stay too far away from 0 except for 1 or 2 outliners, around 1-3 maximum. It is scatter with no pattern, meaning that the residuals are independent. Additionally, the Homoscedasticity of the residuals showed that the data does not meet the standard for homoscedasticity, linearity, and normality since the residual does not form a rectangle but concentrates in one spot.





We calculate everything like Module 1 and see that most the sample size has more variables belong to bin -0.2122 the most rather than any other bins. For the Expected Frequency we take the normal probability value of the bin and multiply to the total number of variables. The Chi Square value is sum of all square of observed value minus the expected value dividing the expected value. The degree of freedom is 16 and the P value is 0. We can safely say the Data is normally distributed.





1. **Conclusion**

In short, this report perfectly sums up that besides the prediction models that we have seen in Model 4 and 5, Chi-Squared testing can be used in the same way to test the correlation between two data sets. Nevertheless, despite its flawless façade, this method’s outcome also depends heavily on the sample. We also executed the Chi Square Test Goodness to Fit. I am confident that after this course, I will be able to carry out research in the future with the good understanding.

**References**

Fahim Zibran, M. (2007). Chi-Squared test of Independence. *Department of Computer Science, University of Calgary, Alberta, Canada*, 1–7. Retrieved from http://pages.cpsc.ucalgary.ca/~saul/wiki/uploads/CPSC681/topic-fahim-CHI-Square.pdf

Gooch, J. W. (2011). Chi-square Goodness of Fit test. *Encyclopedic Dictionary of Polymers*, 973–973. https://doi.org/10.1007/978-1-4419-6247-8\_15175