PROBABILITY THEORY AND INTRODUCTORY STATISTICS

ALY6010, FALL 2019

MODULE 6 PROJECT ASSIGNMENT

REGRESSION AND CORRELATION ANALYSIS & ANOVA

SUBMITTED BY: SHIVANI ADSAR

NUID: 001399374

SUBMITTED TO: DR. DEE CHILUIZA REYES

DATE: 10/27/2019

Introduction

The assignment aims at performing regression and correlation analysis on the US Occupation Data. Using the given data about the occupation's share of employment in a given area and the occupations in the US, we have performed the regression and correlations amongst various parameters. On the basis of the population samples given, we have to calculated the random samples and standardized samples to perform the Chi-Square Test. In order to calculate the values, some parameters like the Observed values, Expected values and their respective statistics have helped in deciding the hypothesis. Moreover, the Location Quotients of both New York and Los Angeles for randomly selected professions, we have used the scatter plots to represent the Normal Probability Plot of Residuals, Independency of Residuals, Homoscedasticity of Residuals and the frequency distribution of the residuals.

Analysis

Part 1: Chi-Square Analysis

Q1. Chi- Square Test for US Population sample

On the basis of the population samples given, we have used the sample size of 600 for New York Location Quotient and a sample size of 520 for Los Angeles Location Quotient. This data was used to calculate the Standardized Sample for New York and Los Angeles using the STANDARDISED function using the formula "=STANDARDIZE(x, mean, standard deviation)" in Excel. The z values were divided into 6 groups ($Z \leq -0.25$, $-0.25 < Z \leq 1.25$, $1.25 < Z \leq 2.25$, $2.25 < Z \leq 3.25$, $3.25 < Z \leq 4.25$, $Z > 4.25$) as observed values in Table A.

Table C: χ2						
	Z ≤ -0.25	.25 < Z ≤ 1.	2.25 < Z ≤ 3.25	3.25 < Z ≤4.25	Z > 4.25	TOTAL:
NY	0.3	1.0	1.8	0.1	1.6	7.9
LA	0.3	1.2	2.1	0.1	1.9	9.1
TOTAL:	0.5	2.3	4.0	0.2	3.5	17.0

Table 1: Displays chi-squared value calculated for NY and LA Location Quotients

According to the conditions given, the Observed values of location quotients were calculated by using the COUNTIFS function.

Moreover, the expected values of location quotients were calculated on the basis of Total values obtained for both Los Angeles and New York.

The test statistic values for chi-squared distribution have been calculated on the basis of the Observed and Expected values.

As per the formula for Chi-Square distribution,

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where, O : Observed Values, E : Expected Values, X^2 : Chi-Square Distribution

Table D: STATISTICS	
Test Statistic χ^2	17.0
Degrees of Freedom DF	5
P-value	0.004439

Table 2: Test Statistics values for testing the hypothesis

- Here, the Null Hypothesis, H0 is LOC Quotients and the locations are Independent Factors and the Alternate Hypothesis, Ha is LOC Quotients and locations are not Independent Factors

- Since, the chi-square calculated P-value ie. 0.004 which is lesser than the critical value of 0.03, we will reject the null hypothesis.
- This shows that there is a dependency amongst standardized location quotient and location with a chi-square test statistics value of 17.

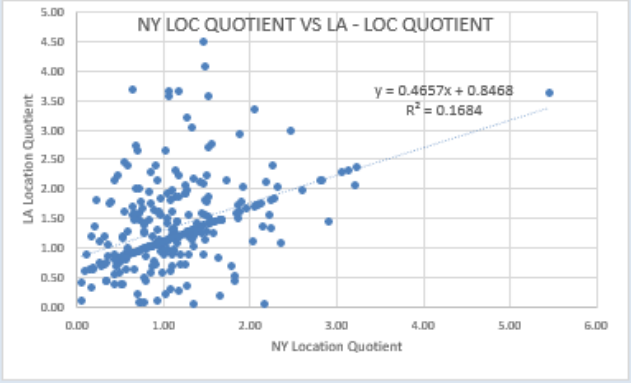
Application

- The chi-square test for categorical variables determines whether there is a difference in the population proportions between two or more groups.
- For example: In medicine, A chi-square test can be used to compare the incidence between the patients that received Ondansetron, Patients that received Droperidol and patients that received a placebo.
- The chi-square test is a statistical hypothesis test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true. A common usage of the Chi-square test is the Pearson's chi-square test, also known as the chi-square goodness-of-fit test or chi-square test for independence. The Chi square test is used to compare a group with a value, or to compare two or more groups, always using categorical data.

Part 2: Regression and Correlation

Q1&Q2 Location Quotients for New York and Los Angeles for randomly selected professions

On the basis of the NY LOC Location Quotient and the LA LOC Location Quotient, the parameters like the Slope, Intercept, Correlation and Determination have been calculated. Using the random samples of 280 location quotient for LA and NY. In the Table 3, the Coefficient of Correlation of 0.41, Determination of 0.16, Slope of 0.46 and Intercept of 0.84 were calculated. The graph shows a positive correlation which means that on increasing the location quotient of NY, the location quotient of LA will decrease.



(1) Table A	
Slope m	0.4657
Intercept b	0.8468
Correlation R	0.4103
Determination R^2	0.1684

Table 3: The table shows the Slope, Correlation, Intercept & Determination

Fig.1: Scatter Plot showing the location quotients for NY and LA

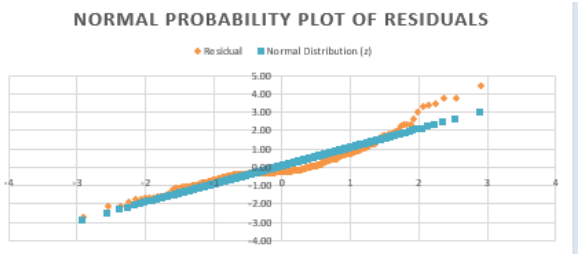
As can be seen from the Figure 1 and Table 3, the slope of the line is 0.4657 which has an intercept of 0.8468 with a correlation of 0.4103. The residual is calculated as 0.1684 using the formula, 'RESIDUAL=OBSERVED-PREDICTED'. As can be seen from the figure, there is a positive correlation. The regression line shows the positive correlation between New York and Los Angeles location quotients.

Q3&Q4 Simple Regression and Normal Probability Plot of Residuals

The slope and intercept of Los Angeles Location Quotient were used to calculate the predicted values. In Table C, the Residual Mean of 0.00, Residual Standard Deviation of 0.67, Redidual Minimum of -1.807, Residual Maximum of 2.973 and Residual Count of 280 were calculated.

Q5. Normal Probability Plot

We have plotted a normal probability plot of residuals on the basis of Table C.



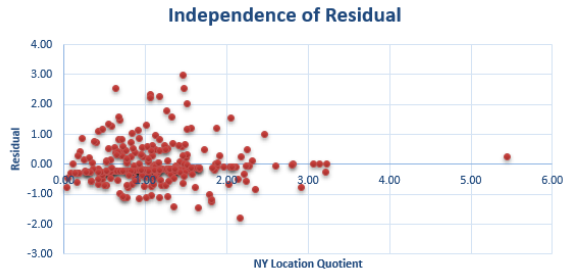
(4) Table C	
Residuals Mean	0.00000
Residuals SD	0.67256187
Residuals Minimum	-1.807
Residuals Maximum	2.973
Residuals Count	280

Fig.2: Normal Probability Plot of Residuals

Table 4: Calculations of Residual Parameters

The Fig.2 indicates the scatter plot of normal probability plot of residuals. As can be seen, the blue dots show the standard normal distribution of values and the yellow line denotes the standardized residuals. It can be observed that the two lines coincide with each other that is a representation that the residuals are normally distributed.

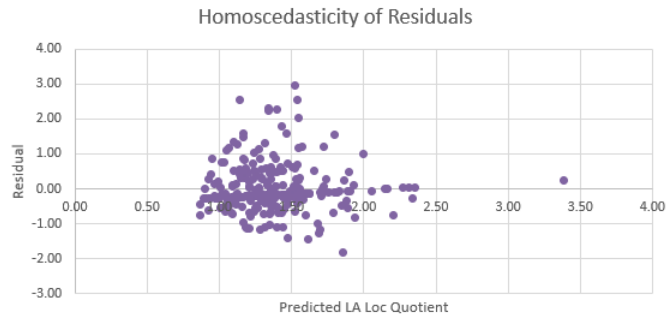
Q6. Calculation of Independency of Residuals



According to Fig.3, it can be observed that the two parameters ie. Residual and NY Location Quotient are independent from each other. The figure does not show any relationship between the two parameters.

Fig.3: Scatter Plot of Independency of Residuals VS X values, NY Location Quotient

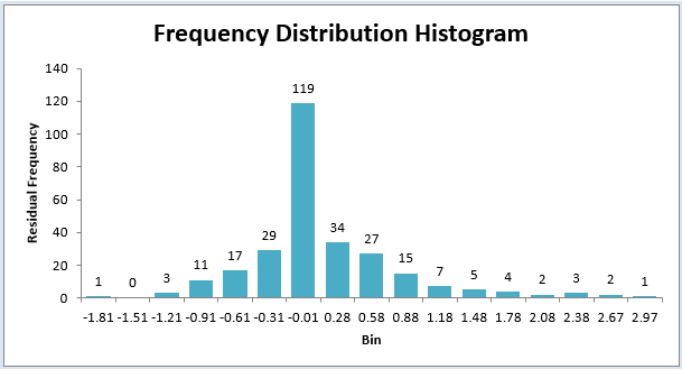
Q7. Homoscedasticity of Residuals



We are using the test of Homoscedasticity to check the Homoscedasticity of variance between the error terms and the independent variables. According to the Fig.4, shows a scatter plot representing the Homoscedasticity of residuals. The purple dots mentioned in the graph show the Residual and Predicted Los Angeles Quotient values. The graph indicates that the condition of Homoscedasticity is satisfied by the residuals as we can see that no patterns exist between the Residual and the Predicted LA Location Quotient.

Fig.4: Homoscedasticity of residuals VS Predicted Y values

Q8. Frequency Distribution of Residuals



The Fig.5, represents the frequency histogram of residual values. It can be observed that the highest frequency of 119 is seen at the bin range of -0.01. Also, it can be observed from the graph that the frequencies around the mean value show higher values of frequencies as compared to the frequencies at the ends.

Fig. 5: Frequency Histogram of Residual Values

Q9. Chi-Square Goodness of Fit Test

(9) Table E	
χ^2	675.12
DF	15.00
P-value	0.00E+00

Table 5: This table shows the chi-square goodness of fit test for normality of residuals. As seen, the chi-square value of 675.12, Degree of Freedom of 15 and P-Value was calculated using the formula, 1-CHISQ.DIST(X^2,DF,1)

Table 5: Chi-Square Goodness of Fit Test

The hypotheses is used to evaluate the normality distribution of residuals data. Considering, (α=0.03)
Ho: Residuals data belongs to a normal distribution , Ha: Residuals data does not belong to a normal distribution
It can be observed that, this is right tailed test as the chi square value is as greater, approx. 675.12. Using that the p value 0.00 is less than the level of significance, hence we will reject the null hypothesis. It can be noted that, the residual data does not follow a normal distribution

Conclusion

1. The null hypothesis that the Location Quotient and location are independent factors was rejected as we noted that the calculated p-value was lesser than the significance level.
2. A positive correlation coefficient of 0.41 was obtained between New York and Los Angeles location quotient, which shows that the two variables are directly proportional to each other.
3. Similarity was observed between the graphs of Independency and Homoscedasticity as there is a positive correlation between the X and Y parameters.
4. As the P-value was lower than the significance level, the null hypothesis was rejected.

Reference

1. Bulman, A. G. (n.d.). Probability and Counting Rules. In ELEMENTARY STATISTICS: A STEP BY STEP APPROACH, TENTH EDITION (10th ed., p. A-440). New York
2. Bruce E. Trumbo California State University, Hayward Journal of Statistics Education v.3, n.2 (1995) Retrieved from https://www. http://jse.amstat.org/v3n2/trumbo.html
3. A. S. Holevo, “On asymptotically optimal hypotheses testing in quantum statistics”, Teor. Veroyatnost. i Primenen., 23:2 (1978), 429–432; Theory Probab. Appl., 23:2 (1979), 411–415