Applied Statistical Analysis - ProblemSet 2

Eoghan O'Sullivan

October 14, 2024

1 INTRODUCTION

The problem set was two-fold insofar as two research scenarios were provided as problems. The first of these two problems required coding in R high-level language and analysis in other tools. It also required supplemental data to be included. The second problem also required data analysis in R in terms of bi-variate regression and conducting a chi-squared test of independence.

1.1 QUESTION 1 - Political Science

This problem I approached by using R to perform statistical analysis on data that were given in the problem set description. I loaded a modified data set in RStudio which had these aforementioned data with an extra observation. I ran a correlation on the two variables that were of interest: Bribe requested and Stopped/given warning. The code I wrote produced an answer to the first part of the problem one. The below test statistic was generated from there. Note that the extra observation included in the data loaded was for middle class in addition to the upper and lower classes already included and that this observation had values equaling zero. These data enclosed with submission.

$$t = 0.57735 \tag{1}$$

The same analysis in R produced a p-value from the test statistic and I interpreted the result.

$$p - value = 0.6667 \tag{2}$$

The p-value shows that the probability is positive because 1-a at 0.1 is not resulting in a Type 1 or Type 2 Error.

The residual sum of squares formula below was used to calculate the residuals for each cell in the table. $RSS = \sum(y-\bar{y})^2$

Table 1: Table of Residual Sums of Squares

	N/a	**Res**	**Bribe**	**Res**	**Stopped/Warned**	**Res**
Upper Class	14.00	28.00	6.00	-16.00	7.00	13.00
Lower Class	7.00	-45.50	7.00	-9.50	1.00	-11.00

I therefore calculated the mean for each variable in the data set. The formula was used in an Excel spreadsheet and converted to formulae interpreted in Excel. I also did regression analysis in Excel using a ToolPak Add-in and the results are inconclusive with these data. However the output results are submitted.

My interpretation of the results of standardised residuals is that the model fits its data.

1.2 QUESTION 2 - Economics

My null hypothesis is that there is no association between gender and reserved variables (Ho = 0) whereas the alternative is that female gender and reserved variable are correlated.

My assumptions are that the type of data is appropriate, that it contains randomisation and that it represents a population enough to infer from it in my analysis. I have calculated the below test statistic to test the null hypothesis.

$$t = 255.59 \tag{3}$$

The resultant p-value was calculated using a chi-square test of independence in R. This had as it's forerunning activity contingency table formulation in R. I compared the Female binary variable to the Reserved variable in my Pearson chi-squared test, having read in these related data using an online location. This p-value was sufficiently small contradict the null hypothesis as it is a small number.

$$p - value = 1.569182e - 57 \tag{4}$$

To interpret and firstly calculate the correlation coefficient, I used Excel using an x variable, in this case, Female and a y variable, Reserved. The following formula represents the correlation coefficient, r.

$$r = \frac{1}{n-1} \sum (\frac{x - \bar{x}}{Sx})(\frac{y - \bar{y}}{Sy}) \tag{5}$$

The results of the correlation analysis points to a negative correlation overall. The results are enclosed. There was a positive correlation in the following West Bengal Gram Panchayat in both the first and second villages: 26,84,86,92,107,119,142, and 145.