CA660 Statistical Data Analysis

December 14, 2016

We declare that this material, which we now submit for assessment, is entirely our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. We understand that plagiarism, collusion, and copying is a grave and serious offence in the university and accept the penalties that would be imposed should we engage in plagiarism, collusion, or copying. We have read and understood the Assignment Regulations set out in the module documentation. We have identified and included the source of all facts, ideas, opinions, viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references.

We have not copied or paraphrased an extract of any length from any source without identifying the source and using quotation marks as appropriate. Any images, audio recordings, video or other materials have likewise been originated and produced by me or are fully acknowledged and identified.

This assignment, or any part of it, has not been previously submitted by us or any other person for assessment on this or any other course of study. We have read and understood the referencing guidelines found at http://www.library.dcu.ie/citing&refguide08.pdf and/or recommended in the assignment guidelines.

We understand that we may be required to discuss with the module lecturer/s the contents of this submission.

I/me/my incorporates we/us/our in the case of group work, which is signed by all of us.

Signed: RK Vidya , Vikas Chhillar Date: 14/12/2016

Exploratory Analysis on population dataset

RK Vidya Priyadharshini
16212606
Group No: 500

Vikas Chhillar
16212887
Group No: 500

December 15, 2016

We declare that this project was a combined effort by both of us. And both of us contributed equally in all phases of the assignment from the selection of the datasets, setting up the hypothesis, discussing which statistical methods to apply for the datasets to the identification of the inference from the results obtained.

Contents

1 Introduction

Statistical Data Analysis involves using the statistical tools to explore and analyse the large datasets generally population census data or financial data. There are six types of analysis that can be performed on datasets. They are descriptive, exploratory, inferential, predictive, causal and mechanistic arranged according to the complexity in performing these analysis. Descriptive analysis is the first step of any analysis that is performed on any dataset. It gives us the big picture of any dataset like the various measures of location (mean, mode, median) and measures of dispersion (variance, standard deviation, skewness, interquartile range, etc.,). Exploratory analysis helps us to find answers to the various questions posed by the dataset based on the critical thinking of the analyst. This helps us in identifying the patterns of the data thereby determining if anomalies or outliers exist in the data. However, the exploratory analysis does not gives us a definite answer to the questions. It is only the beginning of any analysis and further inferential or predictive analysis needs to be done in order to verify and confirm the inferences of the analysis. Let us see the exploratory analysis done on a census data in this assignment.

2 Data Source

The data was downloaded from the Central Statistics Office of Ireland website. The dataset name is "CDD06. Population of each province, county and city, classified by sex, age group and detailed marital status". The dataset is a multivariate dimensional since the values are measurements made on many variables per subject. It consists of discrete variables like Marital Status (Single, Married, Divorced and Widowed), different age groups and the gender based on which the census data is recorded for 8 cities in Ireland.

3 Methodology

The methodology followed in this assignment is a step by step analysis as follows,

1. Descriptive Analysis: The fair idea of the data-points in the dataset is obtained by the descriptive analysis. The basic statistician tools are found out by use of excel or by importing the dataset into R. The central tendency measures are mean, median and mode. The variability measures which give us the spread of the data are range, variance, standard deviation and inter quartile range. Percentile and Quartiles depicts how a particular data point is placed in relation with the other data points in the dataset.

2. Exploratory Analysis: After calculating the measures of location and spread, we understand the how the data is presented in the dataset. Now, we analyse the key questions the data poses. Since the data contains the marital status and the age groups with the categorisation of gender, we can analyse the marital patterns of both male and female in different cities. On a fair analysis, we are able to find out that the city Dublin has the highest population amongst the other cities. And for the various other analysis we can use suitable statistical methods in order to find answers.

3.1 Hypothesis

Based on the questions for which we need answers the hypothesis is set up or assumed in the assignment. Here we do not have a single hypothesis. The various hypotheses we assumed in order to carry out the statistical tests are,

1. The gender of the population affects the divorce decisions.
2. Males and Females have different opinions on when to get married.
3. The idea to remain single differs in various age groups.
4. The idea to remain single differs in various groups based on the gender.
5. The number of people getting married at an early age is less.
6. The idea to get divorce or married is dependent on the age.

3.2 Choice of Methods

The choice of methods depends on how and what variables both independent and dependent, we are considering for evaluating our hypothesis. The choice of methods depend on the following factors,

1. Type of data.
2. Dependency of the data.
3. Availability of control variables.

So, now let us analyse what methods could be used to statistically accept or reject our hypotheses.

3.2.1 Hypothesis 1

The gender of the population affects the divorce decisions.

This hypothesis revolves in identifying the differences by comparing the observed frequencies with the expected frequencies. And both the variables are categorical in nature with no control variables available. So the best method for this statistical analysis would be Chi Square Test.

3.2.2 Hypothesis 2

Males and Females have different opinions on when to get married. This involves identifying the differences between two groups. Considers one independent variable and one dependent variable. Thus the best test in here would be the *t*- Test. Another hypothesis to be considered here is the identification of differences between the age groups 20-29 years and 70-79 years.

3.2.3 Hypothesis 3

The idea to remain single differs in various age groups.
The rates of marriage vs divorced between three different age groups.

The above two hypothesis involves the exploration of more than two different groups of ages and thus ANOVA is the best statistical test that can be used here. It indicates to us only if there is a difference between the groups but not exactly the group which is different.

3.2.4 Hypothesis 4

The idea to remain single differs in various groups based on the gender.

The above hypothesis has a control variable and involves exploration of two or more groups. So the statistical test used here is the ANCOVA which is similar to ANOVA i.e, Analysis of Variance but with an additional control variable.

3.2.5 Hypothesis 5

The city with the highest and the lowest rate of marriage and divorce are Dublin and Kilkenny respectively.

Here, we are using the Linear Regression method to identify the city with highest rate of marriage and divorce.

3.2.6 Hypothesis 6

The idea to get divorce or married is dependent on the age.

3.3 Limitations

The limitations of the choice of methods are that the Analysis Of Variance method tells us if the groups are different but does not indicate to us which group is different. So we have to use the other statistical methods to find out which group is significantly different from the other.

3.4 Implementation

The following steps are part of our implemenation,

3.4.1 Data collection

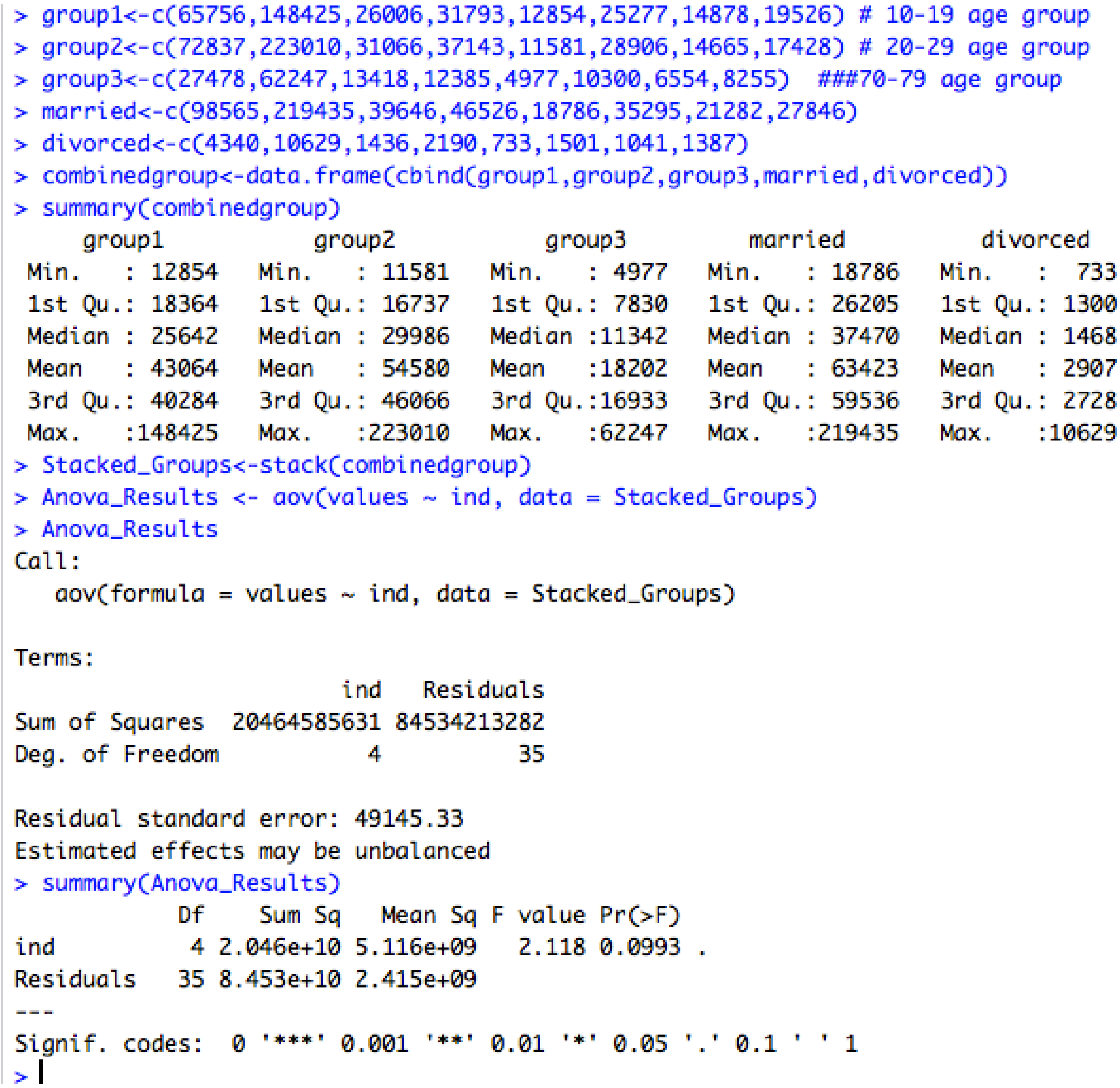
After the collection of our data, when we had tried to import it in R as a .csv file we had encountered difficulties. The error saying that the data had too many headers was displayed. And so we had to re structure the data such that the population count of the Single, Married, Divorced, Widowed, Age groups 0-9 years, 10-19 years, 20-29 years, 30-39 years, 40-49 years, 50-59 years, 60-69 years, 70-79 years , above 80 years , Male and Female based on the 8 cities Cork, Dublin, Dun-Laoghaire-Rathdown, Galway, Kilkenny, Limerick, Waterford and Wexford had been arranged in a single row by column method.

3.4.2 Statistical Test

We had done the following statistical tests,
ANOVA: The five columns are taken for the analysis. The first three group contains the age group of different people from different cities. Two other column gives the data about married people and divorced people. The different column are combined using ”cbind”. All the columns are placed in the stack. The function for using the anova is used for calculating the anova. The null hypothesis is that the the mean for all age group for married and divorced is same. As there are more than 2 columns therefore we are not using t-test. For the analysis, the variation within group is also seen. The larger the ratio of difference between group and within groups, the more likely is that the groups have different mean(rejection of null hypothesis).

Linear Regression: An approach for modelling the relationship between x and y values. Marriage and Divorce rates for women and men are taken from different places in Ireland. From the data, the scatter plot is created, and the relationship between the two variables (i.e., married and divorce rates). The residual plot explains the meaning of the slope and of the y-intercept of the line of best fit, and the effect of outliers on this line can be seen in the graph. The horizontal line in the graph is the mean of the divorce and vertical line is mean of people of the age group (20 to 29).

Which place will have higher or lower rates for marriage and divorce? Are there any states whose rates will be high or low?

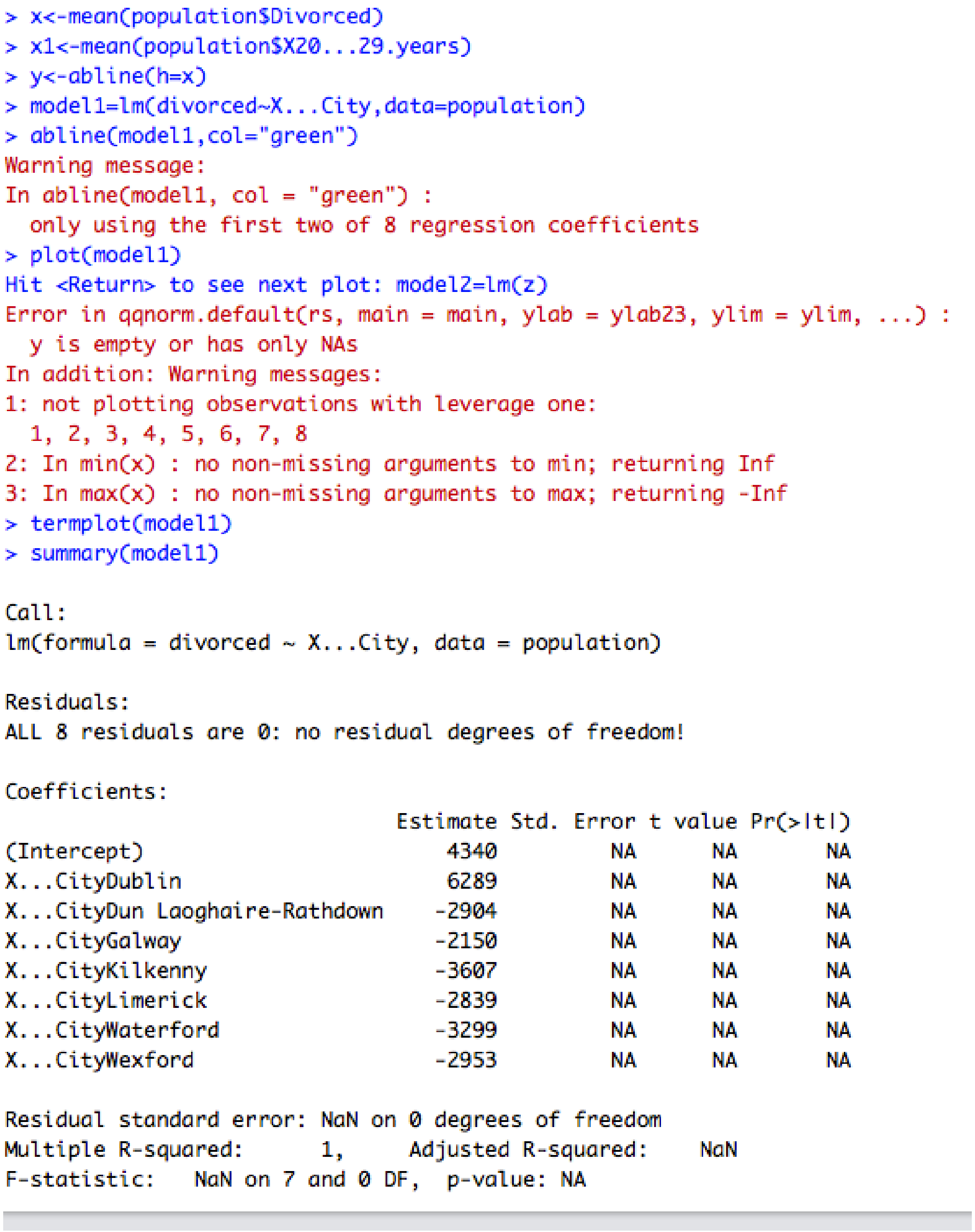


```
> group1<-c(65756,148425,26006,31793,12854,25277,14878,19526) # 10-19 age group
> group2<-c(72837,223010,31066,37143,11581,28906,14665,17428) # 20-29 age group
> group3<-c(27478,62247,13418,12385,4977,10300,6554,8255)  ###70-79 age group
> married<-c(98565,219435,39646,46526,18786,35295,21282,27846)
> divorced<-c(4340,10629,1436,2190,733,1501,1041,1387)
> combinedgroup<-data.frame(cbind(group1,group2,group3,married,divorced))
> summary(combinedgroup)
     group1           group2           group3          married           divorced    
 Min.   : 12854   Min.   : 11581   Min.   : 4977   Min.   : 18786   Min.   :  733  
 1st Qu.: 18364   1st Qu.: 16737   1st Qu.: 7830   1st Qu.: 26205   1st Qu.: 1300  
 Median : 25642   Median : 29986   Median :11342   Median : 37470   Median : 1468  
 Mean   : 43064   Mean   : 54580   Mean   :18202   Mean   : 63423   Mean   : 2907  
 3rd Qu.: 40284   3rd Qu.: 46066   3rd Qu.:16933   3rd Qu.: 59536   3rd Qu.: 2728  
 Max.   :148425   Max.   :223010   Max.   :62247   Max.   :219435   Max.   :10629  
> Stacked_Groups<-stack(combinedgroup)
> Anova_Results <- aov(values ~ ind, data = Stacked_Groups)
> Anova_Results
Call:
   aov(formula = values ~ ind, data = Stacked_Groups)

Terms:
                        ind   Residuals
Sum of Squares  20464585631 84534213282
Deg. of Freedom           4          35

Residual standard error: 49145.33
Estimated effects may be unbalanced
> summary(Anova_Results)
            Df    Sum Sq   Mean Sq F value Pr(>F)  
ind          4 2.046e+10 5.116e+09   2.118 0.0993 .
Residuals   35 8.453e+10 2.415e+09                 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 
```

Figure 1: R code for ANOVA



```
> x<-mean(population$Divorced)
> x1<-mean(population$X20...29.years)
> y<-abline(h=x)
> model1=lm(divorced~X...City,data=population)
> abline(model1,col="green")
Warning message:
In abline(model1, col = "green") :
  only using the first two of 8 regression coefficients
> plot(model1)
Hit <Return> to see next plot: model2=lm(z)
Error in qqnorm.default(rs, main = main, ylab = ylab23, ylim = ylim, ...) :
  y is empty or has only NAs
In addition: Warning messages:
1: not plotting observations with leverage one:
  1, 2, 3, 4, 5, 6, 7, 8
2: In min(x) : no non-missing arguments to min; returning Inf
3: In max(x) : no non-missing arguments to max; returning -Inf
> termplot(model1)
> summary(model1)

Call:
lm(formula = divorced ~ X...City, data = population)

Residuals:
ALL 8 residuals are 0: no residual degrees of freedom!

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                        4340         NA      NA       NA
X...CityDublin                     6289         NA      NA       NA
X...CityDun Laoghaire-Rathdown    -2904         NA      NA       NA
X...CityGalway                    -2150         NA      NA       NA
X...CityKilkenny                  -3607         NA      NA       NA
X...CityLimerick                  -2839         NA      NA       NA
X...CityWaterford                 -3299         NA      NA       NA
X...CityWexford                   -2953         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:    NaN 
F-statistic:   NaN on 7 and 0 DF,  p-value: NA
```

Figure 2: R code for Linear Regression

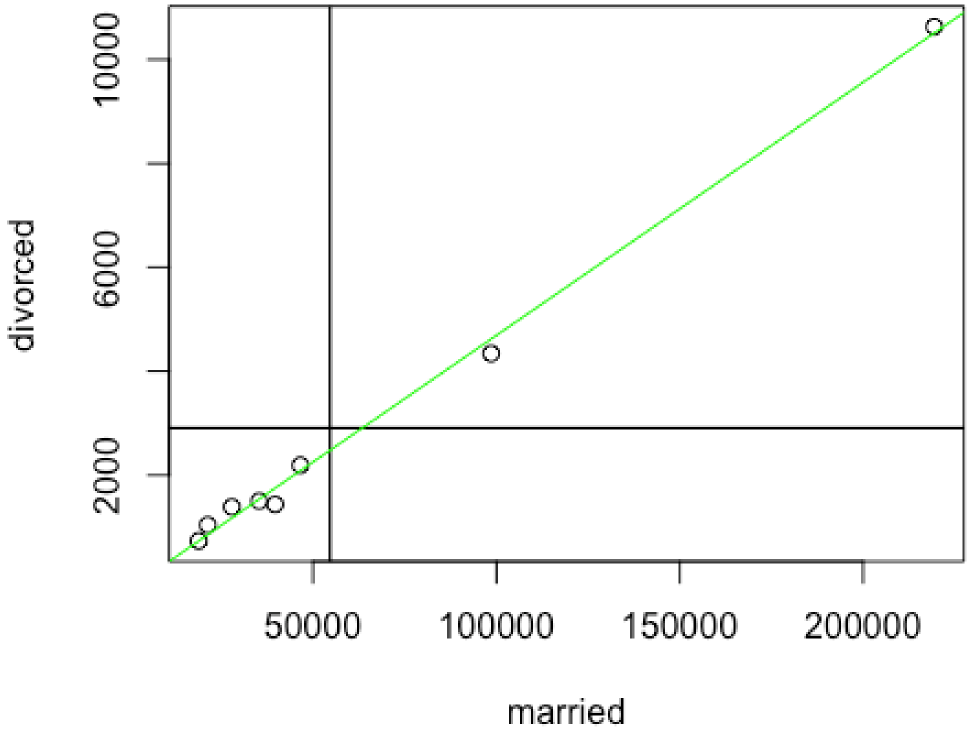


Figure 3: Linear Regression- Married vs Divorced

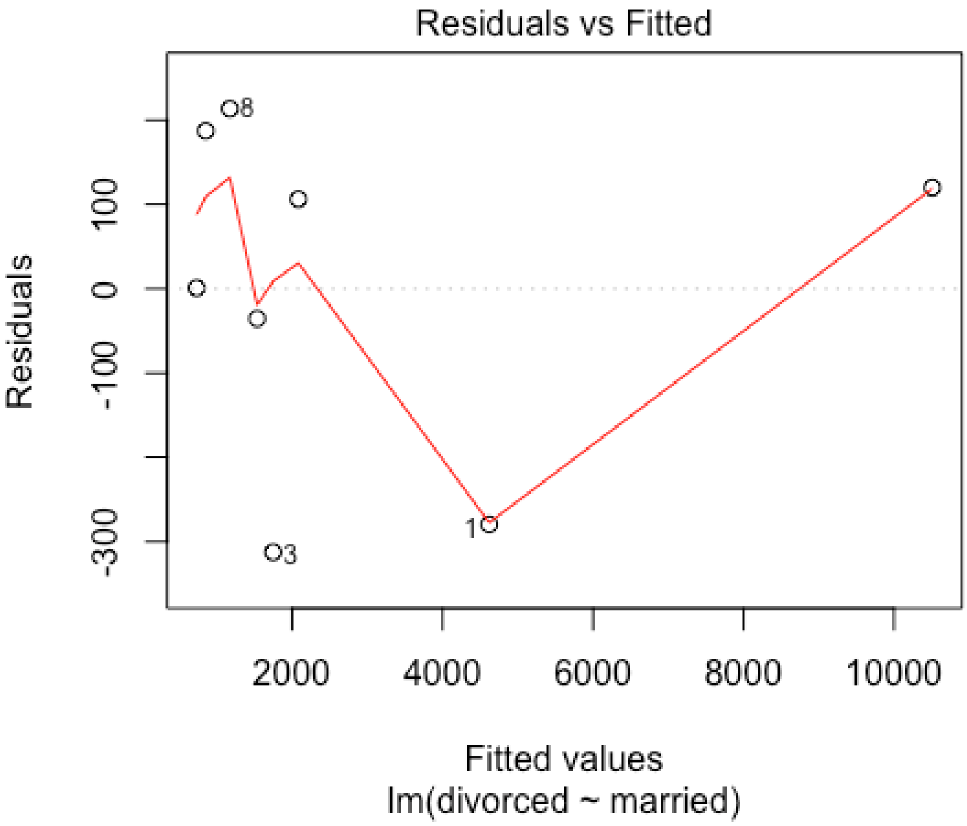


Figure 4: Residuals vs Fitted

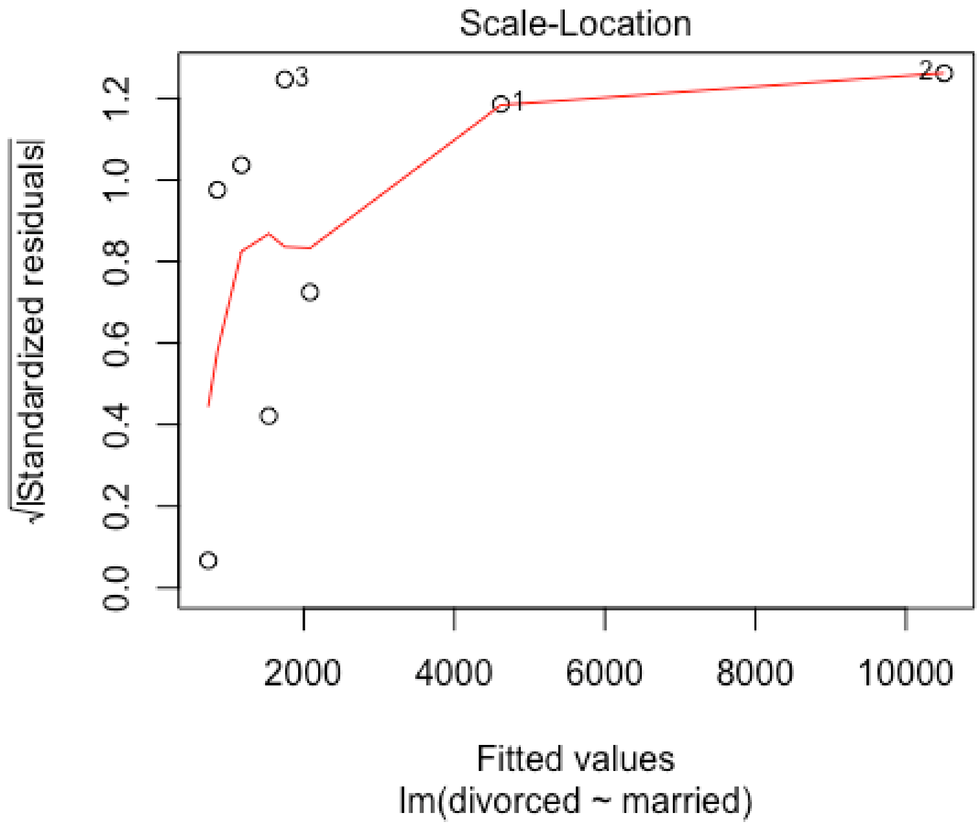


Figure 5: Scale-Location

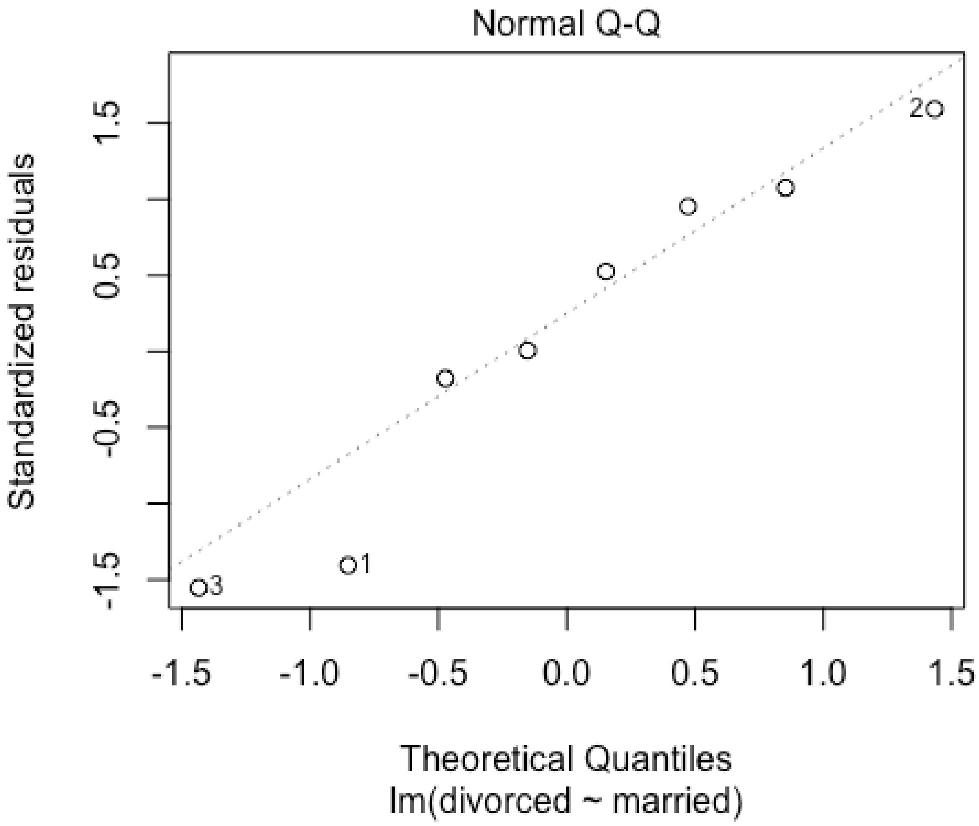
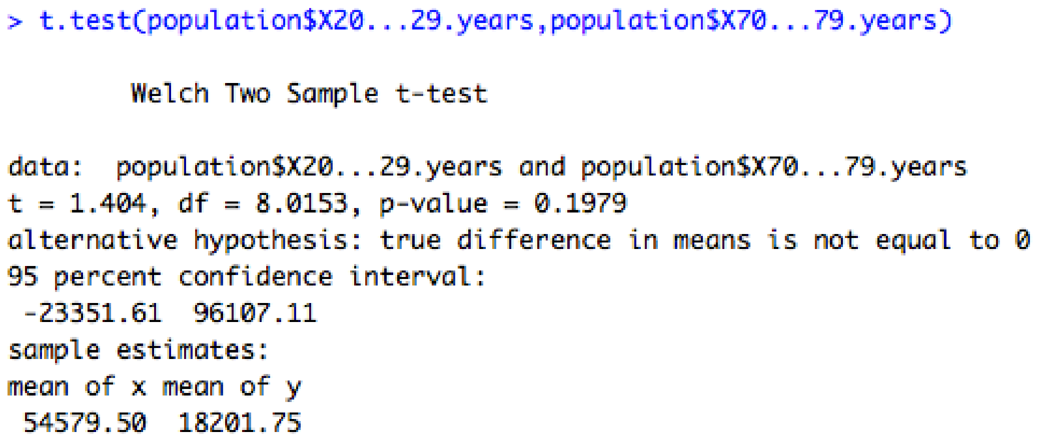


Figure 6: Normal Q-Q

t- Test: This test is done to check if two columns are different from each other. And this consists of one independent variable dichotomous in nature and one dependent variable continuous in nature. In our dataset, we analyse the two age groups 20-29 years and 70-79 years.



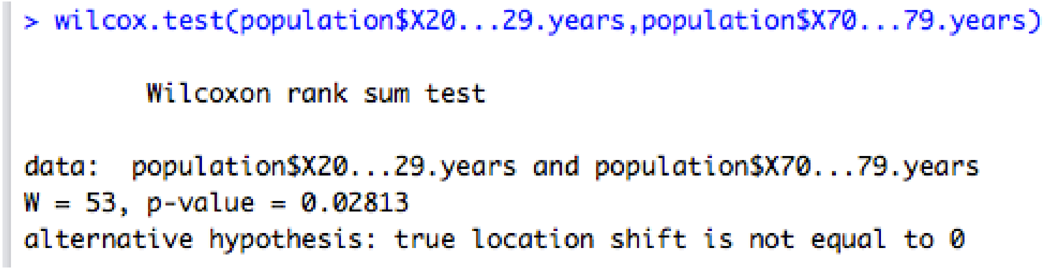
```
> t.test(population$X20...29.years,population$X70...79.years)

        Welch Two Sample t-test

data:  population$X20...29.years and population$X70...79.years
t = 1.404, df = 8.0153, p-value = 0.1979
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -23351.61  96107.11
sample estimates:
mean of x mean of y 
 54579.50  18201.75 
```

Figure 7: R code for t test.

Wilcoxon Rank Test: Here the medians of two age groups are compared by using non-parametric method of assigning ranks to the data values. The age groups 20-29 years and 70-79 years are considered here.



```
> wilcox.test(population$X20...29.years,population$X70...79.years)

        Wilcoxon rank sum test

data:  population$X20...29.years and population$X70...79.years
W = 53, p-value = 0.02813
alternative hypothesis: true location shift is not equal to 0
```

Figure 8: R code for Wilcoxon Rank Test.

4 Results and Discussion

Based on our statistical tests we have found the answers to the several questions that we had asked about the dataset in the initial stage of analysis.
The gender of the population affects the divorce decisions- Yes, the gender has an effect on the rate of divorce in each city.

Males and Females have different opinions on when to get married,
The number of people in the age group between 20-29 years is significantly higher that the number of people in the ages between 70-79 years- Yes by the result of the t- test we can see that both the groups are significantly different.

5 Conclusion

We conclude by saying that the exploratory analysis on the popultion dataset has helped us critically think and resolve various questions posed by the dataset with use of the statistical tests. And we were able to do the calculation using the software R and using the add-in 'Data Analysis' in Microsoft Excel.

References

[1] M. J. Crawley, *Statistics: An Introduction Using R*, Wiley-Blackwell, Chichester, West Sussex, England, 2005.

[2] N. J. Gotelli and A.M. Ellison, *A Primer of Ecological Statistics*, Sinauer Associates Publishers, 2004.

[3] Central Office of Statistics, Ireland, URL:http://www.cso.ie/en/census/census2011reports /census2011thisisirelandpart1/

[4] I. Kabacoff, *R in Action, Data analysis and graphics with R*, Second Edition, 2015.