Influence of Smoking on Death Rate with Respect to Age

Part 1

1. Introduction

A two year cohort study on the relationship between Human smoking Habits and Death rates was conducted on males aged between 50 and 69 years in the year 1954 by E.C Hammond and D.Horn for two major reasons. Firstly, to find out if smoking influences death rate from lung cancer. Secondly, to find out if smoking has a considerable effect on the overall death rates and on the diseases caused for the death. The study was conducted on 187,766 men. The death toll due to smoking in the United States, accounts for more than 480,000 every year, or about 1 in 5 deaths which is a huge health concern and needs more attention in this society. The data from this study, will help us determine the death rates due to smoking in different age groups and the effect it has on death compared to the death rate of men who don’t smoke.

Data source:

http://users.stat.ufl.edu/~winner/data/smkdth.txt

http://users.stat.ufl.edu/~winner/data/smkdth.dat

https://jamanetwork.com/journals/jama/article-abstract/295889

2. Problem Statement

The Problem of our experiment is to find out how smoking influences death rate with respect to age factor.

3. Description and Purpose

Our experiment is aimed at analyzing the effects of smoking on the death of an individual with respect to age. This problem is scientifically significant because if there is a direct relationship between smoking and death rate in a specific age group, we can imply the correlation between deaths of individuals who smoke versus the non-smoking ones. Also, we can analyze the trend of deaths due to this habit over the ages and raise awareness. If there is no direct correlation existing from our experiment, the study must be extended to analyze the cause of the death and if there is any other predominant disease that prevails that needs to be made aware of.

4. Hypotheses

a. What is the prediction, or guess, about the outcome of the experiment?

The prediction is that influence of smoking increases death rate with respect to increase in age factor.

b. Is the prediction logical?

Yes, the prediction is logical.

c. This statement should be written in future tense, using an “If/then” or prediction format.

If an individual smoke, then there is higher chance of their death in early ages.

Part 2

1. Identification of Variables

The dependent variable will be number of cases observed for a behavior. Independent variables consist of Smoking status (0 = No, 1 = yes), Age classification (1=50-54, 2=55-59, 3=60-64, 4=65-69) and Death status (0 = Survival, 1 = Death).

Independent: Smoking status, Age classification, Death status

Dependent: Number of cases

2. Factors and Levels

There are three factors in this study. First, Smoking status with two levels, value 0 for non-smoker and 1 for smoker. Second, Age classification with four levels of age grouping in it, value 1 for age group 50-54, 2 for age group 55-59, 3 for age group 60-64 and 4 for age group 65-69-year-old. Third, Death status with two levels, 0 for survival and 1 for death.

3. Population Specification

The population contains death case numbers of men aged 50-69 years who smoke and don't. Because the study is solely carried out on men, the blocking factor of gender doesn't play its role here. It includes population of people who do not smoke whom we can refer as control group. Using mixed-design analysis of ANOVA, analysis will be performed to test the variances between the groups and with-in the groups.

Part 3

1. Literature Review

Question. Please find 4-6 scholarly articles that helps in determining the need of your experimentation. In another word, find the gap in research done to date that led you to be interested in the research question(s).

The studies done till date state the fatal effects of smoking on the physical and mental fitness of the smokers. The studies have been carried out to understand the associations of smoking with lung cancer, bladder and pancreas cessation and heart diseases leading to an untimely death. Our study, on the other hand focusses on the association of smoking, age and death for men aged 50 years and above and will conclude that smoking in which age bracket is going to affect the smoker the most.

https://jamanetwork.com/journals/jamainternalmedicine/article-abstract/615878

https://www.popline.org/node/510062

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2549910/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2541135/

https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/224836

2. Procedures of the Experiment Testing

Question 1. Selection of the treatment design method.

As described in Part 2 document, the three factors have multiple and different levels in them. Smoking Status has two, Age Classification has four and the factor Death Status has 2 levels. To cater the analysis of multiple factors with multiple levels, we would have to go for two-way ANOVA. To match our treatment to test the variance between the groups, we would have to pick mixed-design analysis of ANOVA to be followed as the treatment design method.

Question 2. Selection of the sampling or experimental design and number of replicates.

The data is formed by sampling men aged between 50-69 years. The experimental design then sub-divides them into four age categories and classifies based on smoking habits. The dataset is an aggregated case for each unique combination of influencing factor. We don’t have replicates.

Question 3. Selection of measurements to be taken.

The measurement will be death rate which will be calculated by dividing total number of deaths who smokes by total number of deaths under that age category.

Question 4. Selection of unit of observation(s).

The dependent variable in our case is “Number of cases” so the unit is the number of people who survived or died in the study.

3. Data Collection Procedures and Methods

Question 1. Interviews, Surveys, Observations, Focus Group, etc.

The data set which we are using are the observations which have been tracked by a study which was conducted by E.C Hammond and D. Horn in the year 1954. The Focus group in this study was around 187,766 men and both the scientist tracked them for 2 years. The study detailed their smoking pattern that is if they smoked or not and tracked their life status, that is if they were dead or alive at the end of this study period. This study is very helpful for our analysis since it was a cohort study with data that relates death, age and smoking.

Question 2. Procedures: How do you plan on acquiring the data using any of the methods?

The data to be used in our analysis was available from the study that was conducted by two scientists who were researching on the relationship between humans smoking habit and reason of deaths caused due to smoking.

Data source: http://users.stat.ufl.edu/~winner/data/smkdth.dat

Attributes: http://users.stat.ufl.edu/~winner/data/smkdth.txt

Part 4

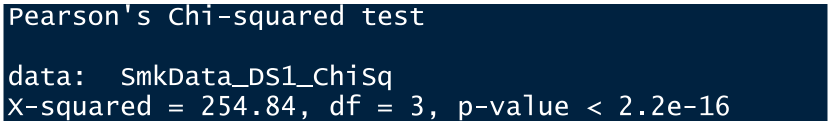
1. Elective Analyses

1. Chi-Square.

Starting with Chi Square test of independence, we would like to test the significance level of independence on the two groups of people who died but differ in their smoking habits with below hypothesis–

H_0 = The age classification and smoking habits are independent in determining the death rate

H_A = The age classification and smoking habits are related to each other in determining the death rate

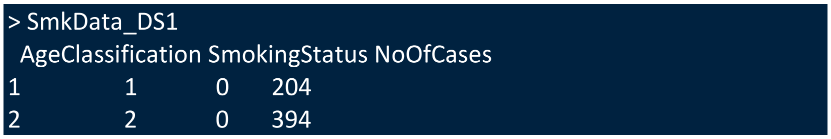


```
Pearson's Chi-squared test

data:  SmkData_DS1_ChiSq
X-squared = 254.84, df = 3, p-value < 2.2e-16
```

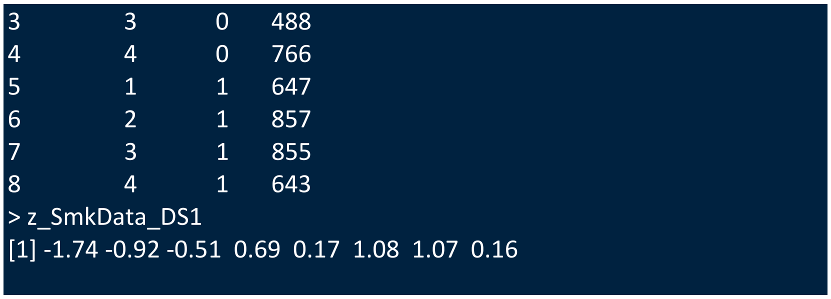
The p value of the test was found to be less than 2.2e-16 which is far below the significance level of 0.05 and thus, reject the null hypothesis and conclude that the two variables, i.e. age classification and smoking habit are in fact dependent on each other in determining the death rate.

2. Z-Score.

Next, we will calculate the z-score of the values to find out the deviation of a score from the mean. Below is the dataset and its z-score.



```
> SmkData_DS1
  AgeClassification SmokingStatus NoOfCases
1             1           0      204
2             2           0      394
```



```
3          3       0    488
4          4       0    766
5          1       1    647
6          2       1    857
7          3       1    855
8          4       1    643
> z_SmkData_DS1
[1] -1.74 -0.92 -0.51  0.69  0.17  1.08  1.07  0.16
```

The z-score gives us the following interpretations

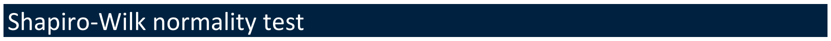
1. For people who smoked and lie in the first and youngest age group, 50-54 years have a death rate 0.17 standard deviations more than the average
2. For people who smoked and lie in the age group, 55-64 years have the highest death rate 1.08 and 1.07 standard deviations more than the average
3. For people who smoked and lie in the eldest age group, 65-69 years have the lowest death rate 0.16 standard deviations more than the average

3. Distributions (Normal, Non-normal).

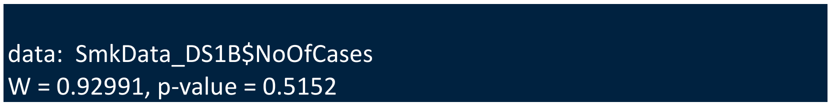
To test the normal distribution of the data, we will perform certain tests on our data –

1. shapiro-wilk test.

The null hypothesis for this test is that the data are normally distributed. For the chosen alpha level 0.05, if the p-value is less than 0.05, then the null hypothesis that the data are normally distributed is rejected. If the p-value is greater than 0.05, then the null hypothesis is not rejected.



```
Shapiro-Wilk normality test
```

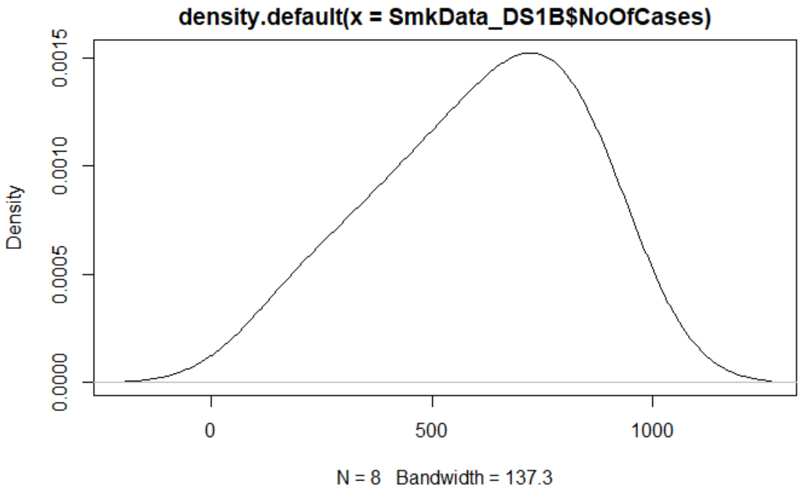


```
data:  SmkData_DS1B$NoOfCases
W = 0.92991, p-value = 0.5152
```

As the p-value is 0.5152 greater than the significance level, we do not reject the null hypothesis.

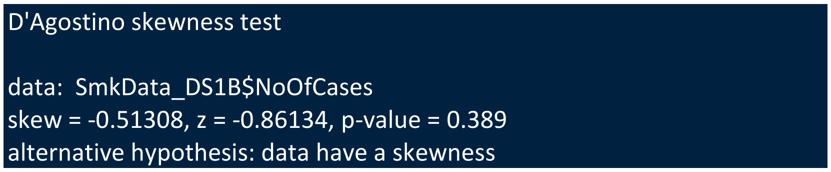
2. ***density plot.***

Plotting a density plot gives a normal distribution at eye level but we would have to perform the skew check.



3. ***skew test (Agostino's test).***

The null hypothesis for this test is that the data are normally distributed. For the chosen alpha level 0.05, if the p-value is less than 0.05, then the null hypothesis that the data are normally distributed is rejected. If the p-value is greater than 0.05, then the null hypothesis is not rejected.

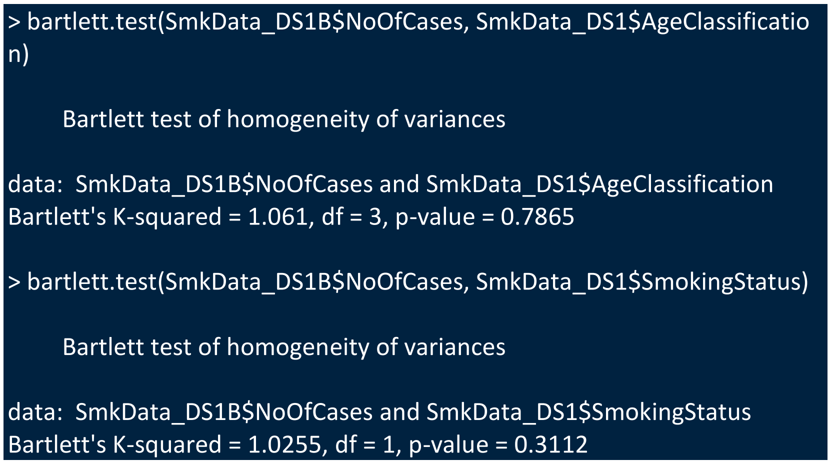


```
D'Agostino skewness test

data:  SmkData_DS1B$NoOfCases
skew = -0.51308, z = -0.86134, p-value = 0.389
alternative hypothesis: data have a skewness
```

As the p-value is 0.389 greater than the significance level, we do not reject the null hypothesis, and conclude that there is not significant amount of skew in the population.

4. ***bartlett's test of equal variances.***

The null hypotheses is that the variance is the same for all the death rates. The alternate hypothesis is that the variances are different for at least death rates.

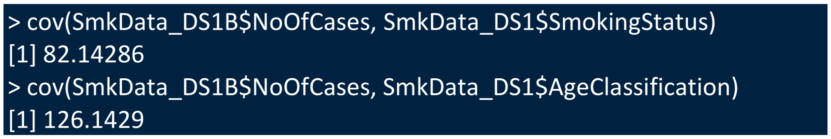


```
> bartlett.test(SmkData_DS1B$NoOfCases, SmkData_DS1$AgeClassificatio
n)

        Bartlett test of homogeneity of variances

data:  SmkData_DS1B$NoOfCases and SmkData_DS1$AgeClassification
Bartlett's K-squared = 1.061, df = 3, p-value = 0.7865

> bartlett.test(SmkData_DS1B$NoOfCases, SmkData_DS1$SmokingStatus)

        Bartlett test of homogeneity of variances

data:  SmkData_DS1B$NoOfCases and SmkData_DS1$SmokingStatus
Bartlett's K-squared = 1.0255, df = 1, p-value = 0.3112
```

The test with both the variables, Age classification and smoking status gives a p-value of 0.7865 and 0.3112 respectively, by which we can reject the null hypothesis and conclude that the variance of both the variables is homogenous throughout the population.

4. **Covariance.**

The covariance of Number of Cases and Age Classification, and Number of Cases and Smoking status in our data set measures how the two pairs are linearly related. A positive covariance would indicate a positive linear relationship between the variables, and a negative covariance would indicate the opposite.



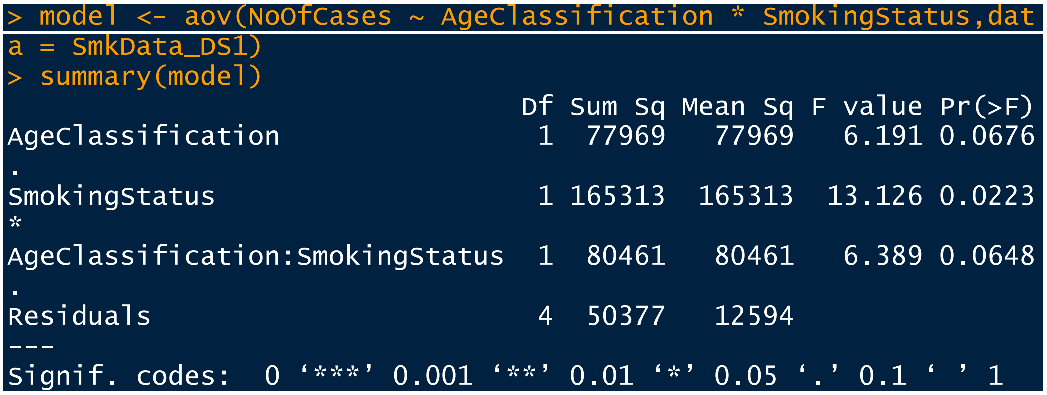
```
> cov(SmkData_DS1B$NoOfCases, SmkData_DS1$SmokingStatus)
[1] 82.14286
> cov(SmkData_DS1B$NoOfCases, SmkData_DS1$AgeClassification)
[1] 126.1429
```

The covariance of Number of Cases and Age Classification, and Number of Cases and Smoking status is about 82 and 126. It indicates a positive linear relationship between the two variables.

2. Mandatory Analyses

1. ANOVA (Analysis of Variance).

To know if the death rate depends on the age classification and smoking status, we performed a two-way ANOVA analysis.



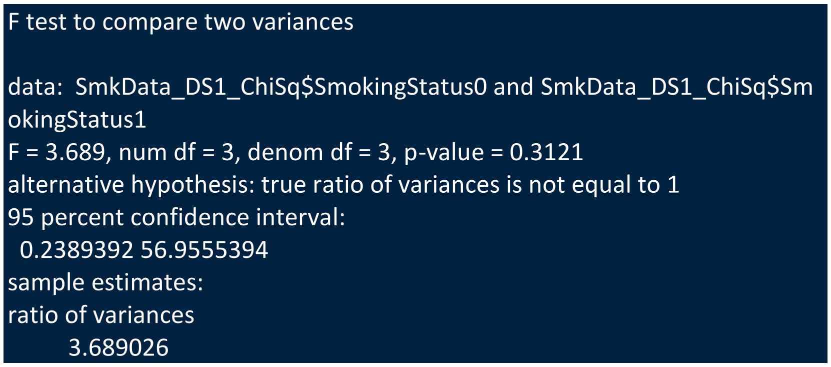
```
> model <- aov(NoOfCases ~ AgeClassification * SmokingStatus,dat
a = SmkData_DS1)
> summary(model)
                                Df Sum Sq Mean Sq F value Pr(>F)
AgeClassification                1  77969   77969   6.191 0.0676
.
SmokingStatus                    1 165313  165313  13.126 0.0223
*
AgeClassification:SmokingStatus  1  80461   80461   6.389 0.0648
.
Residuals                        4  50377   12594
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the ANOVA table, we can conclude that only smoking status is statistically significant among the two predictors. It concludes that age group doesn’t have much significant impact on death rate. These results would lead us to believe that changing smoking habits will impact significantly the survival of the person.

2. F-Test, Sum of Square Error, P-value, Means.

1. F-Test.

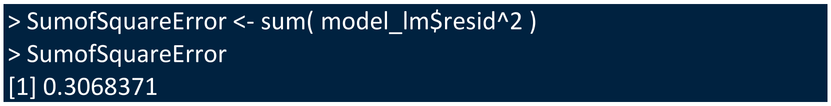
F-test between 2 normal populations with hypothesis that variances of the 2 populations are equal. We tested the variance of the death rates for two groups of who did not survive people who smoked and who did not.



```
F test to compare two variances

data:  SmkData_DS1_ChiSq$SmokingStatus0 and SmkData_DS1_ChiSq$Sm
okingStatus1
F = 3.689, num df = 3, denom df = 3, p-value = 0.3121
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.2389392 56.9555394
sample estimates:
ratio of variances
         3.689026
```

The p-value of F-test is p = 0.3121 which is greater than the significance level 0.05 and f= 3.689 is included in the confidence interval [0.2389, 56.95]. which means we cannot reject the null hypothesis and form the conclusion that there is no significant difference between the two variances.

2. sum of square error

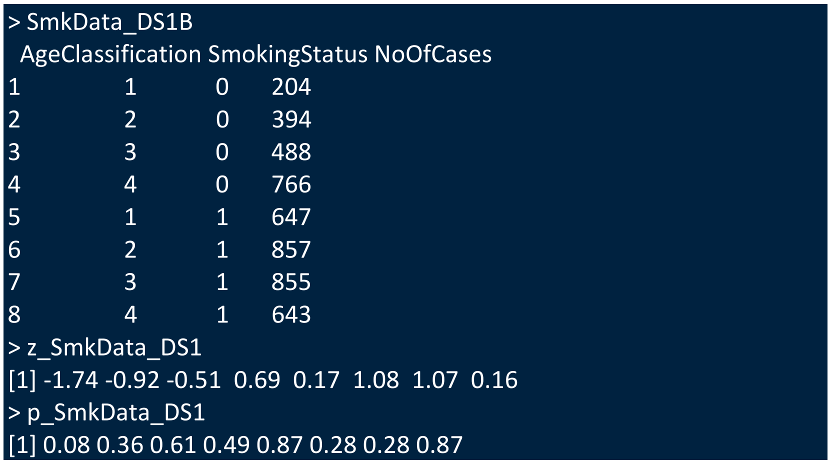


```
> SumofSquareError <- sum( model_lm$resid^2 )
> SumofSquareError
[1] 0.3068371
```

A value closer to 0 indicates that the model has a smaller random error component, and that the fit will be more useful for prediction. The sum of square error is calculated to be 0.30 from the linear model model_lm.

3. p-value.

Since we have obtained the z-scores in the elective part and now, we can calculate p-values, the probabilities against the null hypothesis of the z-scores.

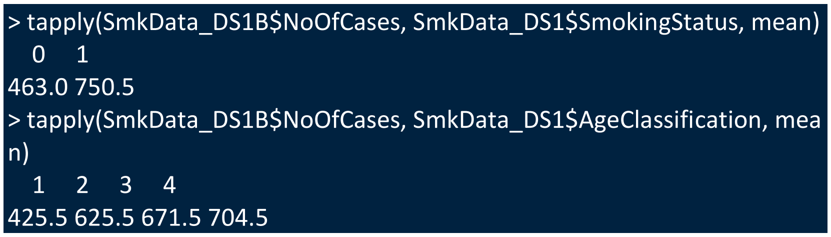


```
> SmkData_DS1B
  AgeClassification SmokingStatus NoOfCases
1                 1             0       204
2                 2             0       394
3                 3             0       488
4                 4             0       766
5                 1             1       647
6                 2             1       857
7                 3             1       855
8                 4             1       643
> z_SmkData_DS1
[1] -1.74 -0.92 -0.51  0.69  0.17  1.08  1.07  0.16
> p_SmkData_DS1
[1] 0.08 0.36 0.61 0.49 0.87 0.28 0.28 0.87
```

The p-values suggests that there is 87% probability of the death rate to be true for the people lying in age group 50-54-year-old and 55-69-year-old and smoking.

4. means.

By calculating the means of the independent variable, Number of cases by sub-dividing it as per the factors of smoking status and age classifications, we obtained the following interpretations.



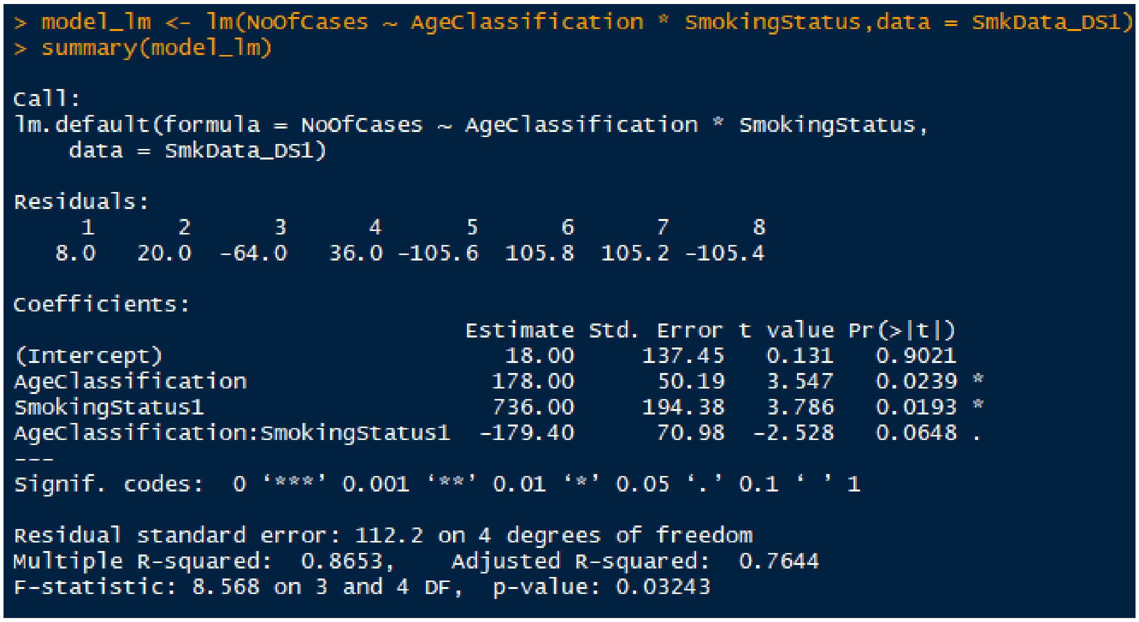
```
> tapply(SmkData_DS1B$NoOfCases, SmkData_DS1$SmokingStatus, mean)
    0     1
463.0 750.5
> tapply(SmkData_DS1B$NoOfCases, SmkData_DS1$AgeClassification, mea
n)
    1     2     3     4
425.5 625.5 671.5 704.5
```

1. With respect to smoking status – The death rate increases to 750.5 from 463.0 when we change the smoking status to yes from no which clearly suggests that non-smokers have higher chances of surviving than smokers.

2. With respect to age classification - There is a positive linear trend here which depicts that with increasing age, the death rate increases.

3. Regression Analysis (Prediction Models).

We ran a simple linear regression model in R and interpret the key components of the R linear model output.



```
> model_lm <- lm(NoOfCases ~ AgeClassification * SmokingStatus,data = SmkData_DS1)
> summary(model_lm)

Call:
lm.default(formula = NoOfCases ~ AgeClassification * SmokingStatus,
    data = SmkData_DS1)

Residuals:
     1      2      3      4      5      6      7      8 
   8.0   20.0  -64.0   36.0 -105.6  105.8  105.2 -105.4 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)  
(Intercept)                          18.00     137.45   0.131   0.9021  
AgeClassification                   178.00      50.19   3.547   0.0239 *
SmokingStatus1                      736.00     194.38   3.786   0.0193 *
AgeClassification:SmokingStatus1  -179.40      70.98  -2.528   0.0648 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 112.2 on 4 degrees of freedom
Multiple R-squared:  0.8653,	Adjusted R-squared:  0.7644 
F-statistic: 8.568 on 3 and 4 DF,  p-value: 0.03243
```

Formula Call: Analyze the interaction and individual effects of Age Classification and Smoking Status on the death rate

Residuals: Residuals are essentially the difference between the actual observed response values (death rate in our case) and the response values that the model predicted. We can see that the distribution of the residuals does not appear to be strongly symmetrical. That means that the model predicts certain points that fall far away from the actual observed points.

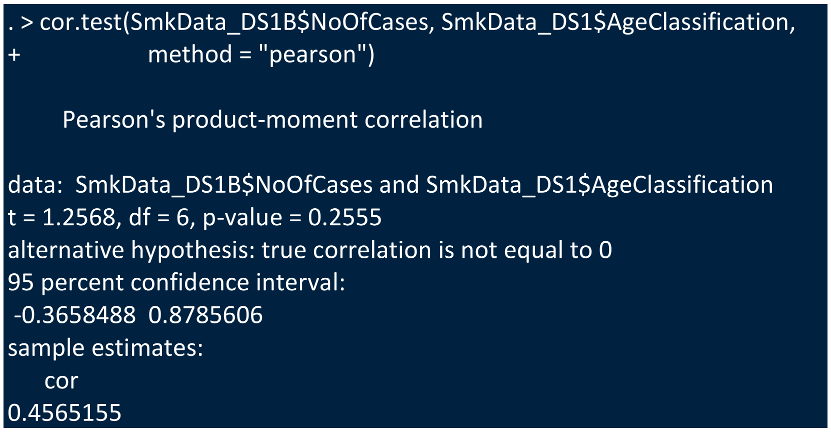
Coefficient Estimate: The coefficient Estimate contains four rows; the first one is the intercept. The intercept, in our example, is essentially the expected value of the death rate when we consider the average age and smoking status in the dataset. So, in an average condition, the death rate is 18. The next rows determine the slope terms which are saying that for every next age group, the death rate goes up by 178, for the smoking status to be true, the death rate jumps to a huge value of 736 which means there is a high chances of the death.

p-value: The p-value for Age classification is significant and less than the alpha which is 0.05 which observes a relationship between age classification and death rates. Similarly, smoking status (1=yes) is also highly significant and indicates that we can reject the null hypothesis which allows us to conclude that there is a relationship between death rate and smoking status (1) as well. The interaction of smoking status and age classification is also borderline significant which suggests there may be an effect of smoking states with specific age groups.

Multiple R^2/Adjusted R^2: The R-squared statistic provides a measure of how well the model is fitting the actual data. We get an R^2 in model as 0.8653. Or roughly 86% of the variance found in the response variable (death rate) can be explained by the predictor variable (smoking status and age classification).

4. Coefficient.

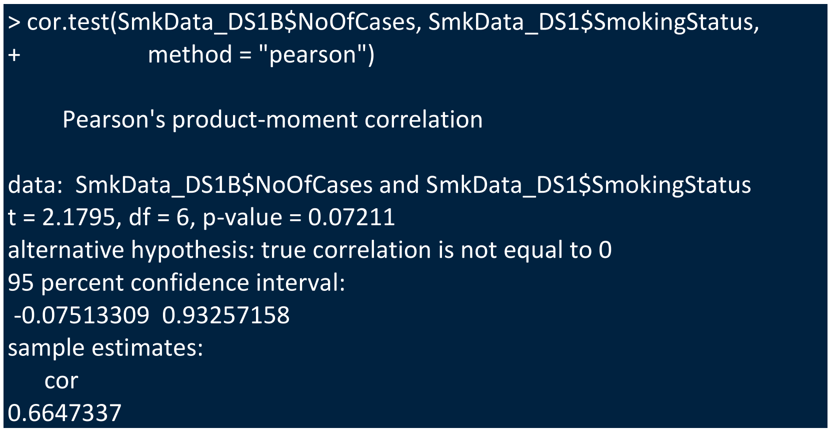
We performed the Pearson’s Correlation Test to evaluate the association between number of cases and age classification.



```
. > cor.test(SmkData_DS1B$NoOfCases, SmkData_DS1$AgeClassification,
+                  method = "pearson")

        Pearson's product-moment correlation

data:  SmkData_DS1B$NoOfCases and SmkData_DS1$AgeClassification
t = 1.2568, df = 6, p-value = 0.2555
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3658488  0.8785606
sample estimates:
      cor
0.4565155
```

The p-value (0.2555) is the significance level of the t-test (t-test statistic value = 1.2568). As the p-value is greater than the significant value (alpha is 0.05), we cannot reject the null hypothesis and conclude that there is no correlation between the two variables.

Next, we performed the Pearson’s Correlation Test to evaluate the association between number of cases and smoking status.



```
> cor.test(SmkData_DS1B$NoOfCases, SmkData_DS1$SmokingStatus,
+                  method = "pearson")

        Pearson's product-moment correlation

data:  SmkData_DS1B$NoOfCases and SmkData_DS1$SmokingStatus
t = 2.1795, df = 6, p-value = 0.07211
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.07513309  0.93257158
sample estimates:
      cor
0.6647337
```

The p-value (0.07211) is the significance level of the t-test (t-test statistic value = 2.1795). As the p-value is greater than the significant value (alpha is 0.05), we cannot reject the null hypothesis and conclude that there is no correlation between the two variables.

Part 5

1. **Bias Control Strategy**

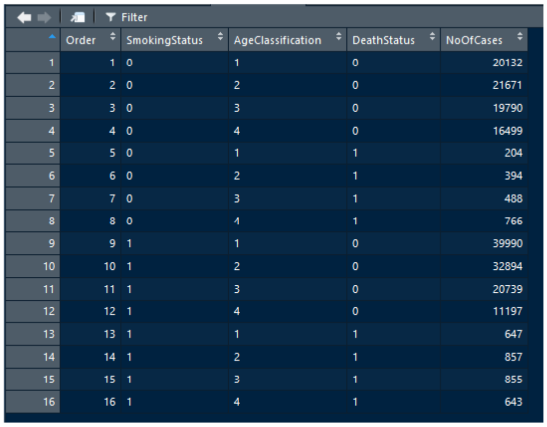
   As we did not perform sampling and took the aggregated results of the collected sample from our data provider. From the aggregated data, we had consistent and identifiable samples (i.e. each sample represent a category).

2. **Data Analysis and discussion of findings: Include data tables, graphs, R-output/findings images, formulas, etc.**

   1. **Data Load**

   The original dataset SmkData is broken down into 2 datasets initially, separated by death status = 0 and 1 to focus our attention to the data having death status = 1

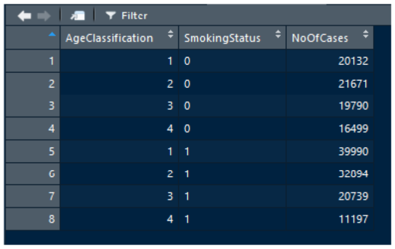
Dataset = SmkData (Original)



	Order	SmokingStatus	AgeClassification	DeathStatus	NoOfCases
1	1	0	1	0	20132
2	2	0	2	0	21671
3	3	0	3	0	19790
4	4	0	4	0	16499
5	5	0	1	1	204
6	6	0	2	1	394
7	7	0	3	1	488
8	8	0	4	1	766
9	9	1	1	0	39990
10	10	1	2	0	32894
11	11	1	3	0	20739
12	12	1	4	0	11197
13	13	1	1	1	647
14	14	1	2	1	857
15	15	1	3	1	855
16	16	1	4	1	643

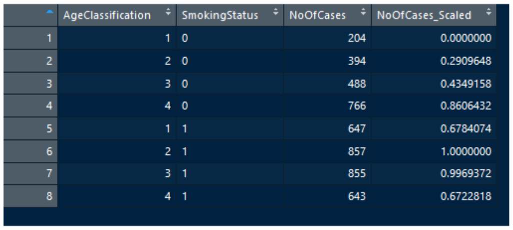
Original dataset flattened into two different datasets

Dataset = SmkData_DS0, DeathStatus = 0



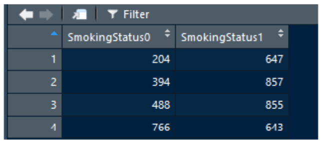
	AgeClassification	SmokingStatus	NoOfCases
1	1	0	20132
2	2	0	21671
3	3	0	19790
4	4	0	16499
5	1	1	39990
6	2	1	32094
7	3	1	20739
8	4	1	11197

Dataset = SmkData_DS1, DeathStatus = 1



	AgeClassification	SmokingStatus	NoOfCases	NoOfCases_Scaled
1	1	0	204	0.0000000
2	2	0	394	0.2909648
3	3	0	488	0.4349158
4	4	0	766	0.8606432
5	1	1	647	0.6784074
6	2	1	857	1.0000000
7	3	1	855	0.9969372
8	4	1	643	0.6722818

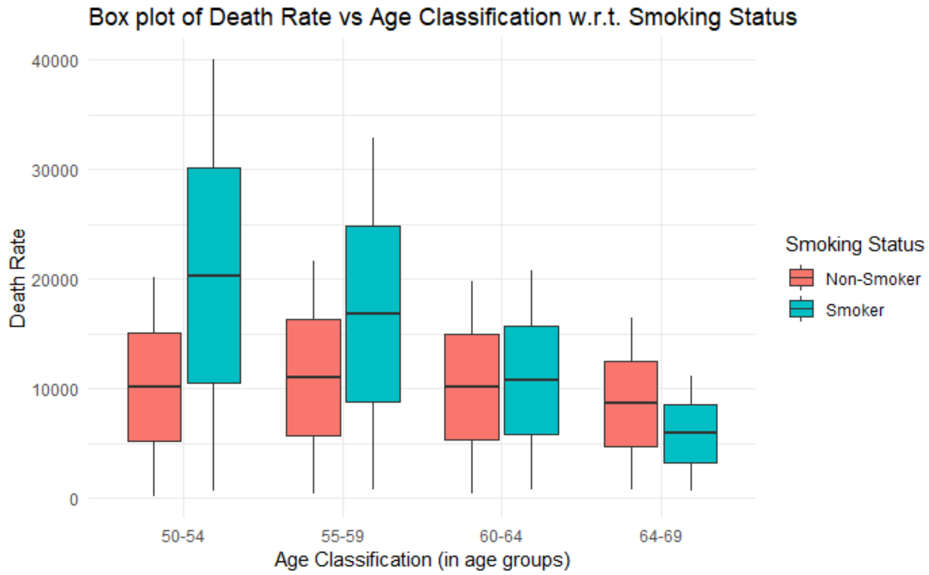
Dataset = SmkData_DS1_ChiSq,
Death Status = 1
(For chi square test)

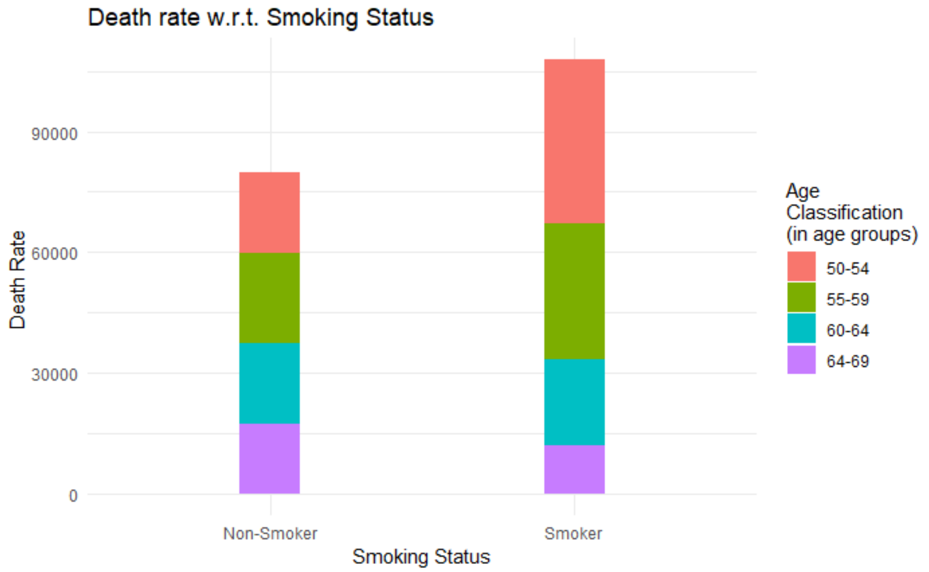


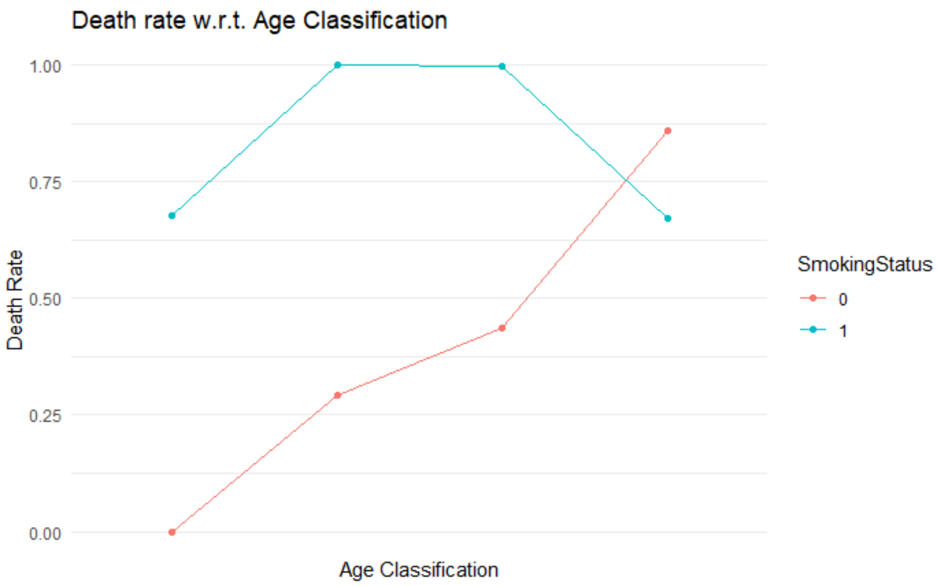
	SmokingStatus0	SmokingStatus1
1	204	647
2	394	857
3	488	855
4	766	643

2. Data Exploration

We want to observe the death rate of smoking people with respect to age. From the plot we observed that the death rate increases with increase in age and then drops, this might be since there are less numbers in the old age category. To analyze further, we plotted the death counts of non-smokers and compared it with the smokers where we observed that there is steep difference in the middle-aged category, the middle-aged people are more affected by smoking than other age groups which result in high death counts comparing to other age groups.





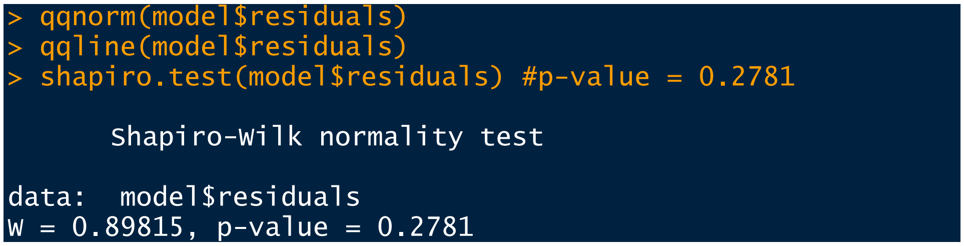


3. ANOVA

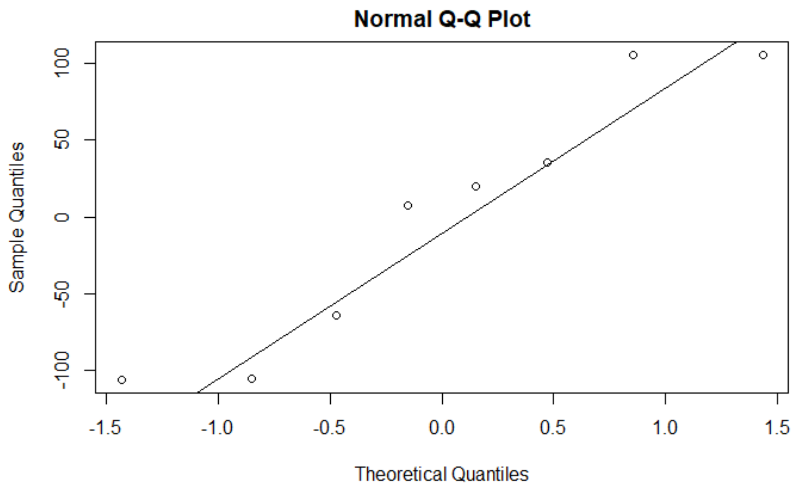
Before running a statistical test, we should run the four assumptions of normality to make sure if the data roughly fits a bell curve. We have performed all the tests to check on the four assumptions of normality in the Part 4, section I, sub-section 3 (Distributions -Normal, Non-normal)

Next is to perform the ANOVA test for which the details and interpretations can be referred in Part 4, section II, sub-section 1 (ANOVA)

After performing the statistical analysis, we looked at the residuals to check if they follow a normal distribution.



```
> qqnorm(model$residuals)
> qqline(model$residuals)
> shapiro.test(model$residuals) #p-value = 0.2781

	Shapiro-Wilk normality test

data:  model$residuals
W = 0.89815, p-value = 0.2781
```



The conclusion above, is supported by the Shapiro-Wilk test on the ANOVA residuals (W = 0.89, p = 0.27) which finds no indication that normality is violated.

Part 6

1. Summary

Smoking has a direct and increasing impact on the death rate among all age groups. Our findings show that smoking habit influence death rate more among middle aged group of men. There should be more awareness programs targeted for middle aged group and regular heath checkup must be prescribed and encouraged among middle aged groups for early detection of health issues.

References

https://en.wikipedia.org/wiki/Bias_(statistics)

http://www.epidemiology.ch/history/PDF%20bg/Hammond%20EC%20and%20Horn%20D%201958%20smoking%20and%20death%20rates%20-%20report%20on%2044%20months%20of%20follow-up.pdf

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1551824/pdf/amjphnation01081-0025.pdf

https://www.cancer.org/cancer/cancer-causes/tobacco-and-cancer/health-risks-of-smoking-tobacco.html

https://www.cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/index.htm