ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

IE266 Case Study

Group 19

“Academic integrity is expected of all students of METU at all times, whether in the presence or absence of members of the faculty.

Understanding this, I declare that I shall not give, use, or receive unauthorized aid in this study.”

1)	**Islam Valehli**	**2349371**
2)	**Kadir Enez Demir**	**2396604**
3)	**Youssef Nsouli**	**2487494**
4)	**Onur Mete Keskin**	**2306454**

Table of Contents

1. Introduction

In this study, a researcher wants to test the effects of various factors related to smoking, which is responsible for most preventable deaths in the world. In addition, the researcher will statistically analyze how well the measures taken to prevent smoking are working. They will use survey data to better analyze the situation.

The survey includes information that may be useful: how long have the people smoked before the age of 18, how family members and close friends smoke, whether they were warned about the harms of smoking and whether they have seen anti-smoking messages in various media. In addition to the questionnaire, the lung capacity of the participants was also tested in order to perform quantitative analysis.

The first task to-be-done is testing the relationship between smoking duration and gender. This is by using contingency tables and performing a chi-square test. Secondly, the researcher will analyze the relationship between smokers in the family and smokers in the friend group using the same method.

The next task he wants to accomplish is construct the linear regression model for his research. With this model, he will explain the effects of various variables on lung capacity and the relationships between each other and then plot these effects.

The third task is constructing confidence interval on the mean response for respondents under certain conditions.

Finally, his final objective is to make predictions about the lung capacity in a certain situation using the constructed model. To do this he will again use confidence intervals. These estimates, analyses, and predictions can help people understand the causes of smoking, how well the measures taken have worked, and how anti-smoking policies need to be developed in the future.

2. Main Body of the Report

Question-1) Test the following claims, using contingency tables and chi-square tests...

a) Duration of smoking and gender are independent.

The required task is proving the independence of the duration of smoking before hitting 18 years old and gender independent of one another. Contingency tables and χ^2 tests will be used.

First, the expected values should be calculated. The probability of two independent events occurring is simply the multiplication of each events probability, namely:

$$P(A \cap B) = P(A)P(B)$$

To find the expected value of a person of a certain gender smoking for a certain period is simply a multiplication of probabilities.

The data is collected and organized into contingency tables to ease calculation. Gender is already a categorized term, and the duration of smoking was categorized into 5 classes, where class 1 is for people who have smoked for less than a year, class 2 for one to two years of smoking, etc., and the last class is for four or more years of smoking.

For observations, the categories in which each participant falls into is counted (Table 1). As for expected values, a more sophisticated algorithm is employed, with the results in Table 2.

Table 1

Observed and Expected Number of Observations, Respectively

	Male	Female
0-1'	252	263
1-2'	297	300
2-3'	196	216
3-4'	88	78
4 and more	82	81

	Male	Female
0-1'	254.3	260.69
1-2'	294.79	302.2
2-3'	203.44	208.56
3-4'	81.97	84.03
4 and more	80.49	82.51

The χ_0^2 is calculated, by using R-Script, as follows:

$$\chi_0^2 = \sum_{i=1}^{n} \frac{(E_i - O_i)^2}{E_i^2}$$

Hypotheses:

H_0: The duration of smoking before 18 years old and gender of a person are independent.

H_a: The two statistics are dependent.

H_0 is rejected if $\chi_0^2 > \chi_{\alpha,(m-1)(n-1)}^2$, where m and n are the number of categories for duration of smoking before 18 and gender respectively.

For this task, $\chi_0^2 = 0.014$ and $\chi_{\alpha,(m-1)(n-1)}^2 = \chi_{0.05,(5-1)(2-1)}^2 = \chi_{0.05,4}^2 = 0.71$.

And since $\chi_0^2 \ngtr \chi_{\alpha,4}^2$, the researcher failed to reject H_0, meaning there isn't evidence, under 0.05 significance, to conclude that the parameters in study are dependent on one another. In addition, the p value is $p = 2.484 \times 10^{-5}$, which is way less than α

b) The number of smokers in the family and the number of smokers in the frie…

The same idea as the previous task is used. The variables (Smoking family members and smoking friends before reaching the age of 18) are already categorized into "Zero", "One", "Two", and "More than 3."

Data points will be counted, and the expected values will be counted using the same algorithm used in the first part. (Tables 3 and 4)

Table 2

Observed and Expectation Number of Observations, Respectively

	Zero Family Members	One Family Members	Two Family Members	More than 3 Family Members
Zero Friends	147	201	277	103
One Friend	63	96	137	69
Two Friends	71	129	144	49
More than 3 Friends	78	100	133	56

	Zero Family Members	One Family Members	Two Family Members	More than 3 Family Members
Zero Friends	141.04	206.65	271.47	108.82
One Friend	70.75	103.61	136.11	54.56
Two Friends	76.14	111.56	146.56	58.75
More than 3 Friends	71.1	104.18	136.86	54.86

Hypotheses:

H_0: The number of smoking family members and friends the person has had before 18 are independent of one another.

H_a: The two variables are dependent.

The χ_0^2 value, as evaluated by R, is $\chi_0^2 = 0.1622$. H_0 is rejected if this value exceeds $\chi^2_{0.05,(4-1)(4-1)} = \chi^2_{0.05,9} = 3.325$.

The test statistic doesn't exceed that number, so, under 95% confidence, there is not enough evidence to conclude that the number of family members and friends who smoke the person had before 18 years old are dependent on one another. In addition, the p value of χ_0^2 is $p = 2.2 \times 10^{-7} \ll \alpha$

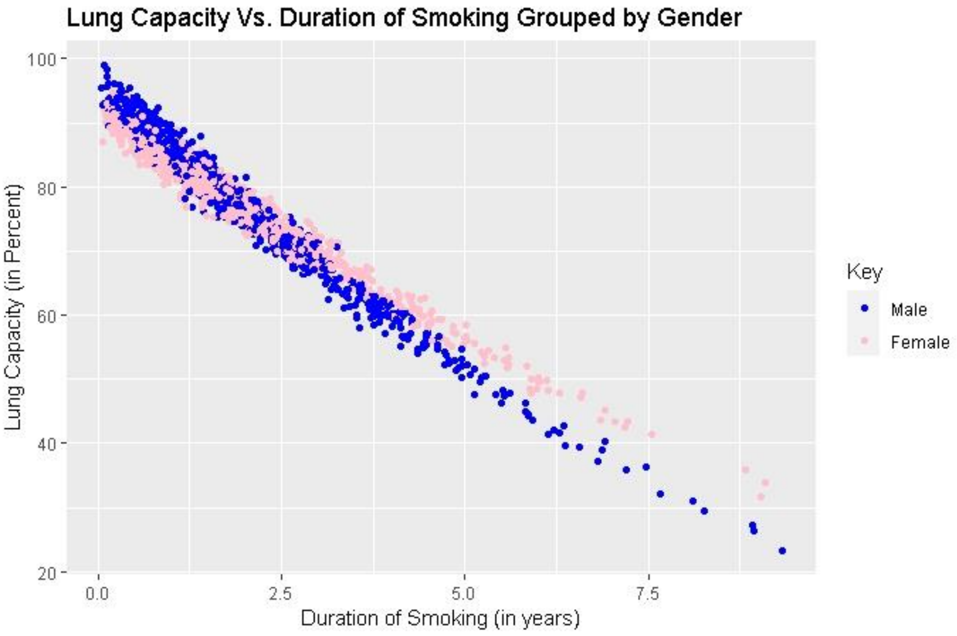
Question-2) Construct the linear regression model using all variables. Then perform…

a) Plot the response against the continuous variables with groups of signific…

To start with the linear regression model, it is imperative to determine what variables to use and add. So, plots will be made by the researcher, who will plot the response against the continuous variables of the data, grouping the data points by the categorical variables. This will ensure that the researcher will be able to find any pattern regarding the interactions between any of the two variables:

Figure 3

Plot of Lung Capacity Vs. Duration of Smoking Before Reaching 18 years of age.

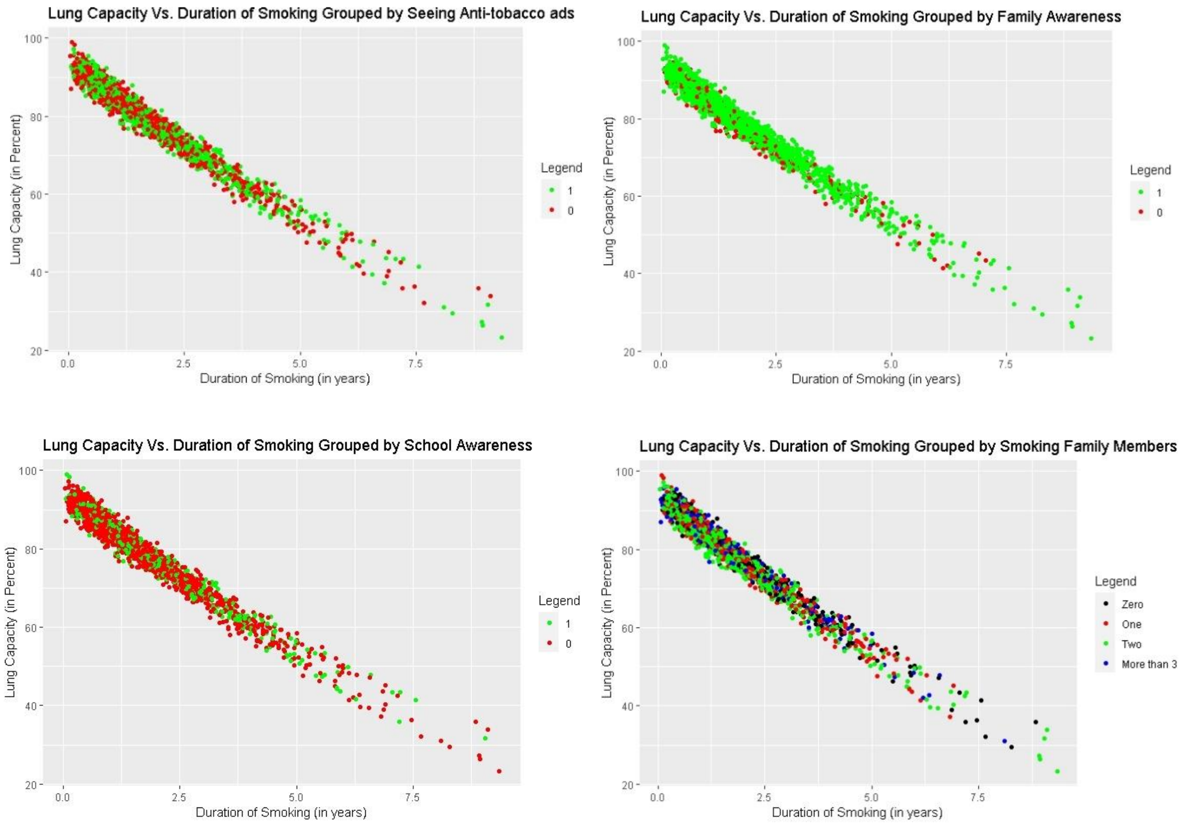


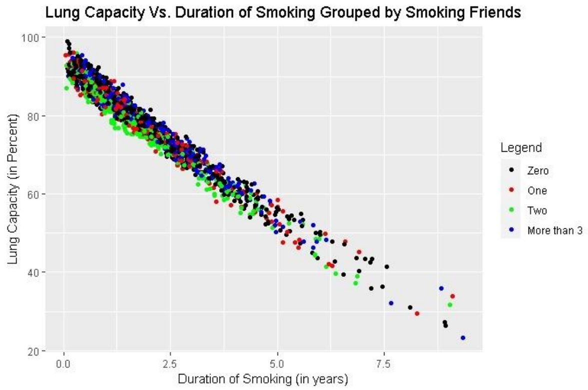
As it can be seen, there is a distinct correlation between gender and duration of smoking, indicating that there is an interaction between these two terms and having a combined effect on lung capacity.

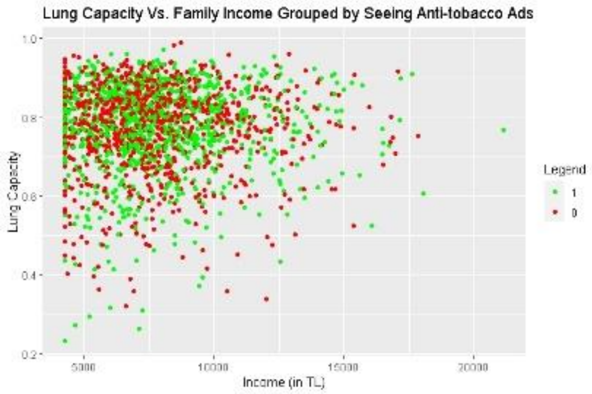
As for other plots, no clear interaction can be seen. Also, the relation between the response and duration of smoking is linear; no quadratic or logarithmic terms need be added. The relation between lung capacity and income is fuzzier; it is not clear if there is any relation at all, but the income's respective variable will be added anyways.

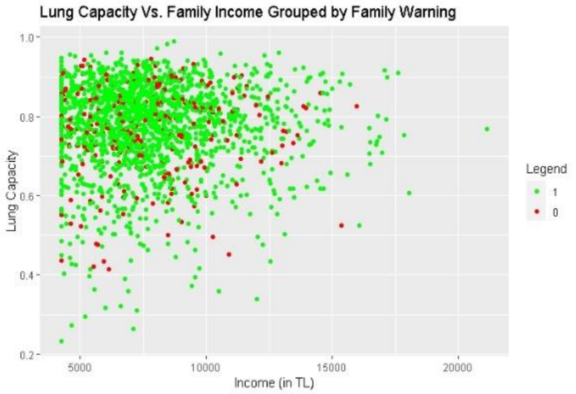
Figures 4 through 14

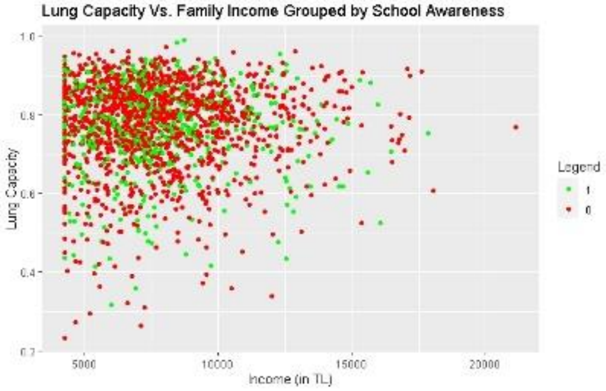
Various Plots of Lung Capacity Vs. Either Duration of Smoking or Family Income Before 18 years of age, Grouped by Various Categorical Variables.

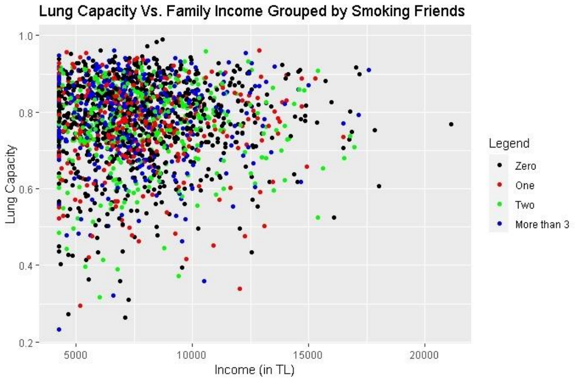


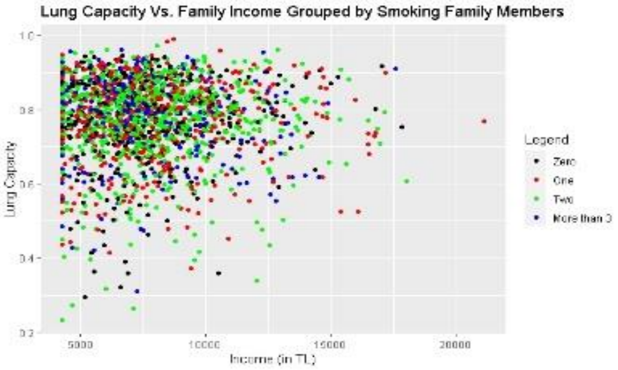

Lung Capacity Vs. Duration of Smoking Grouped by Smoking Friends

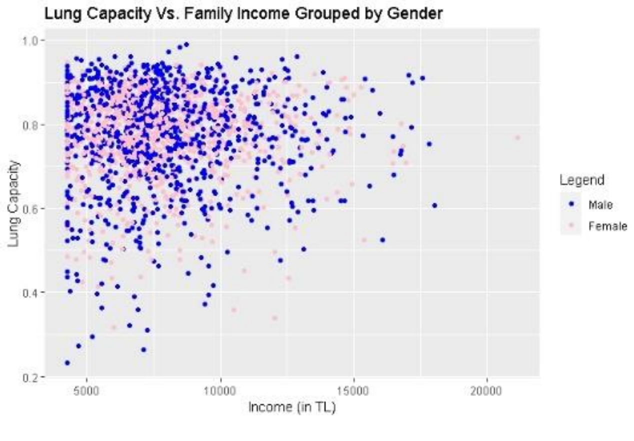

Lung Capacity Vs. Family Income Grouped by Seeing Anti-tobacco Ads


Lung Capacity Vs. Family Income Grouped by Family Warning


Lung Capacity Vs. Family Income Grouped by School Awareness


Lung Capacity Vs. Family Income Grouped by Smoking Friends


Lung Capacity Vs. Family Income Grouped by Smoking Family Members

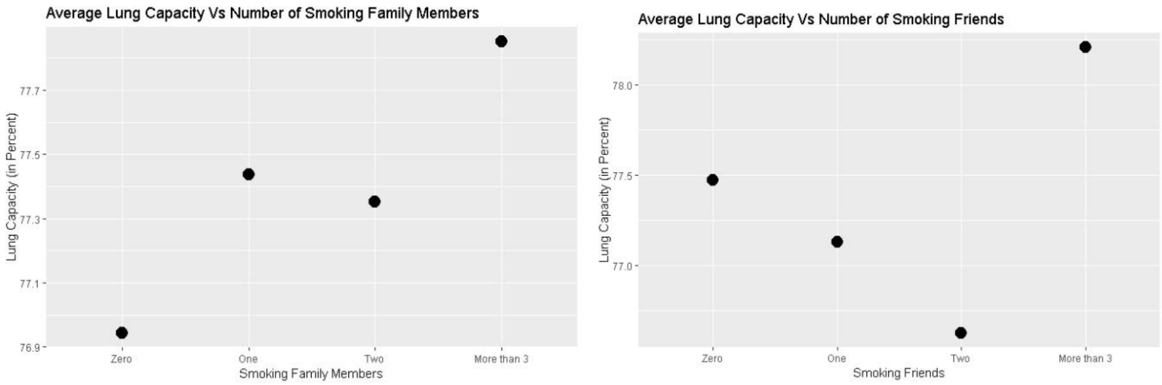

Lung Capacity Vs. Family Income Grouped by Gender

In addition to checking for any interactions between variables, the researcher will numerate the categorical variables, so that the R-Script software can deal with it. For example, the gender parameter will be evaluated as 1 for Male, and 0 for Female. The other three binary parameters, “Had anyone in your family warned you about the hazards of smoking?”, “Have you ever been informed about the hazards of smoking in any of your courses in high school?”, and “Have you seen anti-tobacco messages on TV/on billboards/in newspapers?”, are already provided in 0-1 format and do not need any more alteration.

However, the “When you were in high school, how many of your family members smoked?” and “When you were in high school, how many of your friends in your close friends group smoked?” will be converted to 4 variables each, one for each condition of “Zero”, “One”, “Two”, or “More than 3”. This is to ensure to not assume any linearity of the effect of family or friends who smoke against the response variable. As it can be seen from the following figures, the effect doesn’t seem to follow a linear pattern:

Figures 15 and 16

Average Lung Capacity Vs. Number of Smoking Family Members (Figure 13) and Number of Smoking Friends (Figure 14) a Person had Before 18.



b) Construct the linear regression model using the stepwise regression methods (back…

Since the researcher has prepared the variables that need to be added and all the interaction terms, they are now ready for constructing the linear regression model. Using the stepwise regression model, each variable will be added on its own to the model (given it is not in the model yet), and the F-value and p-value will be tested for each. α_{in} will be chosen as 0.15, and α_{out} will also be 0.15. The variables are arbitrarily named as $x_1, x_2, \ldots x_{15}$. As such, the regression model and regression equation look like this:

$$Y_i = \beta_0 + \beta x_1 + \cdots + \beta_{15} x_{15} + \epsilon_i$$
$$E[Y_i] = \beta_0 + \beta x_1 + \cdots + \beta_{15} x_{15}$$

ϵ is the error variable. $\epsilon_i \sim N(0, \sigma^2)$

At each step, the candidate entering variables will not be chosen arbitrarily, but the one with the highest F-value will be chosen. The steps are as follow:

Step 1:

- Entering Variable: x_1 (Duration of Smoking)
- p-value: 0.000
- F-value: 46131
- R^2_{adj}: 96.14%

Step 2:

- Entering Variable: x_{10} (Two Smoking Friends)
- p-value: 0.000
- F-value: 197.88
- R^2_{adj}: 96.5%

Step 3:

- Entering Variable: x_{15} (Interaction between gender and Duration of Smoking)
- p-value: 0.000
- F-value: 132
- R^2_{adj}: 96.7%

Step 4:

- Entering Variable: x_3 (Gender)
- p-value: 0.000
- F-value: 197.88

- R_{adj}^2: 97.4%

Step 5:

- Entering Variable: x_{12} (Whether the person was warned about smoking by their family or not)
- p-value: 0.000
- F-value: 107.93
- R_{adj}^2: 97.5%

Step 6:

- Entering Variable: x_9 (One Smoking Friend)
- p-value: 0.000
- F-value: 66.64
- R_{adj}^2: 97.6%

Step 7:

- Entering Variable: x_6 (Two Smoking Family Members)
- p-value: 0.000
- F-value: 51.96
- R_{adj}^2: 97.69%

Step 8:

- Entering Variable: x_5 (One Smoking Family Member)
- p-value: 0.008
- F-value: 6.9941
- R_{adj}^2: 97.7%

At all steps, no variable already present in the model had a p-value more than $\alpha_{out} = 0.15$. The model is thus:

$$\hat{y} = 0.909 - 0.07x_1 + 0.03x_3 - 0.002x_5 - 0.007x_6 - 0.008x_9 - 0.019x_{10} + 0.012x_{12} - 0.01x_{15}$$

The summary of the final model and the ANOVA table are provided in Output 15 and 16:

Output box 17

Summary of the Model as Provided by R-Script

```
Residuals:
      Min        1Q    Median        3Q       Max 
-0.053095 -0.011666 -0.000745  0.010544  0.093518 

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)    
(Intercept)  0.9090476  0.0015680  579.736  < 2e-16 ***
x1          -0.0708266  0.0004005 -176.857  < 2e-16 ***
x10         -0.0190077  0.0010022  -18.966  < 2e-16 ***
x15         -0.0151800  0.0005626  -26.983  < 2e-16 ***
x3           0.0323457  0.0013544   23.882  < 2e-16 ***
x12          0.0123617  0.0011610   10.648  < 2e-16 ***
x9          -0.0085560  0.0010285   -8.319  < 2e-16 ***
x6          -0.0070940  0.0009359   -7.580 5.44e-14 ***
x5          -0.0026601  0.0010058   -2.645  0.00825 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01701 on 1844 degrees of freedom
Multiple R-squared:  0.9771,  Adjusted R-squared:  0.977 
F-statistic:  9853 on 8 and 1844 DF,  p-value: < 2.2e-16
```

Output box 18

ANOVA Table of the Model as Provided by R-Script

```
Response: y
            Df  Sum Sq Mean Sq    F value    Pr(>F)    
x1           1 22.4477 22.4477 77558.2557 < 2.2e-16 ***
x10          1  0.0870  0.0870   300.7027 < 2.2e-16 ***
x15          1  0.0543  0.0543   187.4648 < 2.2e-16 ***
x3           1  0.1554  0.1554   537.0654 < 2.2e-16 ***
x12          1  0.0333  0.0333   115.2103 < 2.2e-16 ***
x9           1  0.0199  0.0199    68.6922 < 2.2e-16 ***
x6           1  0.0150  0.0150    51.8634 8.629e-13 ***
x5           1  0.0020  0.0020     6.9941  0.008247 ** 
Residuals 1844  0.5337  0.0003                         
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

c) Check whether the transformation of the response variable is necessary

After completing the mode, it is required to check the validity of the model. This can be done by validating the assumptions made on the error term in the regression mode:

$$Y_i = \beta_0 + \beta x_1 + \cdots + \beta_{15} x_{15} + \epsilon_i$$

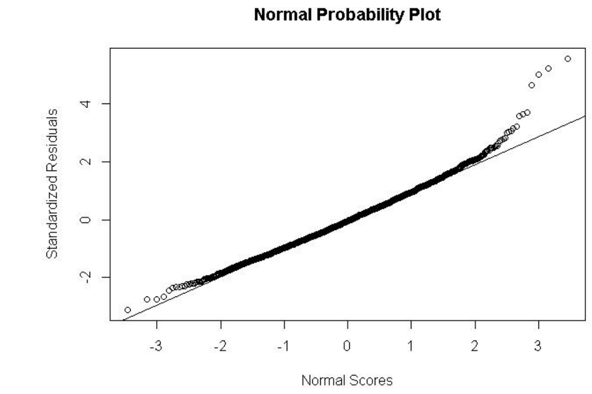
The assumptions are:

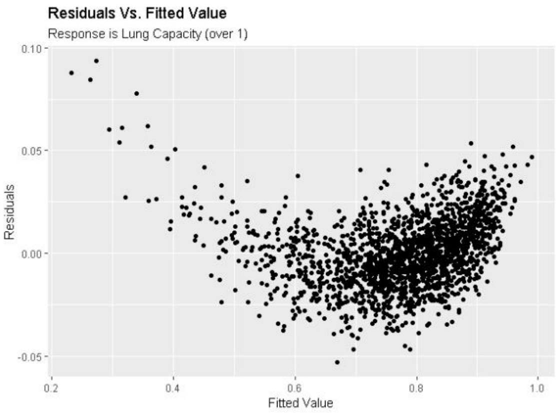
- The error variable is normally distributed
- The mean of the error variable is 0
- The variance does not vary
- There is no correlation between each ϵ_i; they are iid

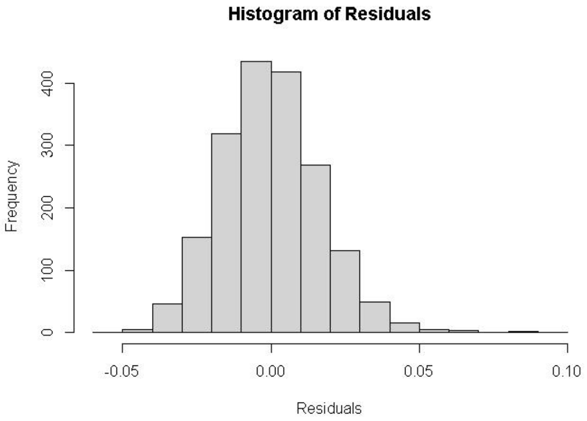
To test these assumptions, various plots with residuals are made. (Figure 16)

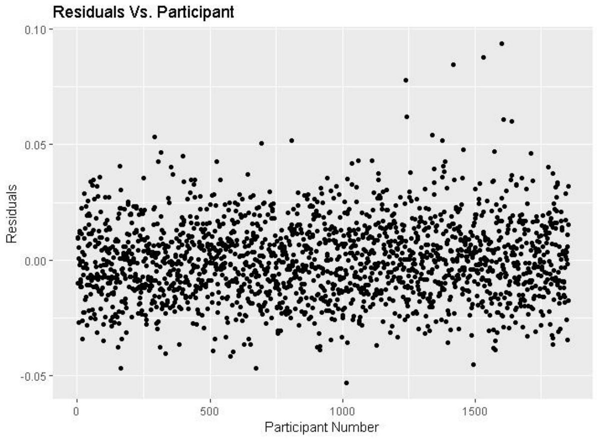
Figure 19

Various Plots Involving Residuals: Normal Probability Plot, Residuals Vs Fitted Value, Histogram, and Residuals Vs Participant









Based on these graphs, it is safe to say that the error variable is normally distributed and each error is independent of the next. Also, the expected value seems to be 0. However, the Residuals Vs. Fitted Value graph is peculiar, as there seems to a quadratic pattern. This indicates that a transformation of the response variable is needed.

So, the model is redesigned by getting the current variable and taking its square root, and adding it to the model. So, if the response variable in the initial model is y, the new model is:

$$\sqrt{y} = \beta_0 + \beta x_1 + \cdots + \beta_{15} x_{15} + \epsilon_i$$

The new model is obtained, of equation:

$$\sqrt{\hat{y}} = 0.9586 - 0.042x_1 + 0.02x_3 - 0.001x_5 - 0.004x_6 - 0.0046x_9 - 0.01x_{10} + 0.007x_{12} - 0.01x_{15}$$

A summary and ANOVA table are also obtained:

Textbox 20

Summary of the Model After Transformation

```
Residuals:
       Min         1Q     Median         3Q        Max 
-0.0283617 -0.0057241 -0.0001535  0.0056790  0.0303553 

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)    
(Intercept)  0.9586331  0.0007848 1221.572  < 2e-16 ***
x1          -0.0424059  0.0002004 -211.581  < 2e-16 ***
x10         -0.0107145  0.0005016  -21.361  < 2e-16 ***
x15         -0.0100585  0.0002816  -35.725  < 2e-16 ***
x3           0.0200761  0.0006778   29.618  < 2e-16 ***
x12          0.0070004  0.0005810   12.048  < 2e-16 ***
x9          -0.0046529  0.0005148   -9.039  < 2e-16 ***
x6          -0.0041981  0.0004684   -8.963  < 2e-16 ***
x5          -0.0013089  0.0005034   -2.600  0.00939 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.008514 on 1844 degrees of freedom
Multiple R-squared:  0.9843,	Adjusted R-squared:  0.9842 
F-statistic: 1.442e+04 on 8 and 1844 DF,  p-value: < 2.2e-16
```

Textbox 21

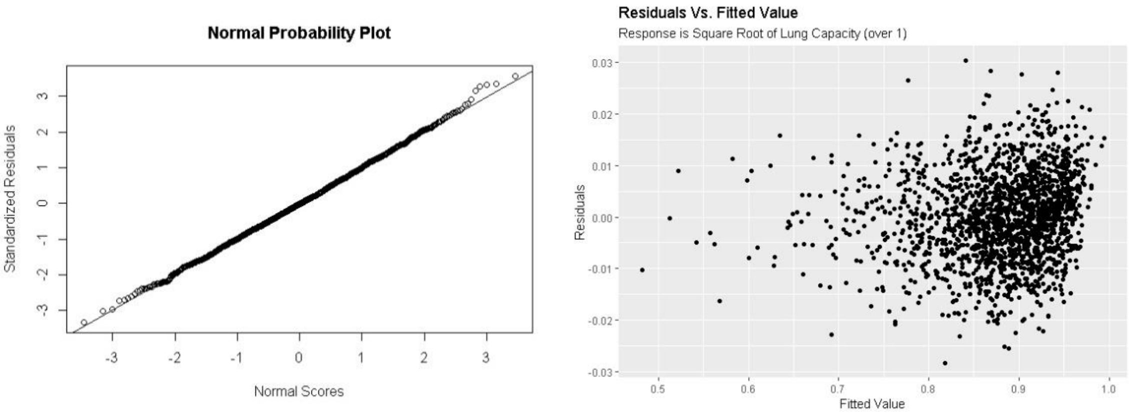
ANOVA Table of the Model After Transformation

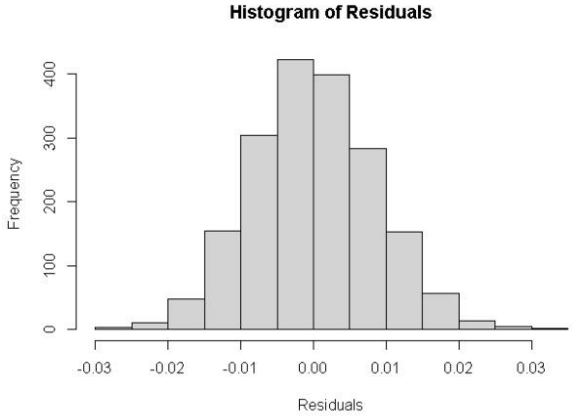
```
Response: sqrt_y
            Df Sum Sq Mean Sq    F value    Pr(>F)    
x1           1 8.2201  8.2201 113391.271 < 2.2e-16 ***
x10          1 0.0277  0.0277    382.604 < 2.2e-16 ***
x15          1 0.0313  0.0313    432.325 < 2.2e-16 ***
x3           1 0.0602  0.0602    830.160 < 2.2e-16 ***
x12          1 0.0106  0.0106    146.710 < 2.2e-16 ***
x9           1 0.0059  0.0059     81.162 < 2.2e-16 ***
x6           1 0.0056  0.0056     77.582 < 2.2e-16 ***
x5           1 0.0005  0.0005      6.761  0.009392 ** 
Residuals 1844 0.1337  0.0001                         
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

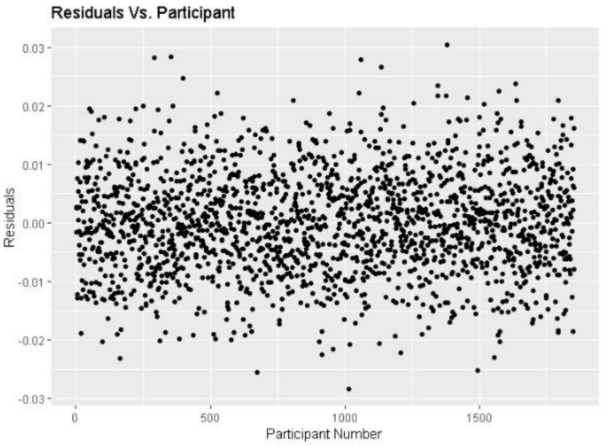
Checking the validity of the model once more by verifying the error variable assumptions:

Figure 22

Various Plots Involving Residuals After Variable Transformation: Normal Probability Plot, Residuals Vs Fitted Value, Histogram, and Residuals Vs Participant







As seen from the plots above, the error variable is normally distributed – even better than that of the y variable model. The histogram is less right-skewed as well. The variance is equal throughout, and the error variables are independent. Hence, the assumptions of the error variable are met; the model is now valid.

Question-3) Construct a 95% confidence interval on the mean responce for...

Here the confidence interval of an individual's lung capacity is generated for a male with family income being 12500 liras who had seen anti-tobacco messages on TV/on billboards/in newspapers and even though had been warned about the side effects of the smoking but did not take a course regarding the hazards of smoking, smoked 1.65 years during his high school years, had one relative who smoked when he was a student while having 2 close friends who also smoked. The result is:

C.I. = [0.8859, 0.8882]

Question-4) Construct a 95% prediction interval for...

Unlike in question 3, here in this case the researcher needs to predict the lung capacity interval for a female with a family that had 2 smokers when she was in high school and earned 10350 liras per month. She also had not been warned by her family. She not only saw advertisements against smoking but also taken a course that explained the hazards of smoking. Even though she took a course, she smoked 3.35 years with more than 3 of her smoking friends. The predicted result is:

C.I. = [0.7956, 0.8291]

3. Conclusion

Overall, the linear regression model pointed out that the strongest factor on harming and decreasing the lung capacity is duration of smoking before the age of 18. It also pointed out that many respondents who had a close person to them who also smokes were affected with lower lung capacity. This may lead to them smoking under peer pressure. It also pointed out that many respondents who were warned about smoking by their family members seemed to have more lung capacity, perhaps an indicator of them smoking less than others.

The model also indicated that gender and duration of smoking have a coupling effect on lung capacity. Perhaps biological differences and the effect of nicotine on each gender cause different effects.