IE266
ENGINEERING STATISTICS

Zühal Ceren Aydın 2444370

Burak Baki Bakır 2530103

İlkay Koç 2444792

Tuna Özkuşaksız 2444875

Question 1

Part a

Are the duration of smoking and gender independent?

To test this claim the hypothesis testing is performed.

Test procedure applied is like the following.

0. Parameter: Are the duration of smoking and gender independent?

1. H_0: Duration of smoking and gender are independent.

H_α: Duration of smoking and gender are not independent.

2. Test statistic: $\chi_0^2 \sim \chi^2_{(k1-1)(k2-1)}$

3. $\chi_0^2 = 7.4554$

```
X-squared = 7.4554, df = 12, p-value = 0.8261
```

4. Comparing chi-squares $\chi_0^2 < \chi^2_{0.05,12}$ $\quad 7.4554 < 21.0260$

5. Comparing p-values $0.05 < 0.82$

To test the independence of the duration of smoking and gender, firstly the data is divided into categories which represents gender and duration. Gender is divided as female and male, duration is divided into 13 categories which are the duration of smoking as intervals from 0 to 10 by increasing 0.5 until 6.5 and going cumulative from 6.5 to 10. By using these categories a contingency table is formed that shows how many people in those categories. Later, by using this table chi-square test is applied at $\alpha = 0.05$. Chi-square values and p-values are found as the result of this test. When these values are compared for the null and the alternative hypotheses, it's seen that $\chi_0^2 < \chi^2_{0.05,12}$, and this implies that the alternative hypotheses should be rejected. Also, when p-values are checked, it's seen that $0.05 < 0.82$, and this implies that the alternative hypotheses should be rejected. With both these comparisons it can be concluded that the null hypothesis is true, so duration of smoking and gender are independent.

	Var1	Var2	Freq
1	1	Female	100
14	1	Male	86
2	2	Female	167
15	2	Male	172
3	3	Female	159
16	3	Male	165
4	4	Female	139
17	4	Male	127
5	5	Female	132
18	5	Male	120
6	6	Female	82
19	6	Male	76
7	7	Female	48
20	7	Male	42
8	8	Female	30
21	8	Male	47
9	9	Female	45
22	9	Male	45
10	10	Female	9
23	10	Male	10
11	11	Female	12
24	11	Male	8
12	12	Female	4
25	12	Male	5
13	13	Female	11
26	13	Male	12

Part b

Are the number of smokers in the family and the number of smokers in the friend group independent?

To test this claim the hypothesis testing is performed.

Test procedure applied is like the following.

0. Parameter: Are the number of smokers in the family and the number of smokers in the friend group independent?

1. H_0: The number of smokers in the family and the number of smokers in the friend group are independent.

H_a: The number of smokers in the family and the number of smokers in the friend group are not independent.

2. Test statistic: χ_0^2 $\chi^2_{(k1-1)(k2-1)}$

3. $\chi_0^2 = 11.762$

```
X-squared = 11.762, df = 9, p-value = 0.2271
```

4. Comparing chi-squares $\chi_0^2 < \chi^2_{0.05,9}$ $11.762 < 16.919$

5. Comparing p-values $0.05 < 0.22$

To test the independence of the number of smokers in the family and the number of smokers in the friend group, firstly the data is divided into categories which represents the number of smokers in the family and the number of smokers in the friend group. Both the number of smokers in the family and the number of smokers in the friend group are divided into four categories: zero, one, two, and more than three. By using these categories a contingency table is formed that shows how many people in those categories. Later, by using this table chi-square test is applied at $\alpha = 0.05$. Chi-square values and p-values are found as the result of this test. When these values are compared for the null and the alternative hypotheses, it's seen that $\chi_0^2 < \chi^2_{0.05,9}$, and this implies that the alternative hypotheses should be rejected. Also, when p-values are checked, it's seen that $0.05 < 0.22$, and this implies that the alternative hypotheses should be rejected. With both these comparisons it can be concluded that the null hypothesis is true, so duration of smoking and gender are independent.

	Var1	Var2	Freq
13	More than 3	Zero	103
9	More than 3	Two	49
5	More than 3	One	69
1	More than 3	More than 3	56
14	One	Zero	201
10	One	Two	129
6	One	One	96
2	One	More than 3	100
15	Two	Zero	277
11	Two	Two	144
7	Two	One	137
3	Two	More than 3	133
16	Zero	Zero	147
12	Zero	Two	71
8	Zero	One	63
4	Zero	More than 3	78

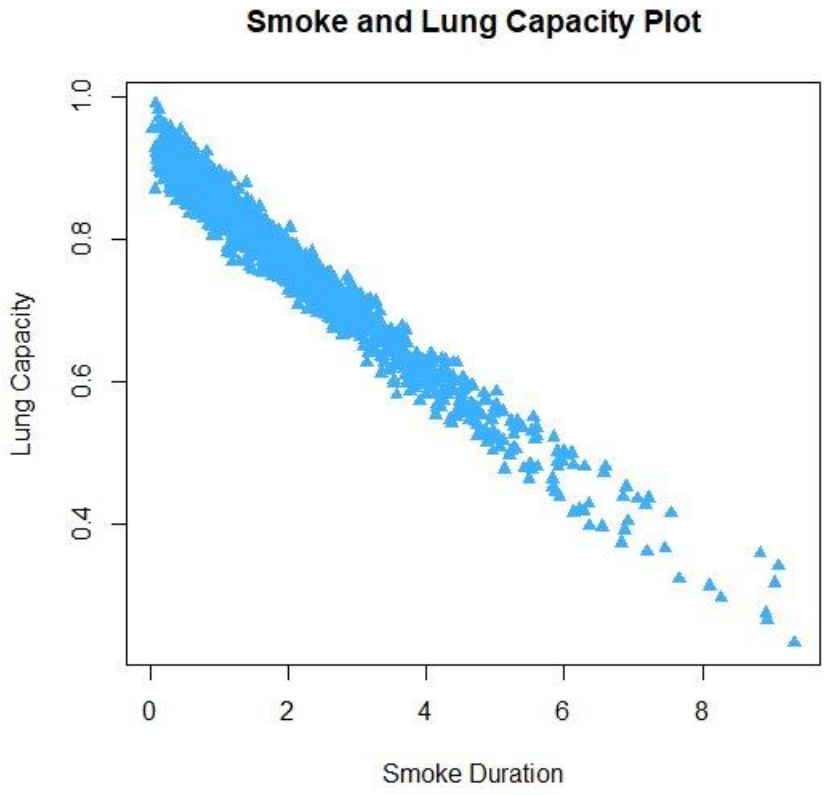
2. Question 2

The linear regression model is created by using all variables and it is plotted the response against the continuous variables with groups of significant categorical variables may cause deviation from normality. We also utilized stepwise regression to build the model, which clearly explained each stage of the selection process and provided the adjusted r2 value. This linear regression give the accurate result that since that r^2 is remarkable high and the p-value of this model is significantly low. On the other hand, some variables have no effect on the model and it may cause deviation from normality with some errors. Hence, interaction terms are necessary and should be added to the model to prevent this.

Continuous Variables

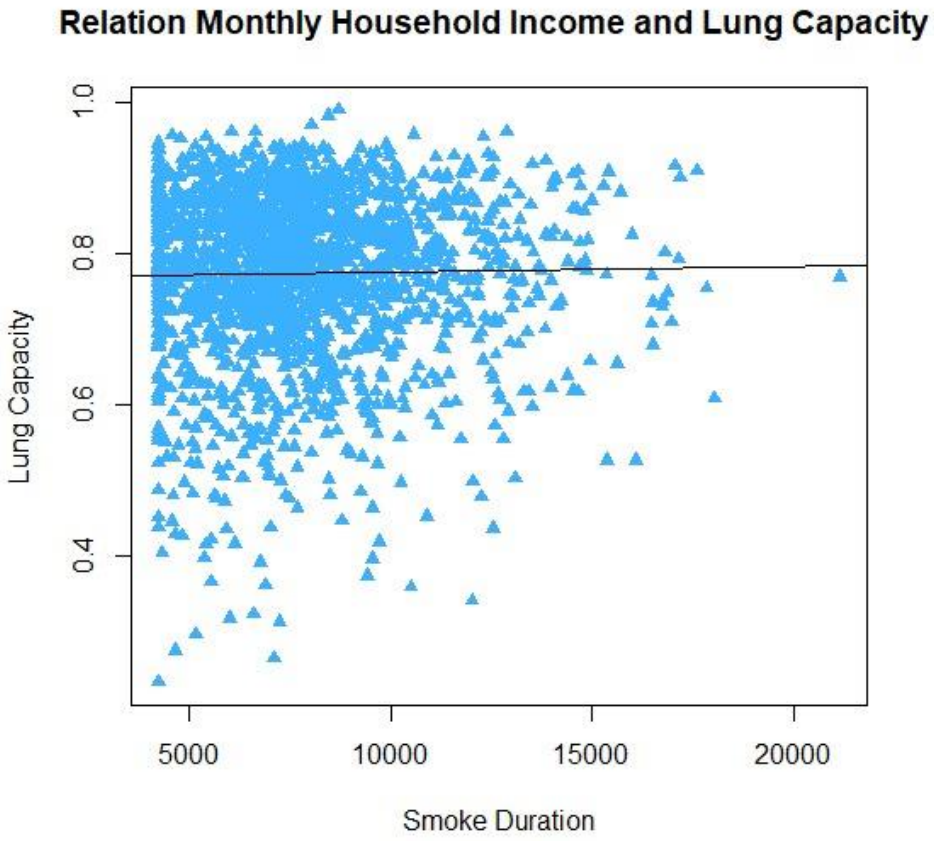
Duration Of Smoking And Lung Capacity

While duration is increasing, the lung capacity is decreasing and the first and the second degree of derivatives of the line is near 0 and negative. Hence it means that the inverse relationship between duration and lung capacity, respectively.



Lung Capacity And Income

Relation monthly household income and the lung capacity does not have a directly relationship. The slope of the fitted line does not have notable. After than that, chi-square independence tests are applied to categorical variables to see if there are any interaction between them.



Categorical Variables

Lung capacity and duration of smoking- gender

In the figure below, residual plot of duration of smoking and lung capacity is formed for two genders. In this plot, the blue color represents males while the red color represents females. Fitted lines formed in this plot have different slopes for males and females. Thus, it can be concluded that duration of smoking and the gender might have an interaction since slope changes when the gender changes. Furthermore, when the slopes are examined, it's seen that fitted line formed for the males have a deeper slope, and this implies smoking may decrease the lung capacity of the males faster than the females.

Lung capacity and duration of smoking- school informing

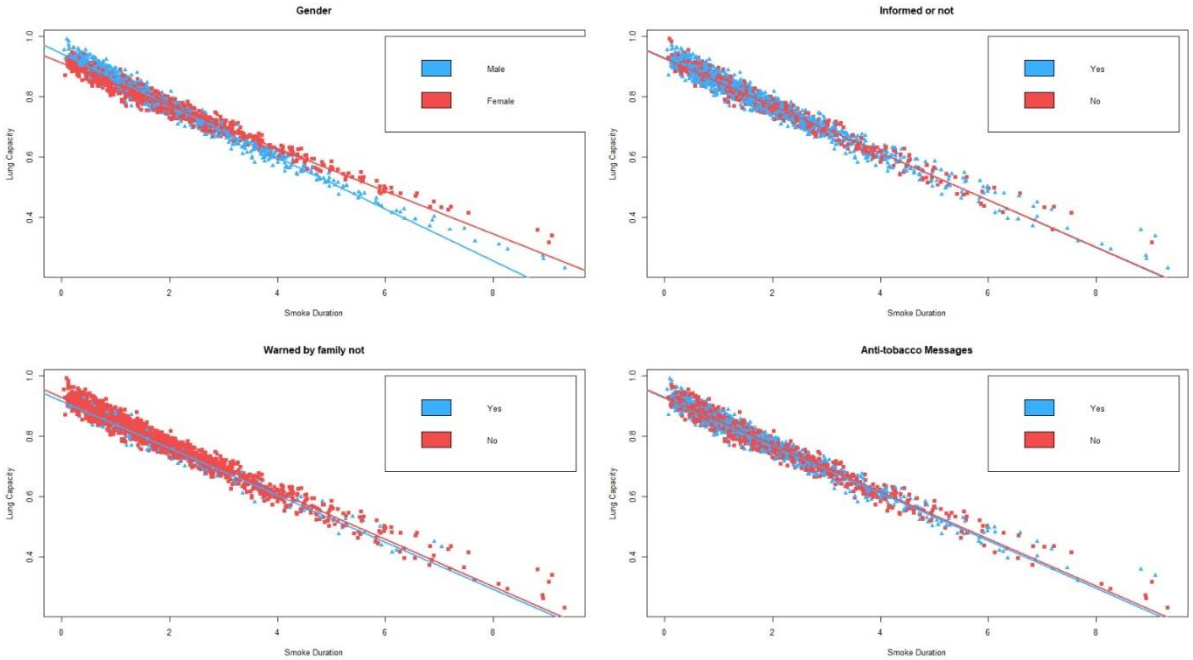
In the figure above, residual plot of duration of smoking and lung capacity is formed for whether the people are informed about the hazards of smoking in any of their courses in high school or not. Blue color represents the people who are informed while the red color represents the people who are not informed. As it's seen in the figure, the slope of the fitted line is the same for both informed and uninformed people. Thus, it can be concluded that there isn't a relation between the duration of smoking and whether a person is informed by their high school or not.

Lung capacity and duration of smoking- family warnings

In the figure, the blue dots represent that the people who had warned about the hazards of smoking in their family, and also the red dots represent people who did not receive this warning. One more time, the slopes of this different groups looks like similar and it refers that there is no relationship between people's smoking and being warned about this issue by their families.

Lung capacity and duration of smoking- Anti Tobacco Messages

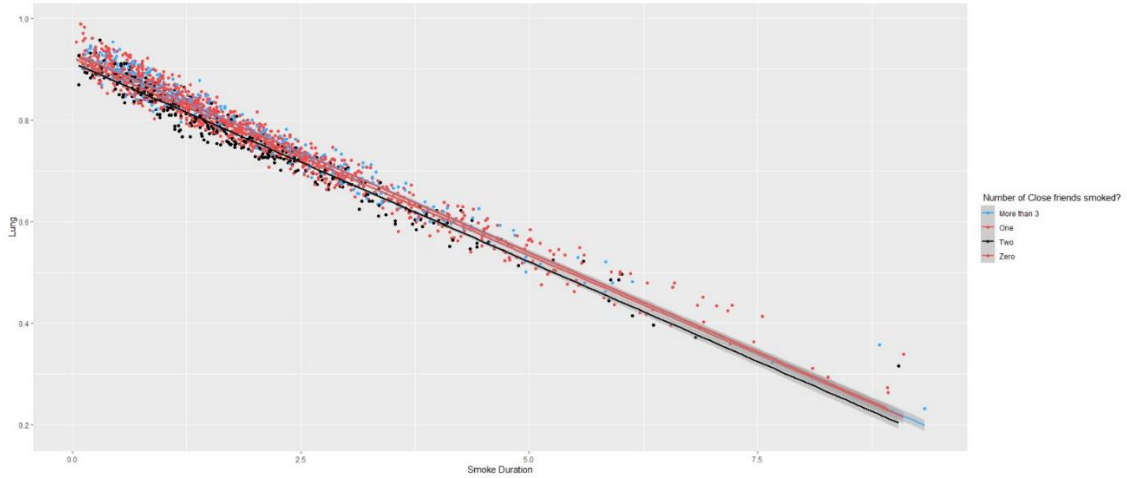
In the figure, the blue dots represent that Have seen anti-tobacco messages on TV/on billboards or in newspapers, and also red dots represent people who have not. The little difference is observed between their slopes. So, we can say that there is no correlation between them.



Lung capacity and duration of smoking- close friends

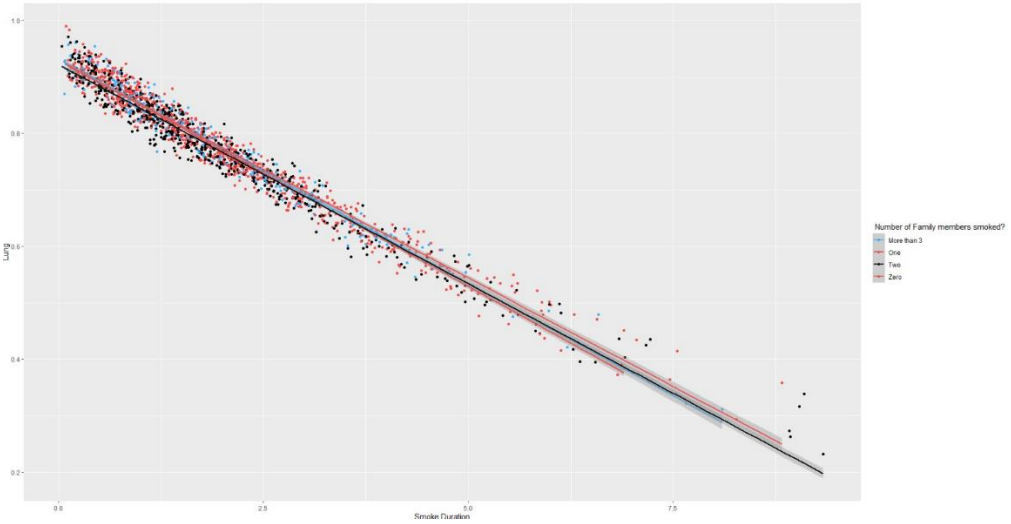
In the below figure, residual plot of duration of smoking and lung capacity is formed for the number of friends in the people's close friends group who smoked in high school. Blue color

represents people having more than three friends, red color represents one, black color represents two, and orange color represents zero. When the figure is examined, it’s seen that although there aren’t significant differences, the slopes are not the same for different numbers of friends. Thus, it can be concluded that there is not a certain relation between the duration of smoking and the number of friends in the people’s close friends group who smoked in high school, so adding this to the model might not give precise opinions.



Lung capacity and duration of smoking- family

The residual plot of smoking time and lung capacity for the number of friends in the smokers' family group is shown in the image below. Orange indicates zero, red represents one, black represents two, and blue denotes persons who have more than three pals. When the graphic is inspected, it can be observed that the slopes are different for various friend counts even if there aren't any appreciable variances. As a result, it may be inferred that there is no clear relationship between the length of smoking and the number of friends in the smokers' family group, and that included this in the model may result in imprecise conclusions.



Lung capacity and income- gender

When we look below table, we observe the relationship between lung capacity and income and the distribution of this relationship by gender. Blue dots represent male while red dots represent female. Although the reds have created a more central image around the lines than the blues, it is not possible to say that there is a clear difference between females and males in terms of income and lung capacity. In addition, the lines formed by female and male are very similar to each other. For these reasons, we don't need to add this interaction to the equation.

Lung capacity and income- warned by family

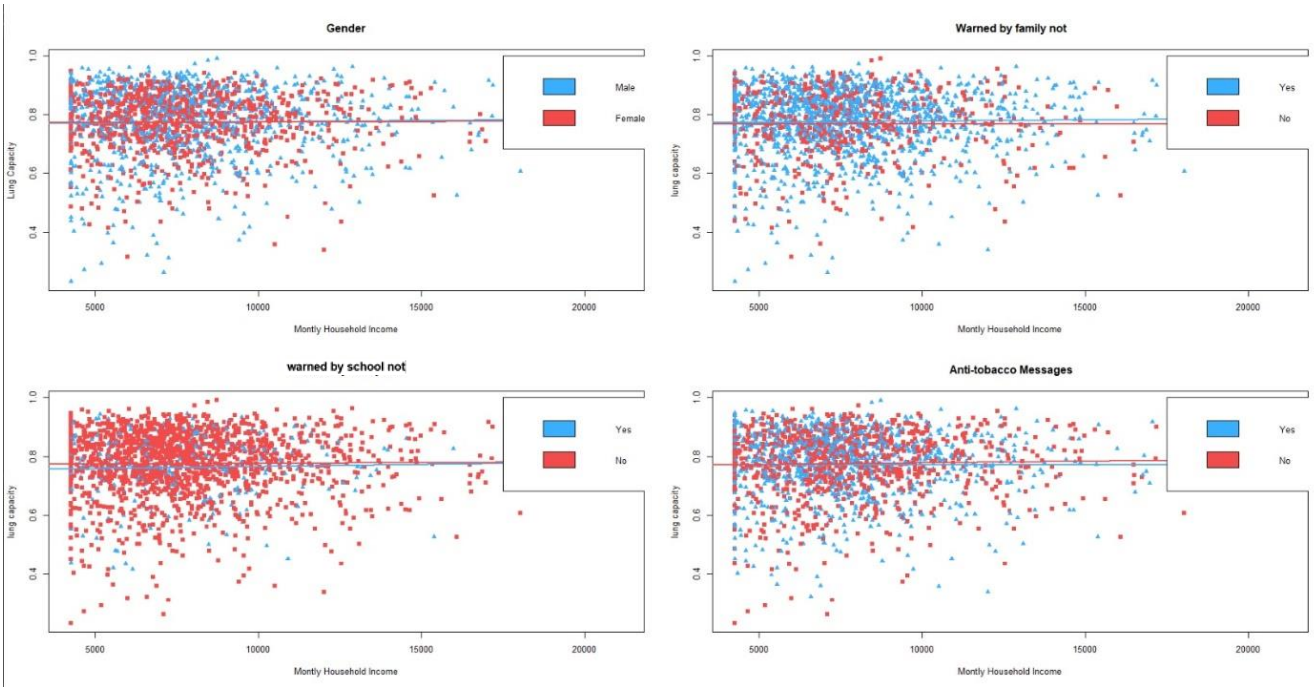
We see a below table similar to the previous two tables. The red dots represent that no family warning was given, and the blue dots represent that the individual's family warned the individual. When we look at the table, we observe that there are more people who are not warned by their families. However, we do not observe a smooth relationship between being warned by the family or not, which affects lung capacity and income. It also shouldn't be included in the model.

Lung capacity and income- warned by school not

In below table, the red dots show that individuals were warned about the harms of smoking at school when they were young, and the blue ones show that they were warned about the harms of smoking. Again, we do not see a significant difference between the slopes of the two lines. Likewise, when we look at the distribution of the points, it is not possible to observe much difference. Therefore, the school warning does not have a strong relationship with the income and lung capacity. As a result, we should not add this interaction to the model.

Lung capacity and income- anti tobacco messages

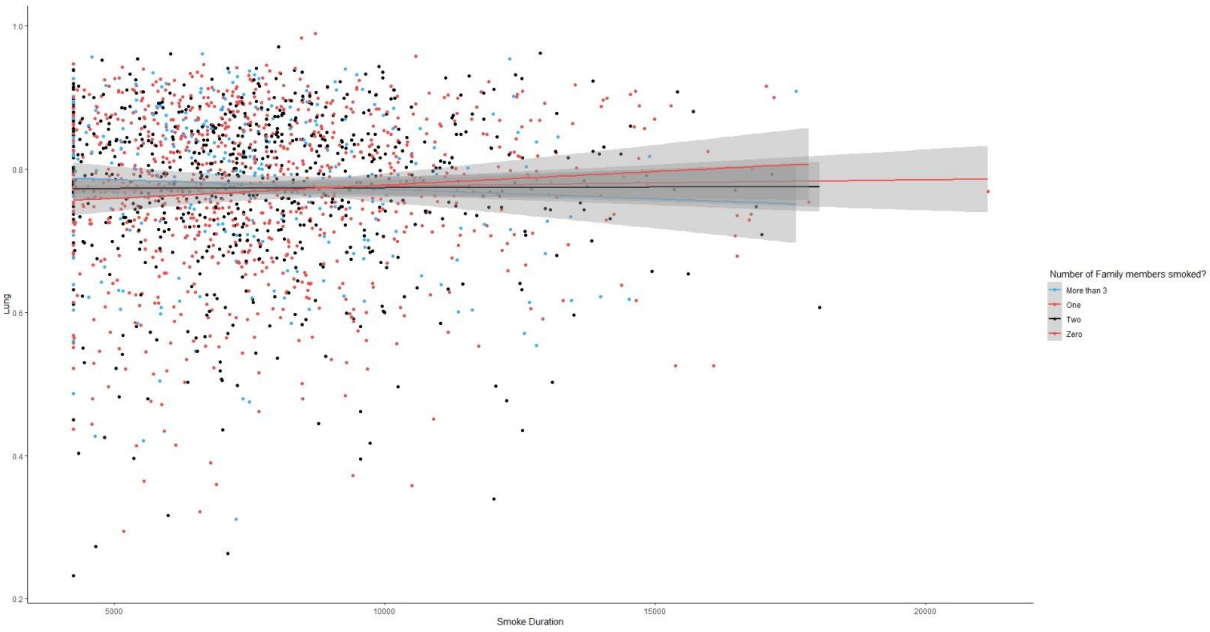
When we look below table, it is possible to say that we obtain an image similar to the female-male table. This table again gives income and lung capacity relationships. This time, the blue dots represent anti-tobacco messages, and the red dots represent no anti-tobacco messages. Since we could not observe any relationship between the distribution of red and blue dots, we should not include this relationship in the model.



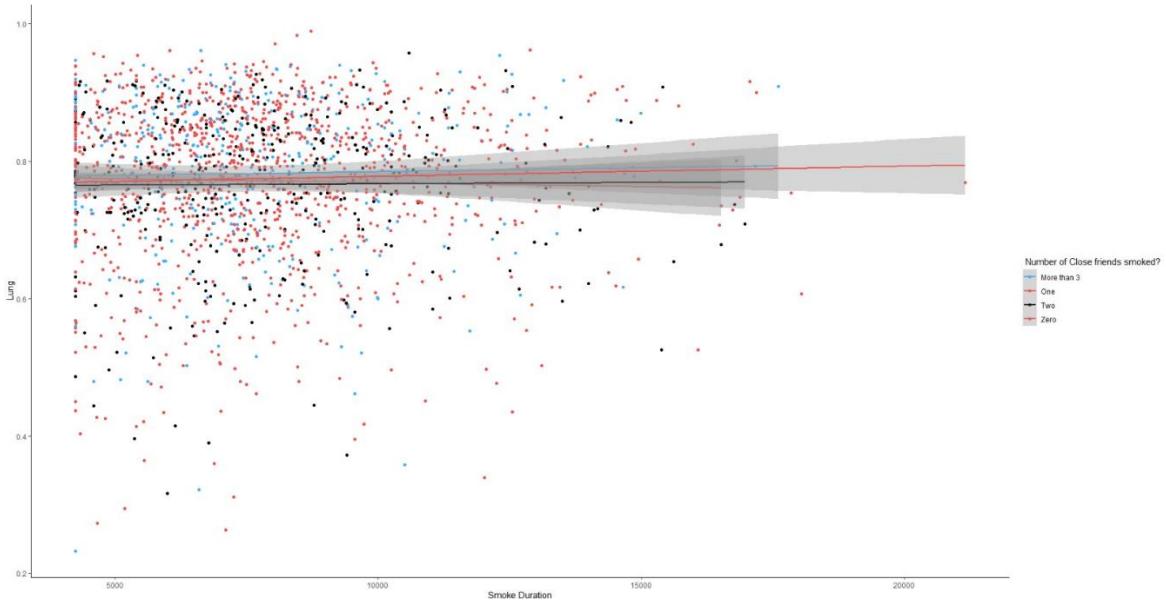
Lung capacity and income- family members

When we look at below table, depending on how many smoking relatives are in the person's immediate family, we witness various conditions in this relationship. We see a highly complex picture when we look at the graph. It makes sense to assume that someone with more smoking family members will have smaller lungs. However, it doesn't appear that way; rather, a friend who smokes more than three cigarettes a day has healthier lungs. Additionally, despite there being 1, 2, and 3 smokers in the household, there is no evidence of a harmonious relationship. In one instance, we observe an increase, whereas in the other, a drop. Since there is no significant correlation between these variables, including them in the model will produce false results.

Lung capacity and income- friends



When we look at below table, we observe the relationship between smoke duration and lung capacity. We observe different situations of this relationship according to the number of smoking friends around the individual. When we look at the graph, we are faced with a very complex picture. We would logically expect an individual with more smoking friends to have less lung capacity. However, the situation does not seem like that, on the contrary, the lungs of a friend who smokes more than 3 cigarettes are in better condition. In addition, a smooth relationship is not observed in the number of smoking friends being 1,2 and 3. We see it increased in one case and decreased in the other. Therefore, there is no strong relationship between these variables, and adding this to the model will give us an incorrect result.



In this part of the report, plots of continuous variables with groups of significant categorical variables are formed to see whether any interaction terms are necessary and adding them to the model gives an opinion. As a result of these plots, adding the interaction between gender and duration of smoking seemed reasonable by looking at the plot they provided. Later, the interaction between the family warning and gender are added to the model by examining the chi-square test applied to them.

```
data:  dataq$Gender and dataq$fWarn
X-squared = 6.0278, df = 1, p-value = 0.01408
```

With these new interactions added to the model, a new model is formed. And the ANOVA table of the initial model(upper) is transformed into a new ANOVA table including interactions added(lower).

```
Analysis of Variance Table

Response: Lung
             Df  Sum Sq Mean Sq    F value    Pr(>F)
Smoke_Dur     1 22.4477 22.4477 55632.8255 < 2.2e-16 ***
Month_house   1  0.0002  0.0002     0.6125   0.43395
Gender        1  0.0015  0.0015     3.6784   0.05528 .
F_Mem_Smoke   3  0.0198  0.0066    16.3209 1.806e-10 ***
Cf_Smoke      3  0.1066  0.0355    88.0461 < 2.2e-16 ***
fWarn         1  0.0290  0.0290    71.9803 < 2.2e-16 ***
Inform        1  0.0002  0.0002     0.4425   0.50603
Tv            1  0.0010  0.0010     2.4294   0.11925
Residuals  1840  0.7424  0.0004
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

```
Analysis of Variance Table

Response: Lung
                 Df  Sum Sq Mean Sq    F value    Pr(>F)
Smoke_Dur         1 22.4477 22.4477 77632.5407 < 2.2e-16 ***
Month_house       1  0.0002  0.0002     0.8547   0.35535
Gender            1  0.0015  0.0015     5.1331   0.02359 *
F_Mem_Smoke       3  0.0198  0.0066    22.7750 1.779e-14 ***
Cf_Smoke          3  0.1066  0.0355   122.8635 < 2.2e-16 ***
fWarn             1  0.0290  0.0290   100.4445 < 2.2e-16 ***
Inform            1  0.0002  0.0002     0.6174   0.43211
Tv                1  0.0010  0.0010     3.3901   0.06575 .
Gender:fWarn      1  0.0009  0.0009     3.0884   0.07902 .
Smoke_Dur:Gender  1  0.2101  0.2101   726.5301 < 2.2e-16 ***
Residuals      1838  0.5315  0.0003
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Categorical values Comparison

All categorical values were compared with each other by chi square test and it was determined that family warning and gender could be related to each other and added to the model.

Question 2.b.

Backward stepwise regression regression is a method for performing stepwise regression that starts with a completed model and gradually removes variables at each stage to find a smaller model that best fits the data. In each step, we eliminate the variable with the highest p-value among the variables and to do that we decided that significance level as 0.05. We continued that until there is notable p-value.

The Initial Model

The initial model is the model that we found by plotting the response against the continuous variables with groups of significant categorical variables and added necessaries to the model.

It consists of smoke duration, monthly household income, gender, number of family member smoke, number of close friends smoked, warned by family, informed in school, anti-tobacco messages, gender times warned by family, gender times smoke duration. It is observed that the highest p-value is belong to monthly household income. To do other step of elimination, we should eliminate the highest p-value that we found as the monthly household income. Additionally, noted that the R-squared is 0.9771.

```
Call:
lm(formula = Lung ~ Smoke_Dur + Month_house + Gender + F_Mem_Smoke +
    Cf_Smoke + fWarn + Inform + Tv + Gender * fWarn + Gender *
    Smoke_Dur, data = dataq)

Residuals:
      Min        1Q    Median        3Q       Max
-0.052076 -0.011648 -0.000808  0.010389  0.093411

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.815e-01  4.323e-03 203.895  < 2e-16 ***
Smoke_Dur        -5.565e-02  8.941e-04 -62.244  < 2e-16 ***
Month_house       4.276e-08  1.537e-07   0.278   0.7809
Gender            2.912e-02  2.442e-03  11.926  < 2e-16 ***
F_Mem_SmokeOne   -1.942e-03  1.267e-03  -1.533   0.1255
F_Mem_SmokeTwo   -6.376e-03  1.212e-03  -5.261 1.60e-07 ***
F_Mem_SmokeZero   1.215e-03  1.363e-03   0.892   0.3727
Cf_SmokeOne      -9.705e-03  1.259e-03  -7.711 2.03e-14 ***
Cf_SmokeTwo      -2.018e-02  1.237e-03 -16.316  < 2e-16 ***
Cf_SmokeZero     -1.770e-03  1.090e-03  -1.624   0.1045
fWarn             6.589e-03  3.813e-03   1.728   0.0842 .
Inform           -2.305e-04  8.646e-04  -0.267   0.7898
Tv                1.018e-03  7.913e-04   1.286   0.1985
Gender:fWarn      3.694e-03  2.333e-03   1.583   0.1135
Smoke_Dur:Gender -1.518e-02  5.630e-04 -26.954  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.017 on 1838 degrees of freedom
Multiple R-squared:  0.9772,    Adjusted R-squared:  0.9771
F-statistic:  5636 on 14 and 1838 DF,  p-value: < 2.2e-16
```

The Second Model

After the first iteration, the second model consists of smoke duration, gender, number of family member smoke, number of close friends smoked, warned by family, informed in school, anti-tobacco messages, gender times warned by family, gender times smoke duration. It is observed that the highest p-value is belong to informed in school. To do other step of elimination, we should eliminate the highest p-value that we found as informed in school. Additionally, noted that the R-squared is 0.9771 again.

```
Call:
lm(formula = Lung ~ Smoke_Dur + Gender + F_Mem_Smoke + Cf_Smoke +
    fWarn + Inform + Tv + Gender * fWarn + Gender * Smoke_Dur,
    data = dataq)

Residuals:
      Min        1Q    Median        3Q       Max
-0.052091 -0.011627 -0.000791  0.010313  0.093276

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.8817586  0.0041967 210.109  < 2e-16 ***
Smoke_Dur        -0.0556519  0.0008939 -62.261  < 2e-16 ***
Gender            0.0291371  0.0024403  11.940  < 2e-16 ***
F_Mem_SmokeOne   -0.0019299  0.0012658  -1.525   0.1275
F_Mem_SmokeTwo   -0.0063644  0.0012109  -5.256 1.65e-07 ***
F_Mem_SmokeZero   0.0012200  0.0013626   0.895   0.3707
Cf_SmokeOne      -0.0097047  0.0012583  -7.713 2.00e-14 ***
Cf_SmokeTwo      -0.0201772  0.0012365 -16.318  < 2e-16 ***
Cf_SmokeZero     -0.0017610  0.0010891  -1.617   0.1060
fWarn             0.0066044  0.0038118   1.733   0.0833 .
Inform           -0.0002355  0.0008642  -0.273   0.7853
Tv                0.0010205  0.0007910   1.290   0.1972
Gender:fWarn      0.0036908  0.0023328   1.582   0.1138
Smoke_Dur:Gender -0.0151780  0.0005628 -26.969  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.017 on 1839 degrees of freedom
Multiple R-squared:  0.9772,    Adjusted R-squared:  0.9771
F-statistic:  6073 on 13 and 1839 DF,  p-value: < 2.2e-16
```

The Third Model

After the second iteration, the third model consists of smoke duration, gender, number of family member smoke, number of close friends smoked, warned by family, anti-tobacco messages, gender times warned by family, gender times smoke duration. It is observed that the highest p-value is belong to anti-tobacco messages. To do other step of elimination, we should eliminate the highest p-value that we found as anti-tobacco messages. Additionally, noted that the R-squared is 0.9771 again.

```
Call:
lm(formula = Lung ~ Smoke_Dur + Gender + F_Mem_Smoke + Cf_Smoke +
    fWarn + Tv + Gender * fWarn + Gender * Smoke_Dur, data = dataq)

Residuals:
      Min        1Q    Median        3Q       Max
-0.052256 -0.011633 -0.000811  0.010271  0.093344

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.8817358  0.0041948 210.198  < 2e-16 ***
Smoke_Dur        -0.0556643  0.0008925 -62.372  < 2e-16 ***
Gender            0.0291093  0.0024375  11.942  < 2e-16 ***
F_Mem_SmokeOne   -0.0019304  0.0012655  -1.525    0.127
F_Mem_SmokeTwo   -0.0063626  0.0012106  -5.256 1.65e-07 ***
F_Mem_SmokeZero   0.0012273  0.0013620   0.901    0.368
Cf_SmokeOne      -0.0097073  0.0012579  -7.717 1.94e-14 ***
Cf_SmokeTwo      -0.0201887  0.0012355 -16.340  < 2e-16 ***
Cf_SmokeZero     -0.0017648  0.0010887  -1.621    0.105
fWarn             0.0065864  0.0038103   1.729    0.084 .
Tv                0.0010258  0.0007906   1.298    0.195
Gender:fWarn      0.0037022  0.0023319   1.588    0.113
Smoke_Dur:Gender -0.0151716  0.0005622 -26.988  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.017 on 1840 degrees of freedom
Multiple R-squared:  0.9772,    Adjusted R-squared:  0.9771
F-statistic:  6582 on 12 and 1840 DF,  p-value: < 2.2e-16
```

The Fourth Model

After the third iteration, the fourth model consists of smoke duration, gender, number of family member smoke, number of close friends smoked, warned by family, gender times warned by family, gender times smoke duration. Actually, 0 family member smoke has the highest p-value, however this variable is one of the component of number of family smoked and since it is observed that the highest p-value is belong to gender times warned by family among the independent variables. To do other step of elimination, we should eliminate the highest p-value that we found as gender times warned by family. Additionally, noted that the R-squared is 0.9771 again.

```
Call:
lm(formula = Lung ~ Smoke_Dur + Gender + F_Mem_Smoke + Cf_Smoke +
    fWarn + Gender * fWarn + Gender * Smoke_Dur, data = dataq)

Residuals:
      Min        1Q    Median        3Q       Max
-0.052772 -0.011527 -0.000722  0.010220  0.093922

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.8822223  0.0041788 211.120  < 2e-16 ***
Smoke_Dur        -0.0556468  0.0008925 -62.348  < 2e-16 ***
Gender            0.0291304  0.0024379  11.949  < 2e-16 ***
F_Mem_SmokeOne   -0.0019471  0.0012656  -1.538   0.1241
F_Mem_SmokeTwo   -0.0063985  0.0012105  -5.286 1.40e-07 ***
F_Mem_SmokeZero   0.0012037  0.0013621   0.884   0.3770
Cf_SmokeOne      -0.0096776  0.0012579  -7.693 2.32e-14 ***
Cf_SmokeTwo      -0.0201989  0.0012357 -16.346  < 2e-16 ***
Cf_SmokeZero     -0.0017434  0.0010888  -1.601   0.1095
fWarn             0.0065618  0.0038109   1.722   0.0853 .
Gender:fWarn      0.0037282  0.0023322   1.599   0.1101
Smoke_Dur:Gender -0.0151857  0.0005622 -27.013  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.017 on 1841 degrees of freedom
Multiple R-squared:  0.9772,    Adjusted R-squared:  0.9771
F-statistic:  7178 on 11 and 1841 DF,  p-value: < 2.2e-16
```

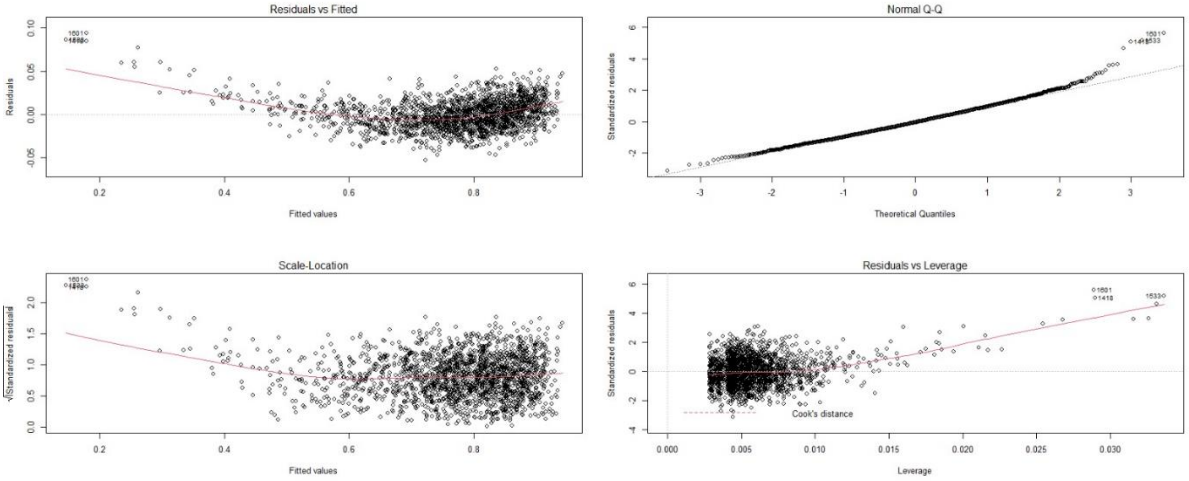
The Fifth Model

After the fourth iteration, we can not find a significantly high p-value and founded as final model.

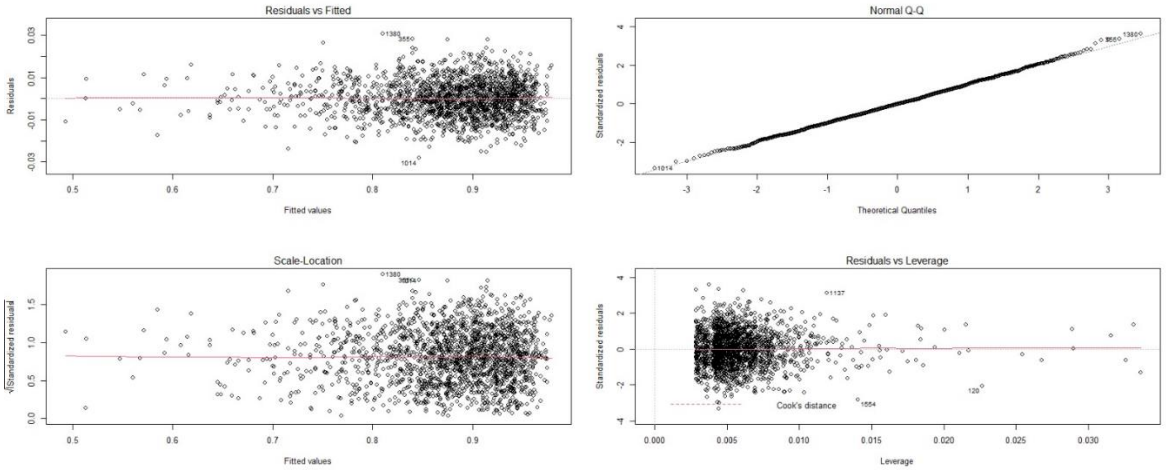
```
Call:
lm(formula = Lung ~ Smoke_Dur + Gender + F_Mem_Smoke + Cf_Smoke +
    fWarn + Gender * Smoke_Dur, data = dataq)

Residuals:
      Min        1Q    Median        3Q       Max
-0.053014 -0.011514 -0.000911  0.010441  0.094124

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.8771144  0.0026941 325.569  < 2e-16 ***
Smoke_Dur        -0.0556282  0.0008928 -62.306  < 2e-16 ***
Gender            0.0323720  0.0013540  23.908  < 2e-16 ***
F_Mem_SmokeOne   -0.0019575  0.0012662  -1.546    0.122
F_Mem_SmokeTwo   -0.0063841  0.0012110  -5.272 1.51e-07 ***
F_Mem_SmokeZero   0.0012108  0.0013627   0.889    0.374
Cf_SmokeOne      -0.0096745  0.0012585  -7.688 2.42e-14 ***
Cf_SmokeTwo      -0.0201755  0.0012362 -16.321  < 2e-16 ***
Cf_SmokeZero     -0.0017538  0.0010892  -1.610    0.108
fWarn             0.0123646  0.0011607  10.653  < 2e-16 ***
Smoke_Dur:Gender -0.0151909  0.0005624 -27.011  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01701 on 1842 degrees of freedom
Multiple R-squared:  0.9772,    Adjusted R-squared:  0.9771
F-statistic:  7889 on 10 and 1842 DF,  p-value: < 2.2e-16
```

Furthermore, the residual plots of the final model of the backward regression can be see below figures.

As can be seen below, the found figures are convex, hence the transformation must be applied to the model and we applied the transformation by taking the square root of the values.



As a result of these procedures the model is become like in the below figure.



Tranformed Model

After the transformation, the outputs of the model have changed as seen below. In addition, the R-squared p value increased to 0.9842.

```
Call:
lm(formula = sqrt(Lung) ~ Smoke_Dur + Gender + F_Mem_Smoke +
    Cf_Smoke + fWarn + Gender * Smoke_Dur, data = dataq)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0282059 -0.0056935 -0.0000674  0.0056252  0.0306938

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.9389406  0.0013481 696.494  < 2e-16 ***
Smoke_Dur        -0.0323354  0.0004468 -72.378  < 2e-16 ***
Gender            0.0200911  0.0006775  29.654  < 2e-16 ***
F_Mem_SmokeOne   -0.0010442  0.0006336  -1.648   0.0995 .
F_Mem_SmokeTwo   -0.0039289  0.0006060  -6.484 1.15e-10 ***
F_Mem_SmokeZero   0.0004493  0.0006819   0.659   0.5100
Cf_SmokeOne      -0.0053162  0.0006297  -8.442  < 2e-16 ***
Cf_SmokeTwo      -0.0113962  0.0006186 -18.424  < 2e-16 ***
Cf_SmokeZero     -0.0010243  0.0005450  -1.879   0.0603 .
fWarn             0.0070049  0.0005808  12.061  < 2e-16 ***
Smoke_Dur:Gender -0.0100647  0.0002814 -35.764  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.00851 on 1842 degrees of freedom
Multiple R-squared:  0.9843,    Adjusted R-squared:  0.9842
F-statistic: 1.155e+04 on 10 and 1842 DF,  p-value: < 2.2e-16
```

The ANOVA table of the initial model is shown below.

```
Analysis of Variance Table

Response: Lung
              Df  Sum Sq Mean Sq    F value    Pr(>F)
Smoke_Dur      1 22.4477 22.4477 55632.8255 < 2.2e-16 ***
Month_house    1  0.0002  0.0002     0.6125   0.43395
Gender         1  0.0015  0.0015     3.6784   0.05528 .
F_Mem_Smoke    3  0.0198  0.0066    16.3209 1.806e-10 ***
Cf_Smoke       3  0.1066  0.0355    88.0461 < 2.2e-16 ***
fWarn          1  0.0290  0.0290    71.9803 < 2.2e-16 ***
Inform         1  0.0002  0.0002     0.4425   0.50603
Tv             1  0.0010  0.0010     2.4294   0.11925
Residuals   1840  0.7424  0.0004
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

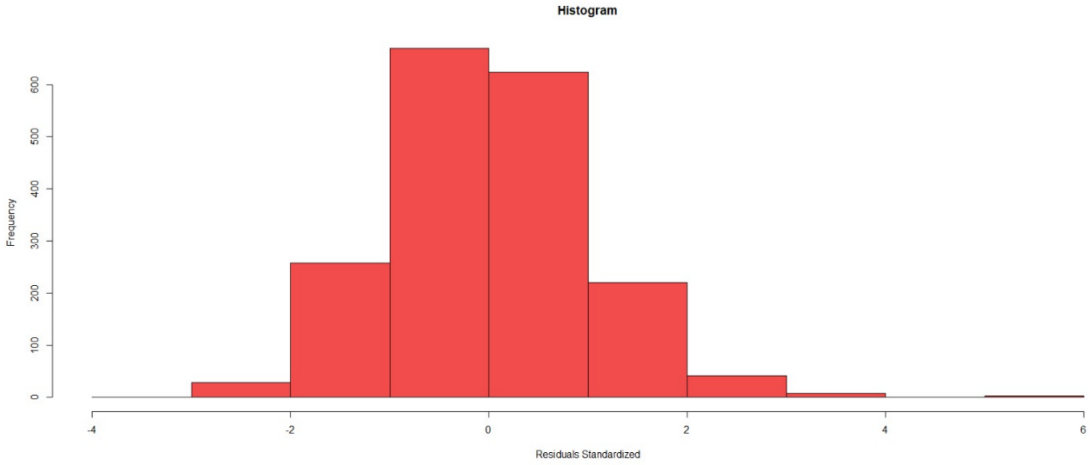
Through the backward elimination steps, the ANOVA table of the final model become the model shown in the below.

```
Analysis of Variance Table

Response: sqrt(Lung)
                 Df Sum Sq Mean Sq    F value Pr(>F)
Smoke_Dur         1 8.2201  8.2201 113511.962 <2e-16 ***
Gender            1 0.0000  0.0000      0.033 0.8559
F_Mem_Smoke       3 0.0068  0.0023     31.410 <2e-16 ***
Cf_Smoke          3 0.0335  0.0112    154.425 <2e-16 ***
fWarn             1 0.0092  0.0092    126.735 <2e-16 ***
Smoke_Dur:Gender  1 0.0926  0.0926   1279.076 <2e-16 ***
Residuals      1842 0.1334  0.0001
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

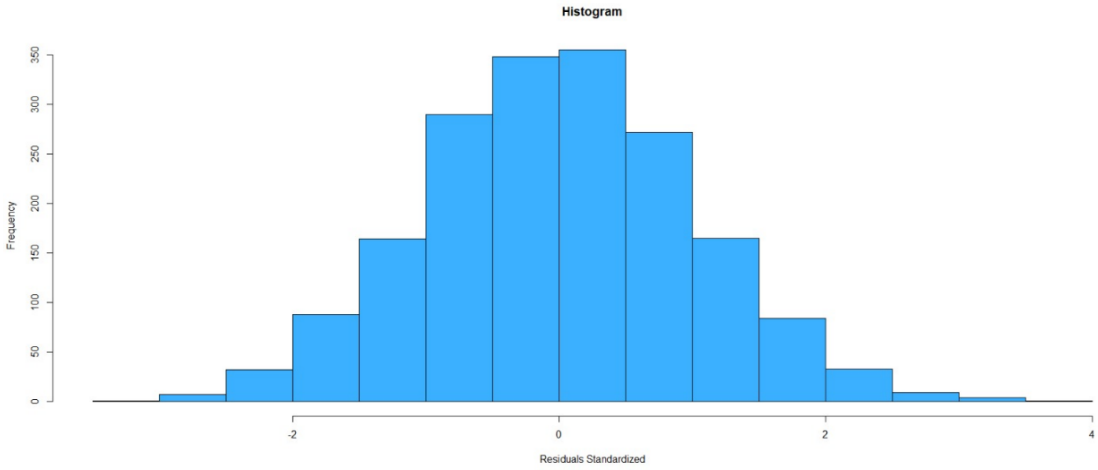
After the transformation operation the ANOVA table of the transformed model is shown in the below.

```
Analysis of Variance Table

Response: sqrt(Lung)
                 Df Sum Sq Mean Sq    F value Pr(>F)
Smoke_Dur         1 8.2201  8.2201 113511.962 <2e-16 ***
Gender            1 0.0000  0.0000      0.033 0.8559
F_Mem_Smoke       3 0.0068  0.0023     31.410 <2e-16 ***
Cf_Smoke          3 0.0335  0.0112    154.425 <2e-16 ***
fWarn             1 0.0092  0.0092    126.735 <2e-16 ***
Smoke_Dur:Gender  1 0.0926  0.0926   1279.076 <2e-16 ***
Residuals      1842 0.1334  0.0001
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Addition to that the histogram of the initial model is shown below.



After the process of backward elimination, the final model's histogram is shown in the below. It is observed that the model approaches the normality.



3. Question 3

$$\hat{y} \pm t_{\frac{\alpha}{2},n-p} \sqrt{(\hat{\sigma}^2 x_0^T (x^T x)^{-1} x_0}$$

The formula for calculating the prediction interval is given above. In final model

$$\hat{y} = \mathbf{sqrt(Lung\ Capacity), smoke\ duration + Gender} + \mathbf{Number\ of\ family\ member\ smoke + Warned\ by\ family\ or\ not} + \mathbf{Gender * Smoke\ Duration}$$

These values represent the linear regression model element if values given as given in the question

Smoke Duration = 1.65

Monthly Income = 12500

Gender = Male

Family Member Smoke = 1

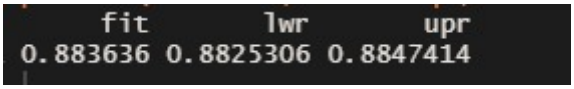
Close Friend Smoke = 2

Warned = 1

Informed By the School = 0

Anti-Tobacco messages = 1

Using these values, the for CI mean fit, lower and upper bounds shown below.



In order to obtain the real response we need to take these values square because in the model while predicting values it takes the squareroot of the responses and fits corresponding to these values. So final result shown below.

4. Question 4

$$\hat{y} \pm t_{\frac{\alpha}{2},n-p} \sqrt{(\hat{\sigma}^2(1 + x_0^T(x^T x)^{-1}x_0)}$$

The formula for calculating the prediction interval is given above. In final model

$$\hat{y} = \mathbf{sqrt(Lung\ Capacity)}, \mathbf{smoke\ duration} + \mathbf{Gender} + \mathbf{Number\ of\ family\ member\ smoke} + \mathbf{Warned\ by\ family\ or\ not} + \mathbf{Gender} * \mathbf{Smoke\ Duration}$$

These values represent the linear regression model element if values given as given in the question

Smoke Duration = 3.35

Monthly Income = 10350

Gender = Female

Family Member Smoke = two

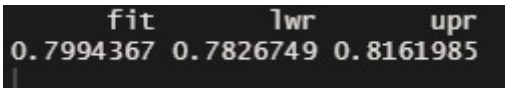
Close Friend Smoke = 3

Warned = 0

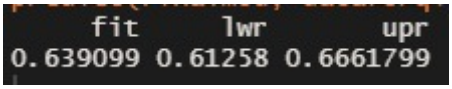
Informed By the School = 1

Anti-Tobacco messages = 1

Using these values, the predicted response mean fit, lower and upper bounds shown below.



In order to obtain the real response we need to take these values square because in the model while predicting values it takes the squareroot of the responses and fits corresponding to these values. So final result shown below.



fit	lwr	upr
0.639099	0.61258	0.6661799