Illinois Institute of Technology

MATH 561 - Algebraic & Geometric Methods in Statistics

Submitted by : Akshay Singh
Amogh Kori
Rahul Machiraju
Ram Vaka

Instructed by : Sonja Petrovic

Table of contents

Introduction

Data Description

Problem Statement

Correlation Plot

ANOVA Analysis

Pearson’s Chi-Squared Test

Fisher’s Exact Test

HC Function

BIC

RMSE

Bayesian Network Graphs

Linear Model

Gussian Model

Conclusion

Introduction:

In this project, we aim to use algebraic and geometric methods to perform data analysis on real-life data using the statistical programming language R. These methods provide powerful tools for exploring relationships between variables and making predictions based on observed data.

We will be conducting various statistical tests, including independence tests, to investigate the relationships between different variables. We will also create Bayesian network graphs to model and predict the probabilistic relationships between variables.

Moreover, we will use linear regression models to identify and quantify the relationships between the response variable and the predictor variables. Finally, we will use goodness of fit tests to evaluate the fit of the models and assess their predictive accuracy.

Our data analysis will involve a wide range of statistical techniques, and we will use the latest tools and packages available in R to carry out our analysis. By applying these techniques, we hope to gain deeper insights into the data and make informed decisions based on our findings.

Data Description:

Dataset - Lung Cancer Prediction - Initially we were working with TikTok song popularity dataset, one major issue we were facing were, a lot of columns were categorical, so it was a bit hard to do independence tests on them. Where as this dataset has most of the columns as numerical, or binary - such as smokes or not, air pollution, age etc.

Desription of the columns 📝

Age	The age of the patient. (Numeric)
Gender	The gender of the patient. (Categorical)
Air Pollution	The level of air pollution exposure of the patient. (Categorical)
Alcohol use	The level of alcohol use of the patient. (Categorical)
Dust Allergy	The level of dust allergy of the patient. (Categorical)
OccuPational Hazards	The level of occupational hazards of the patient. (Categorical)
Genetic Risk	The level of genetic risk of the patient. (Categorical)
chronic Lung Disease	The level of chronic lung disease of the patient. (Categorical)
Balanced Diet	The level of balanced diet of the patient. (Categorical)
Obesity	The level of obesity of the patient. (Categorical)
Smoking	The level of smoking of the patient. (Categorical)
Passive Smoker	The level of passive smoker of the patient. (Categorical)
Chest Pain	The level of chest pain of the patient. (Categorical)
Coughing of Blood	The level of coughing of blood of the patient. (Categorical)
Fatigue	The level of fatigue of the patient. (Categorical)
Weight Loss	The level of weight loss of the patient. (Categorical)
Shortness of Breath	The level of shortness of breath of the patient. (Categorical)
Wheezing	The level of wheezing of the patient. (Categorical)
Swallowing Difficulty	The level of swallowing difficulty of the patient. (Categorical)

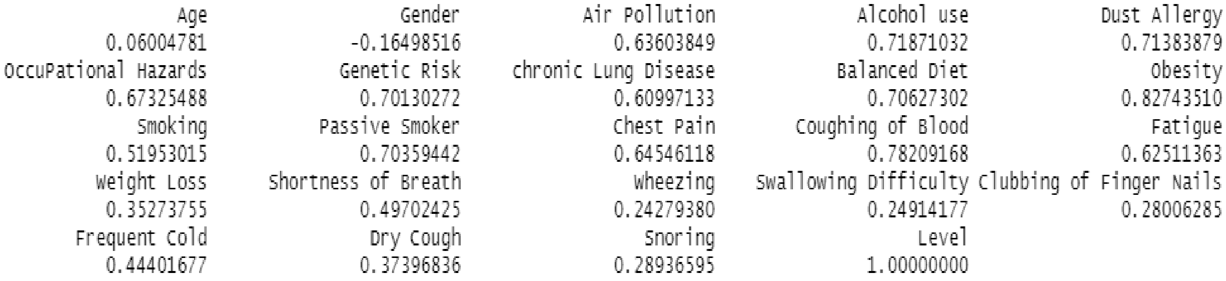
Clubbing of Finger Nails	The level of clubbing of finger nails of the patient. (Categorical)

Problem Statement:

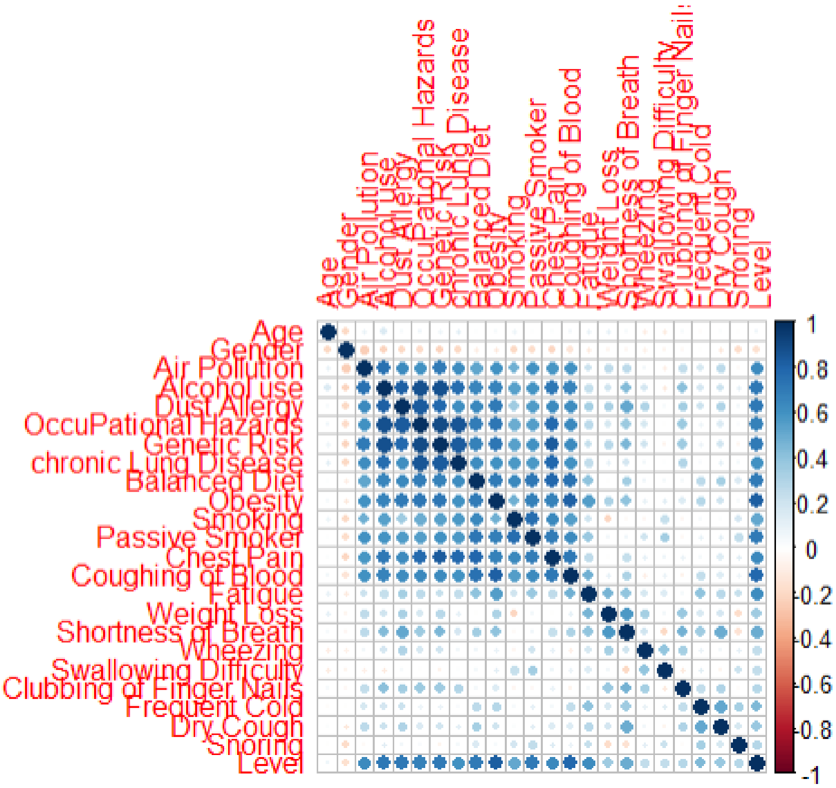
We are conducting data analysis in a real-world setting and experimenting with several models, which includes graphical, linear, and Gaussian models. We are also exploring different approaches to feature selection.

Correlation Plot:

We are drawing correlation plots here to visualize and analyze the correlation between different numerical variables in a dataset. Correlation plots provide a graphical representation of the relationship between variables and help us identify patterns, trends, and potential outliers in the data.

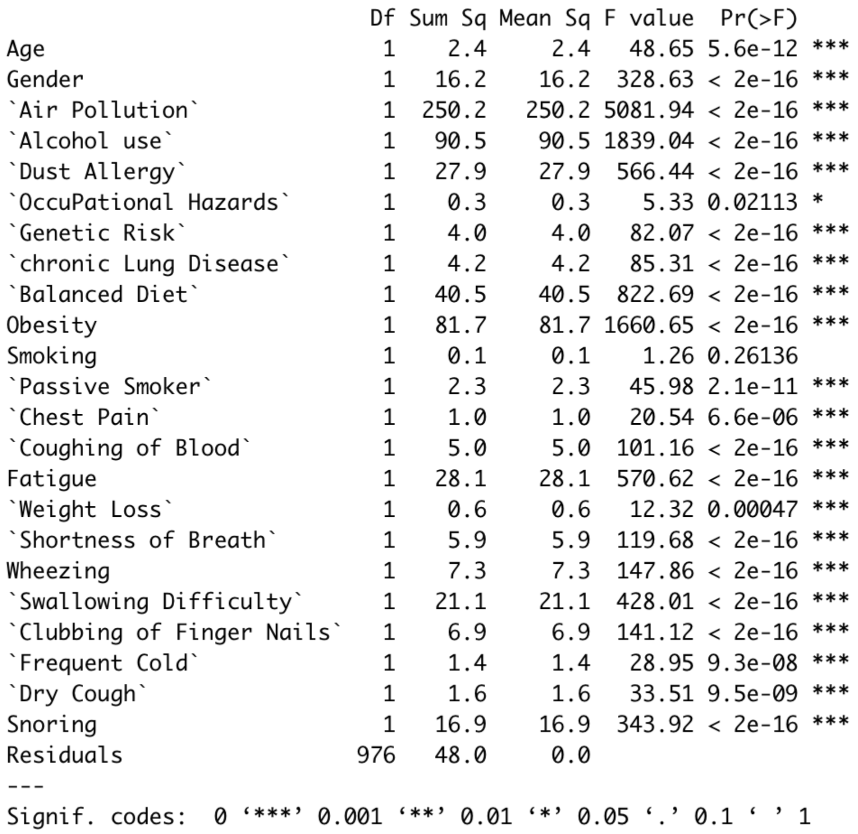


```
                     Age                   Gender            Air Pollution              Alcohol use             Dust Allergy
              0.06004781              -0.16498516               0.63603849               0.71871032               0.71383879
    OccuPational Hazards             Genetic Risk     chronic Lung Disease            Balanced Diet                  Obesity
              0.67325488               0.70130272               0.60997133               0.70627302               0.82743510
                 Smoking           Passive Smoker               Chest Pain        Coughing of Blood                  Fatigue
              0.51953015               0.70359442               0.64546118               0.78209168               0.62511363
             Weight Loss      Shortness of Breath                 Wheezing    Swallowing Difficulty Clubbing of Finger Nails
              0.35273755               0.49702425               0.24279380               0.24914177               0.28006285
           Frequent Cold                Dry Cough                  Snoring                    Level
              0.44401677               0.37396836               0.28936595               1.00000000
```



ANOVA Analysis :

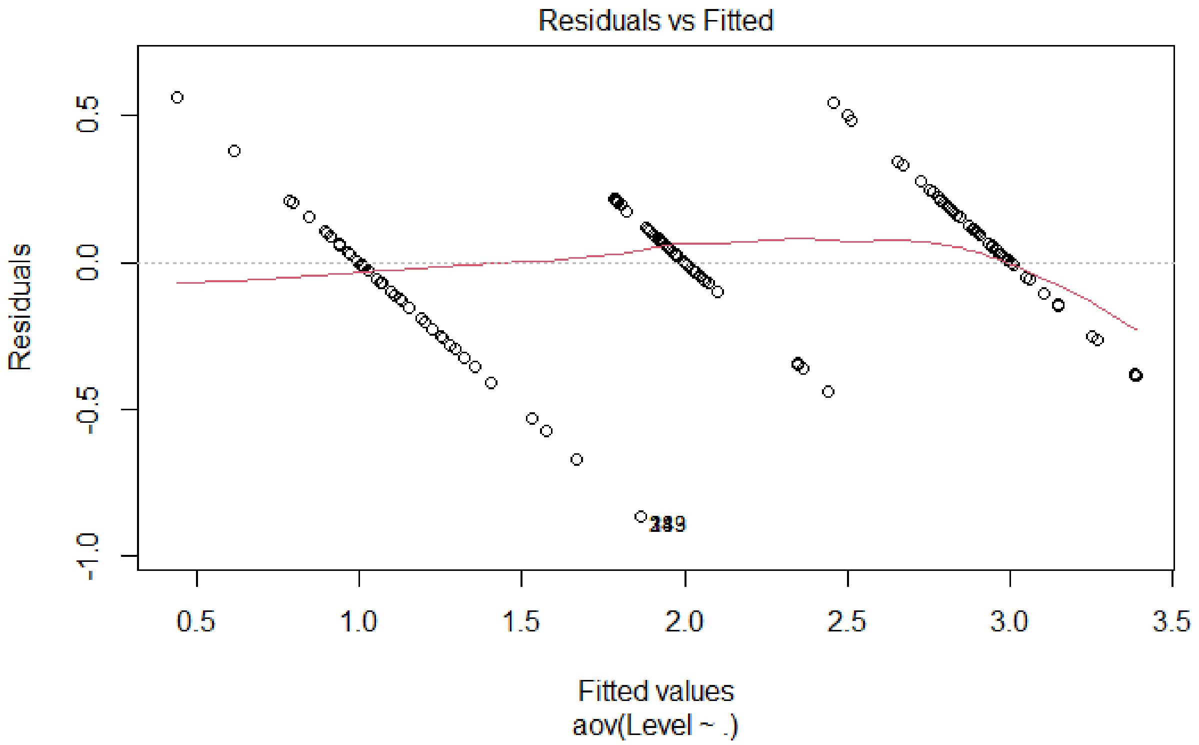
ANOVA (Analysis of Variance) is a statistical test used to determine whether there are any significant differences between the means of two or more groups. We performed a two-way ANOVA on the 'Level' variable with the remaining features in the data frame.

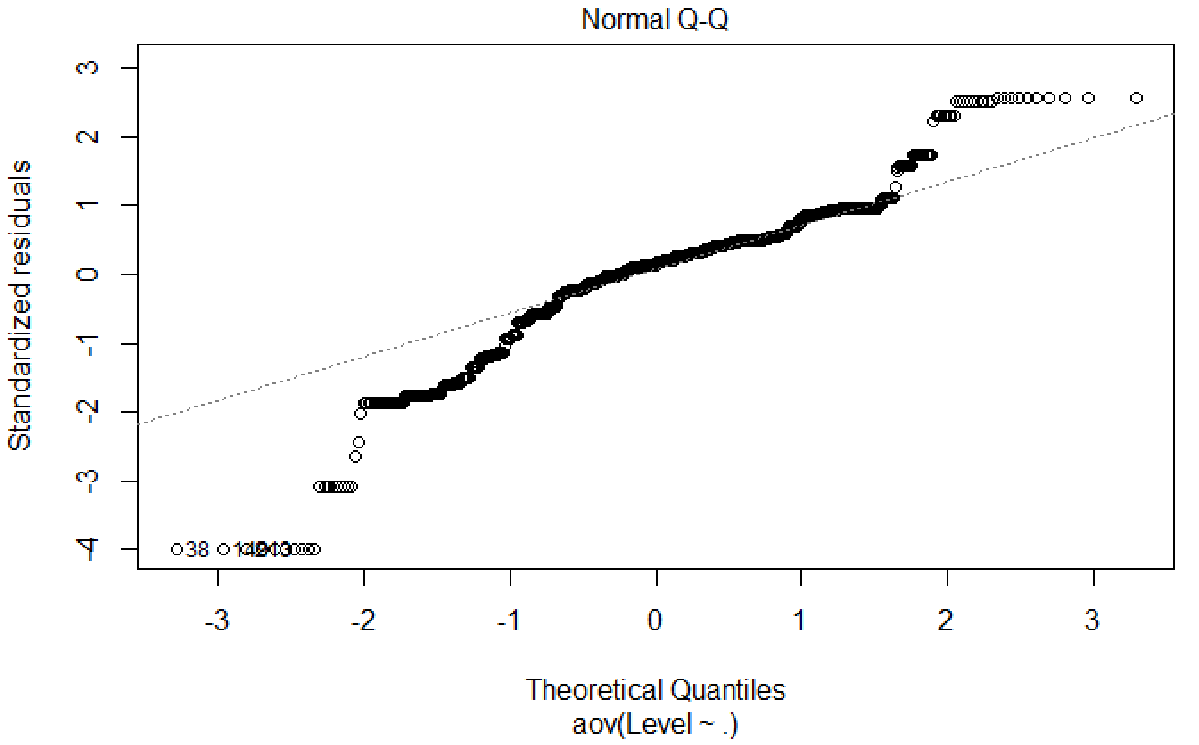


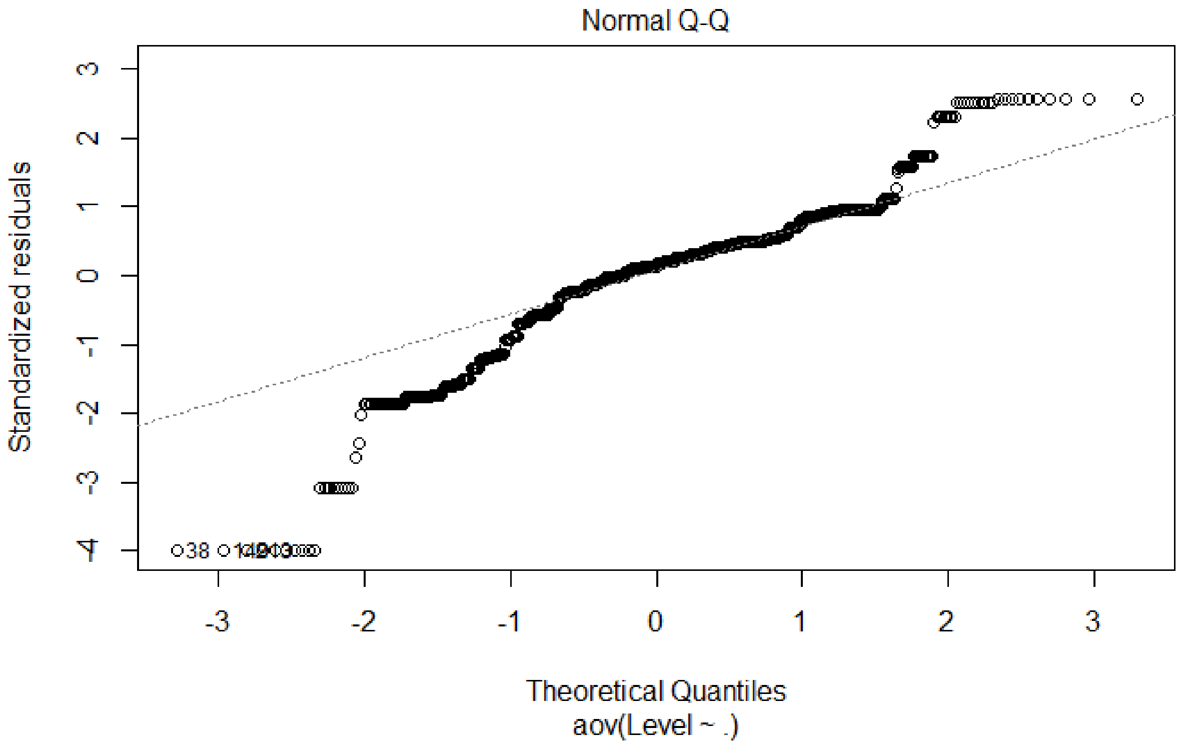
```
                          Df Sum Sq Mean Sq F value  Pr(>F)    
Age                        1    2.4     2.4   48.65 5.6e-12 ***
Gender                     1   16.2    16.2  328.63 < 2e-16 ***
`Air Pollution`            1  250.2   250.2 5081.94 < 2e-16 ***
`Alcohol use`              1   90.5    90.5 1839.04 < 2e-16 ***
`Dust Allergy`             1   27.9    27.9  566.44 < 2e-16 ***
`OccuPational Hazards`     1    0.3     0.3    5.33 0.02113 *  
`Genetic Risk`             1    4.0     4.0   82.07 < 2e-16 ***
`chronic Lung Disease`     1    4.2     4.2   85.31 < 2e-16 ***
`Balanced Diet`            1   40.5    40.5  822.69 < 2e-16 ***
Obesity                    1   81.7    81.7 1660.65 < 2e-16 ***
Smoking                    1    0.1     0.1    1.26 0.26136    
`Passive Smoker`           1    2.3     2.3   45.98 2.1e-11 ***
`Chest Pain`               1    1.0     1.0   20.54 6.6e-06 ***
`Coughing of Blood`        1    5.0     5.0  101.16 < 2e-16 ***
Fatigue                    1   28.1    28.1  570.62 < 2e-16 ***
`Weight Loss`              1    0.6     0.6   12.32 0.00047 ***
`Shortness of Breath`      1    5.9     5.9  119.68 < 2e-16 ***
Wheezing                   1    7.3     7.3  147.86 < 2e-16 ***
`Swallowing Difficulty`    1   21.1    21.1  428.01 < 2e-16 ***
`Clubbing of Finger Nails` 1    6.9     6.9  141.12 < 2e-16 ***
`Frequent Cold`            1    1.4     1.4   28.95 9.3e-08 ***
`Dry Cough`                1    1.6     1.6   33.51 9.5e-09 ***
Snoring                    1   16.9    16.9  343.92 < 2e-16 ***
Residuals                976   48.0     0.0                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

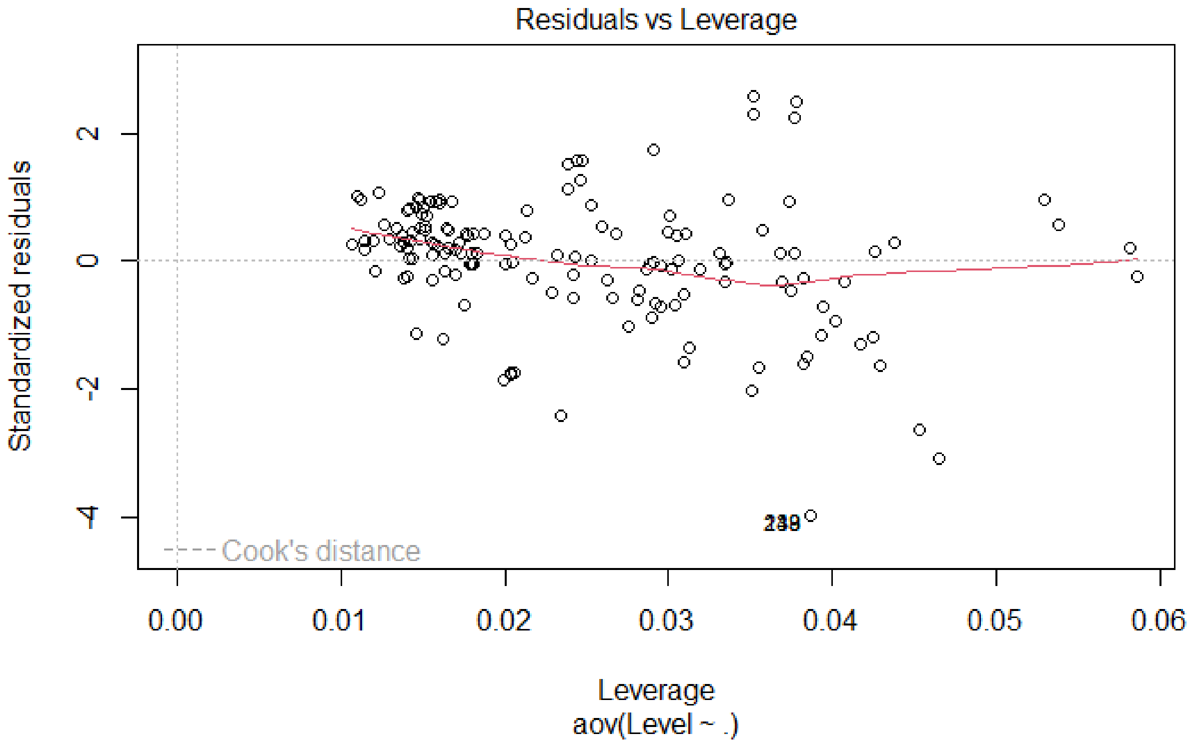
The summary includes the F-statistic, p-value, and degrees of freedom for each variable in the model. It also includes the overall model F-statistic and p-value, as well as the R-squared value, which indicates the proportion of variance explained by the model.

The plot shows the estimated means for each level of each predictor variable, along with error bars indicating the confidence interval. It also shows the interaction between the two predictor variables, if there is one.


Residuals vs Fitted
Residuals
0.5
0.0
-0.5
-1.0
289
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Fitted values
aov(Level ~ .)


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
-4
38
149
143
-3
-2
-1
0
1
2
3
Theoretical Quantiles
aov(Level ~ .)


Normal Q-Q
Standardized residuals
-4
-3
-2
-1
0
1
2
3
-3
-2
-1
0
1
2
3
38
149
13
Theoretical Quantiles
aov(Level ~ .)


Residuals vs Leverage
Standardized residuals
-4
-2
0
2
0.00
0.01
0.02
0.03
0.04
0.05
0.06
238
Cook's distance
Leverage
aov(Level ~ .)

Pearson’s Chi-Squared Test

The chi-squared test is a statistical test used to determine whether there is a significant association between two categorical variables in a contingency table. It calculates a test statistic that measures the difference between the observed frequencies in the contingency table and the frequencies that would be expected under the assumption of independence.

In the case of our example, the null hypothesis assumes that there is no significant association between the levels (of risk of cancer) and age. The alternative hypothesis assumes that there is a significant association between the two variables.

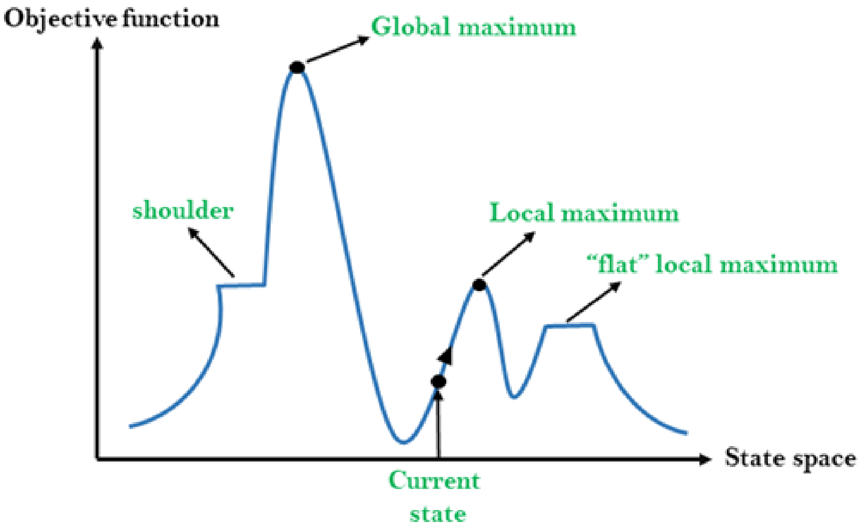
To perform the chi-squared test, we first calculate the expected frequencies for each cell in the contingency table under the assumption of independence. We then calculate the chi-squared test statistic by comparing the observed and expected frequencies. If the calculated test statistic is greater than the critical value at a given level of significance, we reject the null hypothesis and conclude that there is a significant association between the two variables.

Fisher’s Exact Test

Fisher's exact test is a statistical test that calculates the exact probability of obtaining the observed frequencies in a contingency table or a 2x2 table, given the marginal totals. The test calculates the probability of obtaining the observed frequencies and all other possible tables that are as or more extreme than the observed table, assuming the null hypothesis of independence between the two variables.

In contrast to the chi-squared test, Fisher's exact test does not rely on asymptotic assumptions and is suitable for small sample sizes or when the expected frequencies are low. The test is based on a hypergeometric distribution, which takes into account the exact number of observations in each category.

HC Function



Hill climbing is a heuristic optimization algorithm used for solving problems by iteratively improving a candidate solution in a given search space. The hill climbing algorithm works by starting with a random candidate solution and iteratively improving it by making small adjustments. At each iteration, the algorithm evaluates the new solution and compares it to the previous one. If the new solution is better, it is accepted as the new candidate solution, and the process continues. If the new solution is worse, it is rejected and the algorithm terminates.

The hc function in R implements the hill-climbing algorithm for optimizing the parameters of a model. The function takes as input an initial set of parameter values, an objective function to be optimized, and a set of constraints on the parameter values. The hc function updates the parameters iteratively by trying out different parameter values in the vicinity of the current value and selecting the one that results in the best objective function value. The algorithm stops when it reaches a local optimum, which is a set of parameter values that cannot be improved further. The hc function returns an object of class "optim," which contains information about the optimization process, such as the final parameter values, the objective function value, and the convergence status.

BIC - Bayesian Information Criterion

BIC is a statistical measure used to compare different statistical models based on how well they fit the data while also taking into account the complexity of the models.

The BIC score is a numerical value that reflects the relative quality of a statistical model compared to other models. The BIC score is calculated using the following formula:

BIC = -2 * log(L) + k * log(n)

where L is the maximum likelihood estimate of the model, k is the number of parameters in the model, and n is the sample size.

The lower the BIC score is, the better our model is.

BIC is commonly used in fields such as statistics, and machine learning for model selection. It is especially useful when comparing models with different numbers of parameters or when the sample size is relatively small.

RMSE - Room Mean Squared Error

RMSE stands for Root Mean Squared Error, which is a statistical measure used to evaluate the accuracy of a prediction model or estimator. RMSE measures the difference between the predicted values and the actual values of the variable being predicted, and it is expressed in the same units as the variable.

The RMSE is calculated by taking the square root of the average of the squared differences between the predicted and actual values. The formula for calculating RMSE is:

RMSE = sqrt(1/n * sum((y_pred - y_actual)^2))

where n is the number of observations, y_pred is the predicted value, and y_actual is the actual value.

The lower the RMSE is, the better our model is.

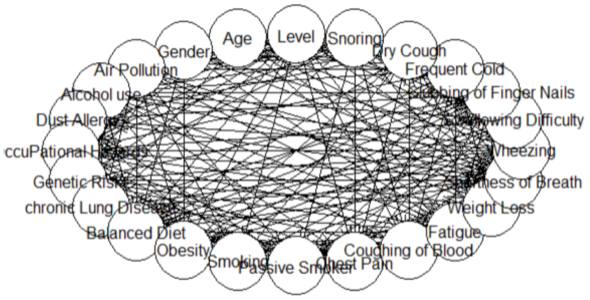
Note: We will be using BIC and RMSE scores throughout our code to evaluate different models.

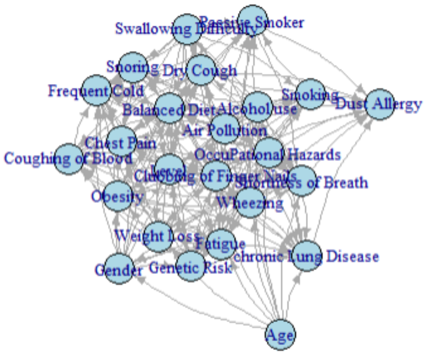
Model 1 : Graphical Model

A graphical model is a statistical model that represents the conditional dependencies between variables using a graph. In the context of the lung cancer prediction dataset, a graphical model could be used to visualize the relationships between different variables, such as smoking habits, age, and cancer diagnosis. This model can provide a comprehensive understanding of how different variables are related to each other and can help identify which variables are most predictive of lung cancer. Additionally, a graphical model can be used to perform causal inference, allowing us to determine the causal relationships between variables and better understand the underlying mechanisms driving the development of lung cancer.

A. For Entire Dataset

The summary of a Bayesian network learned via score-based methods. The model consists of 24 nodes and 186 arcs, with an average Markov blanket size of 21.75 and an average neighborhood size of 15.50. The average branching factor is 7.75, indicating that each node has an average of 7.75 children in the network. The learning algorithm used was hill-climbing, and the score used was BIC (Gaussian). The penalization coefficient used was 3.453878, and 5060 tests were used in the learning procedure. The model is optimized, indicating that the learning algorithm has found the best structure and parameters for the model given the data. The model describes relationships between various health-related variables such as age, wheezing, gender, air pollution, alcohol use, swallowing difficulty, snoring, occupational hazards, genetic risk, obesity, level, fatigue, weight loss, smoking, coughing of blood, passive smoker, dry cough, frequent cold, chest pain, balanced diet, shortness of breath, chronic lung disease, clubbing of finger nails, and dust allergy.





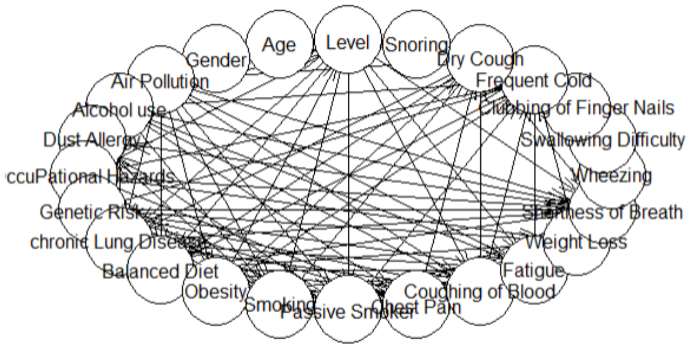
In analyzing a Bayesian network with 10 nodes, it appears that the nodes are being modeled through the use of Gaussian distributions. These nodes (Parameters) include Age, Gender, Air Pollution, Alcohol use, Dust Allergy, OccuPational Hazards, Genetic Risk, chronic Lung Disease, Balanced Diet, and Level. The parameters for each node consist of the coefficients of the Gaussian distribution, as well as the standard deviation of the residuals. In addition to these parameters, each node has conditioning variables for the Gaussian distributions that are specific to each individual node. For example, the node Gender has conditioning variables Age and Wheezing, whereas the node Alcohol use has conditioning variables Age, Air Pollution, and Wheezing.

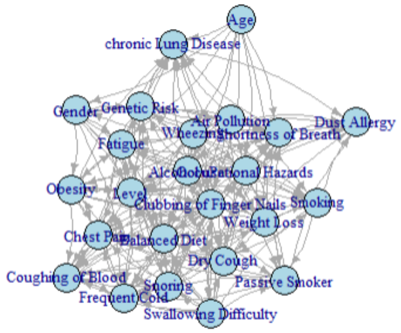
The BIC score of the graphical model for entire dataset is -39899.92

B. Graphical Model for Marginalized data

We selected a subset of columns from our data frame based on the correlation with the 'Level' variable. Specifically, we used a threshold of 0.3 to select only columns with an absolute correlation value below this threshold.

The Bayesian network learned via score-based methods consists of 24 nodes and 105 arcs, with an average markov blanket size of 11.75 and an average neighbourhood size of 8.75. The learning algorithm used was Hill-Climbing, with BIC (Gauss.) as the score and a penalization coefficient of 3.453878. The procedure used 2921 tests and was optimized. The network includes nodes such as Age, Gender, Air Pollution, Wheezing, Smoking, and Balanced Diet, and their relationships are modeled using directed arcs. The average branching factor is 4.38. These results suggest that the network is complex and may have some degree of predictive power for the data.



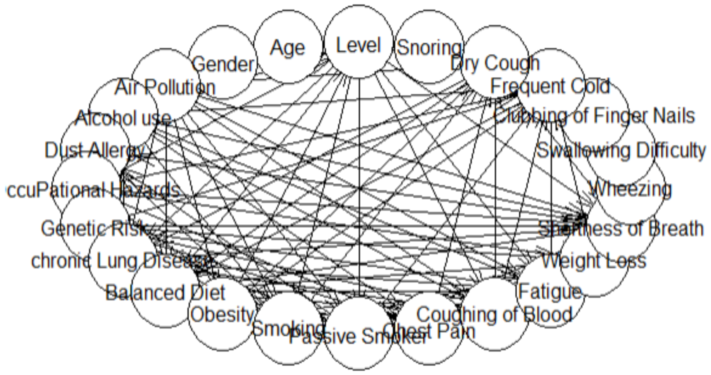


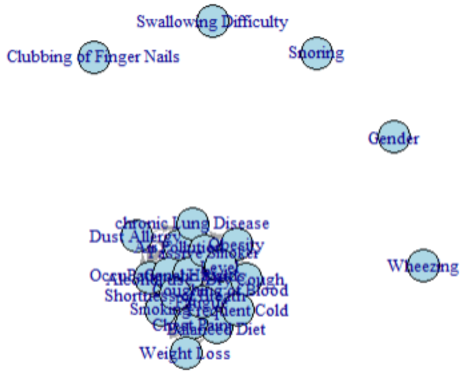
The conditional density and coefficients are provided for each node, along with the standard deviation of the residuals.

The BIC score of the graphical model for entire dataset is infinity.

C. Graphical Model for selected features from ANOVA analysis

The model consists of 22 nodes and 140 arcs, with an average Markov blanket size of 19.55, average neighborhood size of 12.73, and average branching factor of 6.36. The nodes in the model represent variables related to lung cancer, such as age, smoking habits, genetics, and symptoms. The learning algorithm used a penalization coefficient of 3.45, and the model was optimized. A total of 3,297 tests were used in the learning procedure. Overall, this model provides a graphical representation of the conditional dependencies between different variables related to lung cancer, which can be used to better understand the underlying mechanisms driving the development of the disease.





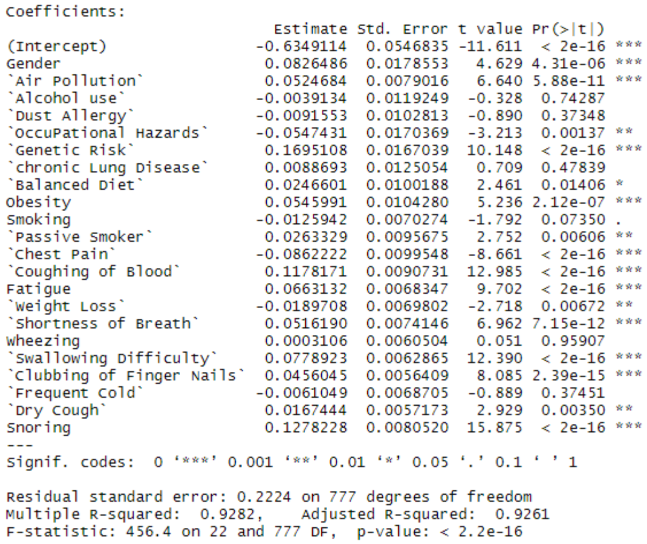
The BIC score for this model is -37552

Model 2 : Linear Model

A linear model is a statistical approach used to describe the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, where a change in the independent variable is directly proportional to a change in the dependent variable. Linear models are commonly used in various fields such as economics, finance, and biology to make predictions and understand the underlying relationships between variables.

For the lung cancer prediction dataset, we are using a few variations of the linear model. The first one uses the entire dataset, where all variables are included in the model. The second model uses marginalized data, which means that some variables are excluded from the model to reduce their impact on the outcome. The third model uses analysis of variance (ANOVA), which is a statistical technique used to determine the differences between groups of variables. By using these three variations of the linear model, we can determine which model best predicts the likelihood of lung cancer in patients based on the available data.

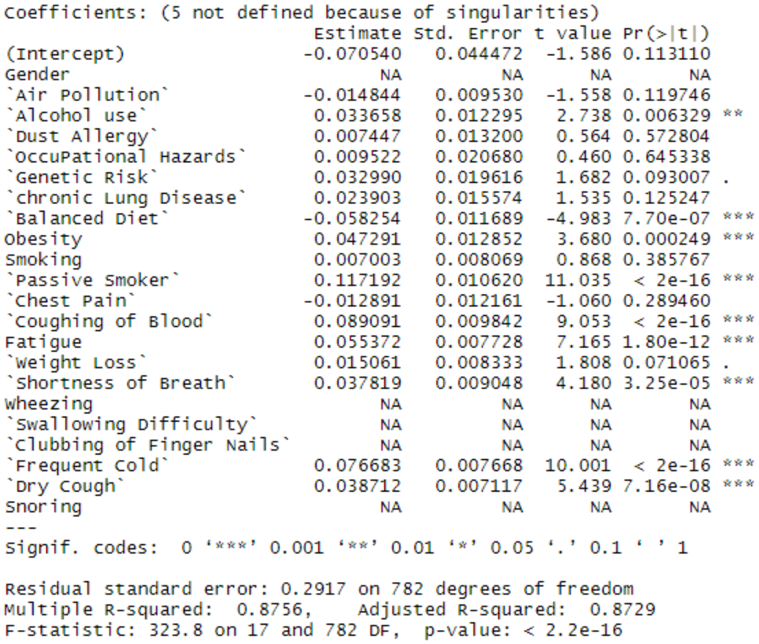
A. For Entire Dataset



```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                -0.6349114  0.0546835 -11.611  < 2e-16 ***
Gender                      0.0826486  0.0178553   4.629 4.31e-06 ***
`Air Pollution`             0.0524684  0.0079016   6.640 5.88e-11 ***
`Alcohol use`              -0.0039134  0.0119249  -0.328  0.74287    
`Dust Allergy`             -0.0091553  0.0102813  -0.890  0.37348    
`OccuPational Hazards`     -0.0547431  0.0170369  -3.213  0.00137 ** 
`Genetic Risk`              0.1695108  0.0167039  10.148  < 2e-16 ***
`chronic Lung Disease`      0.0088693  0.0125054   0.709  0.47839    
`Balanced Diet`             0.0246601  0.0100188   2.461  0.01406 *  
Obesity                     0.0545991  0.0104280   5.236 2.12e-07 ***
Smoking                    -0.0125942  0.0070274  -1.792  0.07350 .  
`Passive Smoker`            0.0263329  0.0095675   2.752  0.00606 ** 
`Chest Pain`               -0.0862222  0.0099548  -8.661  < 2e-16 ***
`Coughing of Blood`         0.1178171  0.0090731  12.985  < 2e-16 ***
Fatigue                     0.0663132  0.0068347   9.702  < 2e-16 ***
`Weight Loss`              -0.0189708  0.0069802  -2.718  0.00672 ** 
`Shortness of Breath`       0.0516190  0.0074146   6.962 7.15e-12 ***
Wheezing                    0.0003106  0.0060504   0.051  0.95907    
`Swallowing Difficulty`     0.0778923  0.0062865  12.390  < 2e-16 ***
`Clubbing of Finger Nails`  0.0456045  0.0056409   8.085 2.39e-15 ***
`Frequent Cold`            -0.0061049  0.0068705  -0.889  0.37451    
`Dry Cough`                 0.0167444  0.0057173   2.929  0.00350 ** 
Snoring                     0.1278228  0.0080520  15.875  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2224 on 777 degrees of freedom
Multiple R-squared:  0.9282,	Adjusted R-squared:  0.9261 
F-statistic: 456.4 on 22 and 777 DF,  p-value: < 2.2e-16
```

RMSE : 0.2233527

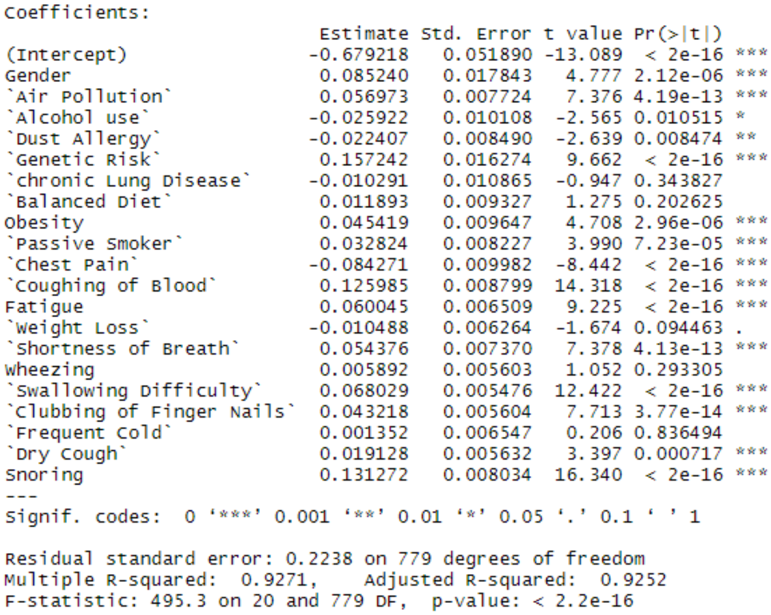
B. For Marginalized Data



```
Coefficients: (5 not defined because of singularities)
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               -0.070540   0.044472  -1.586 0.113110    
Gender                           NA         NA      NA       NA    
`Air Pollution`           -0.014844   0.009530  -1.558 0.119746    
`Alcohol use`              0.033658   0.012295   2.738 0.006329 ** 
`Dust Allergy`             0.007447   0.013200   0.564 0.572804    
`OccuPational Hazards`     0.009522   0.020680   0.460 0.645338    
`Genetic Risk`             0.032990   0.019616   1.682 0.093007 .  
`chronic Lung Disease`     0.023903   0.015574   1.535 0.125247    
`Balanced Diet`           -0.058254   0.011689  -4.983 7.70e-07 ***
Obesity                    0.047291   0.012852   3.680 0.000249 ***
Smoking                    0.007003   0.008069   0.868 0.385767    
`Passive Smoker`           0.117192   0.010620  11.035  < 2e-16 ***
`Chest Pain`              -0.012891   0.012161  -1.060 0.289460    
`Coughing of Blood`        0.089091   0.009842   9.053  < 2e-16 ***
Fatigue                    0.055372   0.007728   7.165 1.80e-12 ***
`Weight Loss`              0.015061   0.008333   1.808 0.071065 .  
`Shortness of Breath`      0.037819   0.009048   4.180 3.25e-05 ***
Wheezing                         NA         NA      NA       NA    
`Swallowing Difficulty`          NA         NA      NA       NA    
`Clubbing of Finger Nails`       NA         NA      NA       NA    
`Frequent Cold`            0.076683   0.007668  10.001  < 2e-16 ***
`Dry Cough`                0.038712   0.007117   5.439 7.16e-08 ***
Snoring                          NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2917 on 782 degrees of freedom
Multiple R-squared:  0.8756,	Adjusted R-squared:  0.8729 
F-statistic: 323.8 on 17 and 782 DF,  p-value: < 2.2e-16
```

RMSE : 0.314533

C. For ANOVA

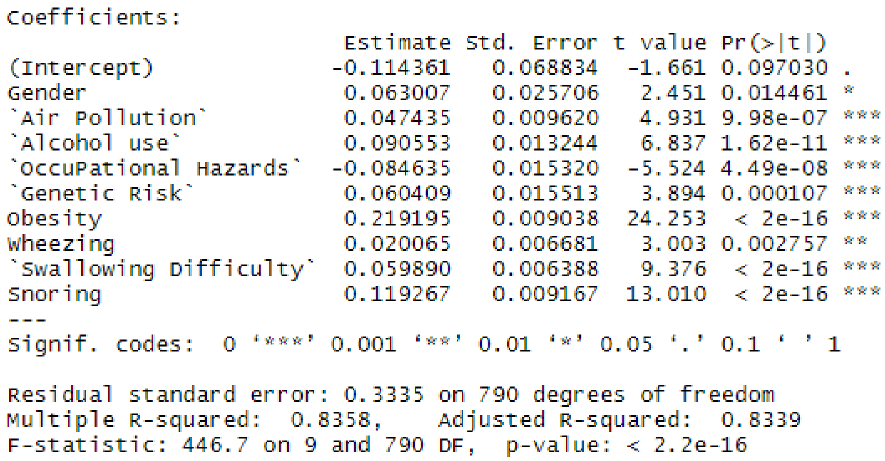


```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                -0.679218   0.051890 -13.089  < 2e-16 ***
Gender                      0.085240   0.017843   4.777 2.12e-06 ***
`Air Pollution`             0.056973   0.007724   7.376 4.19e-13 ***
`Alcohol use`              -0.025922   0.010108  -2.565 0.010515 *  
`Dust Allergy`             -0.022407   0.008490  -2.639 0.008474 ** 
`Genetic Risk`              0.157242   0.016274   9.662  < 2e-16 ***
`chronic Lung Disease`     -0.010291   0.010865  -0.947 0.343827    
`Balanced Diet`             0.011893   0.009327   1.275 0.202625    
Obesity                     0.045419   0.009647   4.708 2.96e-06 ***
`Passive Smoker`            0.032824   0.008227   3.990 7.23e-05 ***
`Chest Pain`               -0.084271   0.009982  -8.442  < 2e-16 ***
`Coughing of Blood`         0.125985   0.008799  14.318  < 2e-16 ***
Fatigue                     0.060045   0.006509   9.225  < 2e-16 ***
`Weight Loss`              -0.010488   0.006264  -1.674 0.094463 .  
`Shortness of Breath`       0.054376   0.007370   7.378 4.13e-13 ***
Wheezing                    0.005892   0.005603   1.052 0.293305    
`Swallowing Difficulty`     0.068029   0.005476  12.422  < 2e-16 ***
`Clubbing of Finger Nails`  0.043218   0.005604   7.713 3.77e-14 ***
`Frequent Cold`             0.001352   0.006547   0.206 0.836494    
`Dry Cough`                 0.019128   0.005632   3.397 0.000717 ***
Snoring                     0.131272   0.008034  16.340  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2238 on 779 degrees of freedom
Multiple R-squared:  0.9271,	Adjusted R-squared:  0.9252 
F-statistic: 495.3 on 20 and 779 DF,  p-value: < 2.2e-16
```

RMSE : 0.227054

In addition to the three variations of the linear model mentioned earlier, we are also using two more variations. The first one uses the model from the hc-1 dataset for the entire lung cancer prediction dataset, and the second one uses the hc-1 model for the marginalized dataset. The hc-1 model is a commonly used linear regression model that assumes that the errors in the data are independent and identically distributed. By incorporating these two additional variations, we can assess the performance of the hc-1 model on our lung cancer prediction dataset and compare it to the other linear models used. This will provide us with a more comprehensive understanding of the predictive power of different linear models on the dataset.

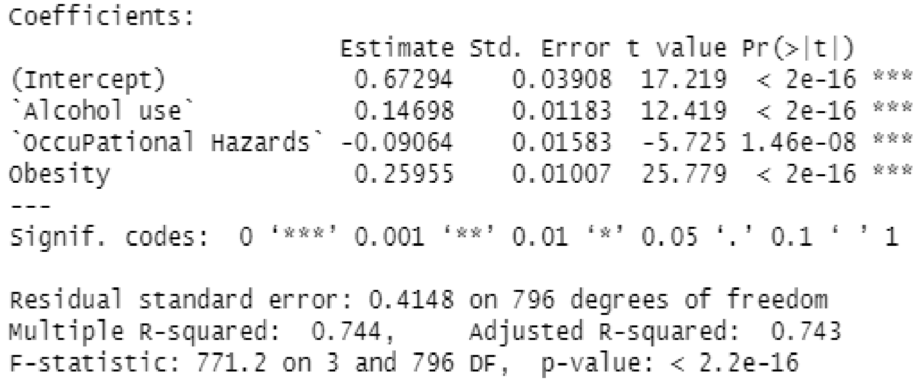
D. Model From HC-1 For Entire Dataset



```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)             -0.114361   0.068834  -1.661 0.097030 .  
Gender                   0.063007   0.025706   2.451 0.014461 *  
`Air Pollution`          0.047435   0.009620   4.931 9.98e-07 ***
`Alcohol use`            0.090553   0.013244   6.837 1.62e-11 ***
`OccuPational Hazards`  -0.084635   0.015320  -5.524 4.49e-08 ***
`Genetic Risk`           0.060409   0.015513   3.894 0.000107 ***
Obesity                  0.219195   0.009038  24.253  < 2e-16 ***
Wheezing                 0.020065   0.006681   3.003 0.002757 ** 
`Swallowing Difficulty`  0.059890   0.006388   9.376  < 2e-16 ***
Snoring                  0.119267   0.009167  13.010  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3335 on 790 degrees of freedom
Multiple R-squared:  0.8358,	Adjusted R-squared:  0.8339 
F-statistic: 446.7 on 9 and 790 DF,  p-value: < 2.2e-16
```

RMSE : 0.3409274

E. Model From HC-1 For Marginal Dataset



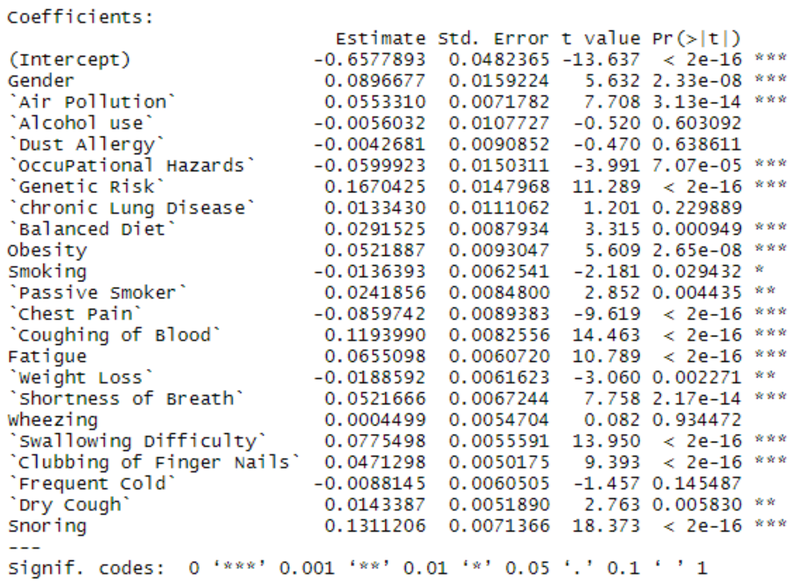
```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)             0.67294    0.03908  17.219  < 2e-16 ***
`Alcohol use`           0.14698    0.01183  12.419  < 2e-16 ***
`OccuPational Hazards` -0.09064    0.01583  -5.725 1.46e-08 ***
Obesity                 0.25955    0.01007  25.779  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4148 on 796 degrees of freedom
Multiple R-squared:  0.744,	Adjusted R-squared:  0.743 
F-statistic: 771.2 on 3 and 796 DF,  p-value: < 2.2e-16
```

RMSE : 0.4106685

Model 3 : Gaussian Model

A Gaussian model, also known as a normal distribution model, is a statistical approach used to describe the distribution of a continuous variable. In the context of the lung cancer prediction dataset, a Gaussian model could be used to model the distribution of a particular variable, such as the age of patients or the number of cigarettes smoked per day. This model assumes that the variable is normally distributed, with a bell-shaped curve and a symmetrical distribution around the mean. By fitting a Gaussian model to the data, we can estimate the mean and standard deviation of the variable and make predictions based on the likelihood of different values occurring within the distribution.

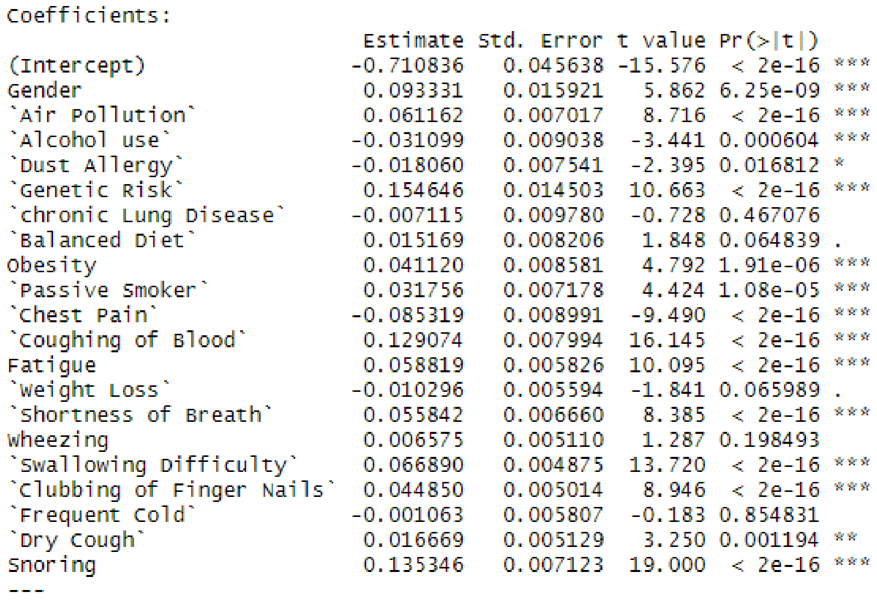
A. Model For Entire Dataset



```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 -0.6577893  0.0482365 -13.637  < 2e-16 ***
Gender                       0.0896677  0.0159224   5.632 2.33e-08 ***
`Air Pollution`              0.0553310  0.0071782   7.708 3.13e-14 ***
`Alcohol use`               -0.0056032  0.0107727  -0.520 0.603092    
`Dust Allergy`              -0.0042681  0.0090852  -0.470 0.638611    
`OccuPational Hazards`      -0.0599923  0.0150311  -3.991 7.07e-05 ***
`Genetic Risk`               0.1670425  0.0147968  11.289  < 2e-16 ***
`chronic Lung Disease`       0.0133430  0.0111062   1.201 0.229889    
`Balanced Diet`              0.0291525  0.0087934   3.315 0.000949 ***
Obesity                      0.0521887  0.0093047   5.609 2.65e-08 ***
Smoking                     -0.0136393  0.0062541  -2.181 0.029432 *  
`Passive Smoker`             0.0241856  0.0084800   2.852 0.004435 ** 
`Chest Pain`                -0.0859742  0.0089383  -9.619  < 2e-16 ***
`Coughing of Blood`          0.1193990  0.0082556  14.463  < 2e-16 ***
Fatigue                      0.0655098  0.0060720  10.789  < 2e-16 ***
`Weight Loss`               -0.0188592  0.0061623  -3.060 0.002271 ** 
`Shortness of Breath`        0.0521666  0.0067244   7.758 2.17e-14 ***
Wheezing                     0.0004499  0.0054704   0.082 0.934472    
`Swallowing Difficulty`      0.0775498  0.0055591  13.950  < 2e-16 ***
`Clubbing of Finger Nails`   0.0471298  0.0050175   9.393  < 2e-16 ***
`Frequent Cold`             -0.0088145  0.0060505  -1.457 0.145487    
`Dry Cough`                  0.0143387  0.0051890   2.763 0.005830 ** 
Snoring                      0.1311206  0.0071366  18.373  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

RMSE : 0.2212494

B. Model For ANOVA



```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                -0.710836   0.045638 -15.576  < 2e-16 ***
Gender                      0.093331   0.015921   5.862 6.25e-09 ***
`Air Pollution`             0.061162   0.007017   8.716  < 2e-16 ***
`Alcohol use`              -0.031099   0.009038  -3.441 0.000604 ***
`Dust Allergy`             -0.018060   0.007541  -2.395 0.016812 *  
`Genetic Risk`              0.154646   0.014503  10.663  < 2e-16 ***
`chronic Lung Disease`     -0.007115   0.009780  -0.728 0.467076    
`Balanced Diet`             0.015169   0.008206   1.848 0.064839 .  
Obesity                     0.041120   0.008581   4.792 1.91e-06 ***
`Passive Smoker`            0.031756   0.007178   4.424 1.08e-05 ***
`Chest Pain`               -0.085319   0.008991  -9.490  < 2e-16 ***
`Coughing of Blood`         0.129074   0.007994  16.145  < 2e-16 ***
Fatigue                     0.058819   0.005826  10.095  < 2e-16 ***
`Weight Loss`              -0.010296   0.005594  -1.841 0.065989 .  
`Shortness of Breath`       0.055842   0.006660   8.385  < 2e-16 ***
Wheezing                    0.006575   0.005110   1.287 0.198493    
`Swallowing Difficulty`     0.066890   0.004875  13.720  < 2e-16 ***
`Clubbing of Finger Nails`  0.044850   0.005014   8.946  < 2e-16 ***
`Frequent Cold`            -0.001063   0.005807  -0.183 0.854831    
`Dry Cough`                 0.016669   0.005129   3.250 0.001194 ** 
Snoring                     0.135346   0.007123  19.000  < 2e-16 ***
---
```

RMSE : 0.2247093

Conclusion

The project proposes various models to predict the RMSE scores based on the features selected for the various models and finally deciding which performs better among all. Initially, The two approaches used to select the features are based on Correlation and ANOVA analysis and then after identifying the features to drop we marginalize those features and add them back into the model and using the various combinations of features identified we build various models namely Graphical Model, Linear Model and Gaussian model and finally we conclude that using the features from the ANOVA analysis was the best approach as it helped in reducing the overall computational complexity of the models and a low RMSE value of 0.22.