Universitat Autònoma de Barcelona

Anàlisi de dades complexes

LUNG CANCER PREDICTION

Nora Alquézar Presta

Resum

Aquest projecte té com a objectiu predir el grau de tumors pulmonars mitjançant l'anàlisi de l'historial mèdic, la predisposició genètica i l'estil de vida d'un pacient. Utilitzant tècniques de bootstrap paramètric i no paramètric, s'explorarà la relació entre aquestes variables i les característiques dels tumors per desenvolupar un marc predictiu per als pacients amb càncer de pulmó.

Resumen

Este proyecto tiene como objetivo predecir el grado de tumores pulmonares mediante el análisis del historial médico, la predisposición genética y el estilo de vida de un paciente. Utilizando técnicas de bootstrap paramétrico y no paramétrico, se explorará la relación entre estas variables y las características de los tumores con el fin de desarrollar un marco predictivo para pacientes con cáncer de pulmón.

Abstract

This project aims to predict lung tumor grade by analyzing medical history, genetic predisposition, and lifestyle factors of a patient. Utilizing parametric and non-parametric bootstrap techniques, the relationship between these variables and tumor characteristics will be explored in order to develop a predictive framework for lung cancer patients.

Contents

1 Introduction

Lung cancer is the leading cause of cancer death worldwide and its impact is significant in Spain as well. In 2023, there were an estimated 22,712 deaths from lung cancer in the country. This high mortality rate is primarily attributed to smoking, which remains the main risk factor. However, exposure to air pollution and genetic predispositions also contribute significantly to the incidence of lung cancer.

All the factors mentioned before requiere an accurate diagnosis and timely treatment. With advancements in medical research and technology, there is a growing interest in utilizing data analysis techniques to predict characteristics of lung diseases. That's why this project aims to contribute to this field by exploring the relationship between medical history, genetic predisposition, lifestyle factors, and lung disease characteristics, in order to develop a predictive framework for patients affected by this disease using parametric and non-parametric bootstrap techniques.

Two models will be analyzed to examine the specified response variable of interest along with their respective independent variables. Each model will utilize a different response variable and will include the most appropriate predictor variables for that specific response.

First, we will focus on the regression model based on the patient's age at the time of lung cancer detection. This is important because the age at which cancer is detected can influence treatment and disease management. We're investigating how factors such as genetic risk and respiratory issues like the shortness of breath, snoring or chest pain may affect.

- **Age regression model:**
  Response variable: age of the patient at the time of tumor detection.
  Predictor variables: genetic risk, chest pain, snoring and shortness of breath.

Secondly, the lung cancer stage regression model will assist in predicting the stage of cancer, whether it's in its early or advanced stages. This is crucial as it significantly impacts treatment options and patient prognosis. We're exploring how factors such as smoking, air pollution, genetic predisposition and alcohol consumption may correlate with cancer progression.

- **Lung cancer stage regression model:**
  Response variable: lung cancer level, which ranges from low to high.
  Predictor variables: smoking, air pollution, genetic risk and alcohol use.

These two regression models are essential for helping us better understand lung cancer, predict its progression and identify factors influencing its detection. With this information, we hope to improve methods for diagnosing, treating, and preventing lung cancer, which could make a significant difference in the lives of patients and their families.

2 Dataset analysis

For this project we will use the dataset Lung cancer prediction sourced from Kaggle. It consists of 999 rows, and 25 columns. Each row stores one patient's individual characteristics, lifestyle habits, genetic dispositions, environmental factors and symptoms, and each column stores one of the mentioned factors.

To begin with, there are three individual characteristics:

- *Patient Id*: Unique numeric identification of each patient (Numeric).
- *Age*: Age of the patient, ranging from 14 to 73 years (Numeric).
- *Gender*: Gender of the patient, female (0) or male (1) (Categorical).
- *Level*: Severity level of the patient's cancer which can be either low, medium or high (Categorical).

Secondly, the dataset holds the lifestyle habits of the patient, which include:

- *Alcohol use*: Level of alcohol consumption by the patient (Categorical).
- *Smoking*: this column shows whether the patient drinks alcohol where 1 is low and 8 high (Categorical).
- *Passive Smoker*: Exposure to passive smoking (Categorical).
- *Balanced Diet*: Quality of the patient's balanced diet (Categorical).
- *Obesity*: Level of obesity in the patient (Categorical).

Following, the genetic disposition variables of each patient:

- *Genetic Risk*: Tells us if any patient's relative has suffered from a cancer (Categorical).
- *Chronic Lung Disease*: Presence of chronic lung disease (Categorical).

Next, the environmental factors:

- *Air Pollution*: The level of air pollution exposure of the patient (Categorical).
- *Dust Allergy*: Indicates if the patient has dust allergy or not (Categorical).
- *Occupational Hazards*: Risks related to the patient's occupation (Categorical).

Finally, the symptoms of the patient:

- *Chest Pain, coughing of blood, fatigue, weight loss, shortness of breath, wheezing, swallowing difficulty, clubbing of finger nails, frequent cold, dry cough, snoring*

3 Procedures used to analyse the data

3.1 Parametric analysis

3.1.a Age regression model

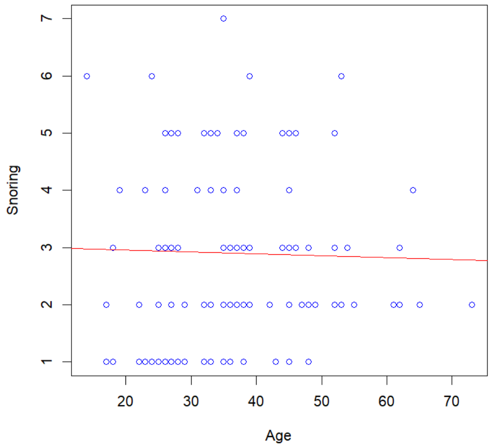
First, our focus is on the age of the patient when the cancer is detected. In order to predict at what age a patient may first exhibit symptoms.

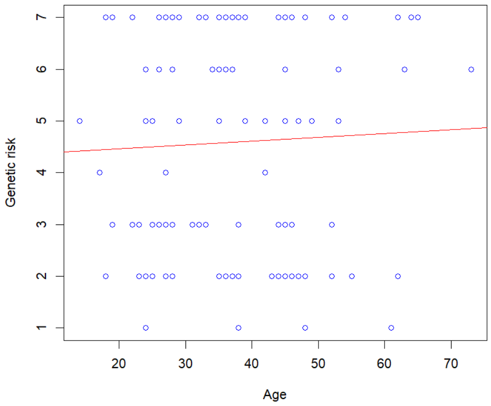
The independent variables which will be analysed will be:

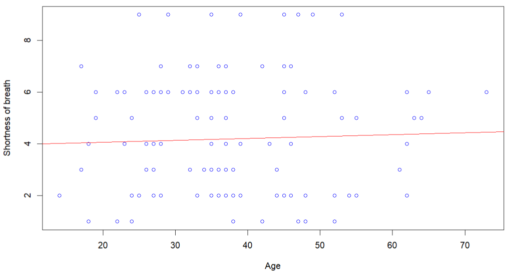
- Genetic risk
- Chest pain
- Snoring
- Shortness of breath

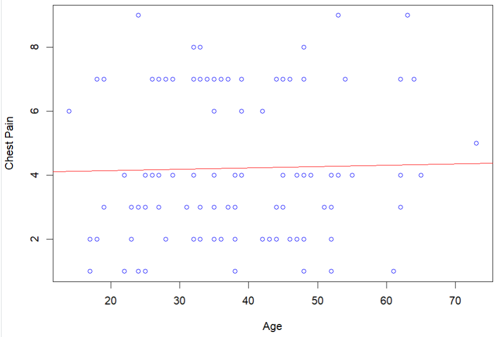
Since all the values in these columns are numeric, it makes the analysis of our parameters easier.

We analyse the parameters by plotting each of the four variables against the patient's age at the time of cancer detection and the regression line between these two.









We observe that, as a person grows older, there is typically an increasing presence of family cancer (genetic risk).

When considering snoring, the plots indicate that it can be detected at any age and persist throughout a person's life.

In addition, it's clear that, although not very pronounced, there is a positive correlation between shortness of breath and chest pain. This suggests that over time, there's a higher chance of cancer becoming more severe and symptoms worsening.

3.1.b Lung cancer stage regression model

The feature we are focusing on is the grade of the lung cancer. The values of which can be: "Low", "Medium" or "High".
To facilitate the analysis of our parameters, we convert the values to numeric format, assigning the following numbers:

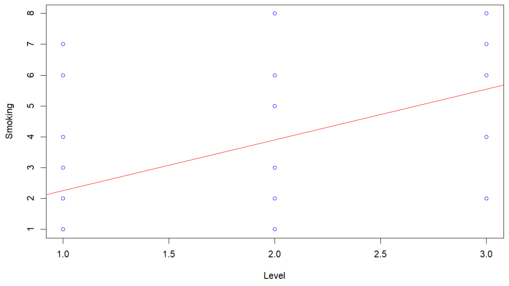
- Low -> 1
- Medium -> 2
- High -> 3

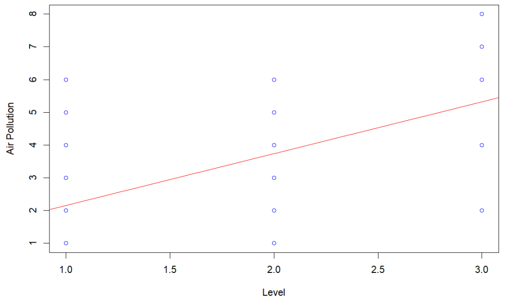
This characteristic will be studied analysing various symptoms of the patient. The independent variables which will be analysed will be:

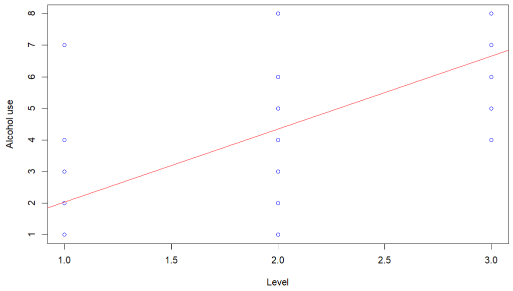
- Air pollution
- Genetic risk
- Smoking
- Alcohol use

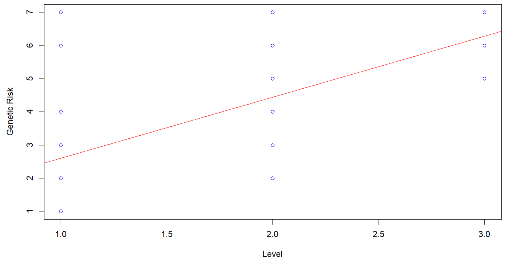
Just like before, since all the values in these columns are numeric, it makes the analysis of our parameters easier.

We analyse the parameters by plotting each of the four given symptoms against the patient's cancer stage and the regression line between these two.









As observed in the plots, all graphs exhibit nearly identical behavior and have a directly proportional relationship between the severity of cancer and the symptoms. This means that typically, higher level indicates a more aggressive cancer with increasingly severe symptoms.

3.2 Section analysis

Now, we are going to start with the study of the dataset. First, we will store in a dataset called selected_data only those variables from the original dataset that we consider necessary for the model, in order to facilitate the access.

Then, we are going to generate another dataset called data_sample, by selecting only 1000 random patients, using replacement. The main goal in this section is to calculate the simulation from the error and then compute the 95% confidence intervals and the histograms of both models.

3.2.a Parametric bootstrap

The bootstrap method is a resampling technique used in statistics to estimate the distribution of a statistic by repeatedly sampling, with replacement, from the original data.

In our code, we have chosen to apply bootstrapping specifically to analyze the dataset based on the error.

The use of errors in this process is crucial because it allows the generation of new datasets that reflect the variability and unpredictability inherent in the original data.

By adding random errors to the fitted values, the bootstrap simulations create a range of possible scenarios that the model could encounter. This helps to assess how sensitive the model's coefficients are to variations in the data, assessing their correlation, and determining confidence intervals.

By employing bootstrapping in this manner, we aim to gain deeper insights into the dataset's characteristics and obtain more reliable estimations.

We first estimate the parameters by fitting a regression model to the dataset (data_sample). Subsequently, we employ the estimated distribution to simulate new samples, noting that the choice of regression model we'll employ depends on the model we've established.

- **Age regression model**
  We've utilized linear regression to model a continuous dependent variable (Age). In this case, we generated errors using a normal distribution with the rnorm function.

- **Lung cancer stage regression model**
  We've used multinomial logistic regression to handle a categorical dependent variable (Level) with multiple categories. This type of regression is an extension of logistic regression and is specifically designed to handle situations where there are more than two possible outcomes or levels in the response variable. The error follows a Multinomial(len,5,prob) distribution. To compute the probabilities we will fit the model using pp <-fitted(grade_regression). These errors align with the categorical nature of the dependent variable.

Once we have the regression models, we extract the residual standard deviation (sigma) from the summary of the model.

Then, we generate the evaluate function which computes the response variable lineal regressions formula. As we already know the coefficients and the estimating parameters, we can calculate the error.

Afterwards, we perform the bootstrapping loop, which goes up to 1000 simulations. It starts by generating the errors from the Normal or multinomial distribution and generating the new observed value.

These errors are added to the original response variable, simulating variability in the data and a new regression model is fitted to the modified dataset. The coefficients and correlations of the fitted models are then extracted and stored in vectors (c1, c2...) for coefficients, and (cor1, cor2...) for correlations. These correlations will tell us how good the regression is (we want the correlations to be either close to -1 or 1). This process is then repeated for a specified number of iterations (n_sim).

Subsequently, we will calculate the confidence intervals for each of the vectors in order to acknowledge the quality of our new Dataset after Bootstrapping.

3.2.b Non-parametric bootstrap

The non-parametric bootstrap is a resampling technique used to estimate the distribution of a statistic without assuming any specific parametric form for the underlying population distribution.

To carry out this method, we start by setting up similar vectors as used in the parametric bootstrap. Then, in the main loop, we create a new sample by randomly picking observations from the data_sample with replacement.

Once we have this new sample, we find its length and run a new regression using it as our dataset. We grab the coefficients from this regression summary and save them in the respective vectors.

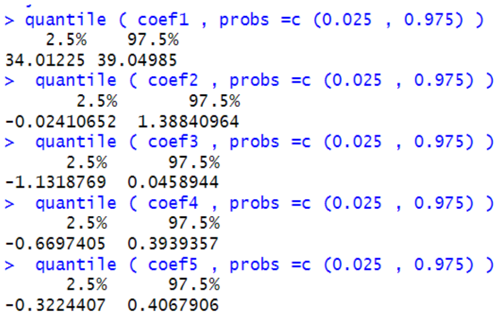
To calculate correlations, we consider both estimated and estimating parameters. Finally, we compute confidence intervals and histograms for the estimated parameters to gain a better understanding of the model's behavior and the distribution of coefficients.

4 Analysis of the results

The results obtained from the bootstrap of the models are presented below.

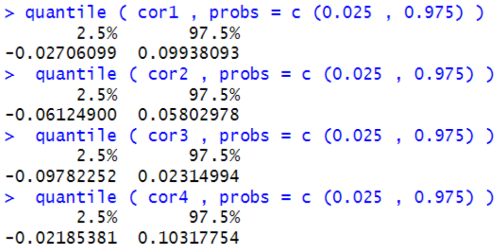
4.0.a Age regression model

To begin with, the obtained 95% confidence intervals using Parametric Bootstrapping can be observed hereunder.



```
> quantile ( coef1 , probs =c (0.025 , 0.975) )
    2.5%    97.5% 
34.01225 39.04985 
>  quantile ( coef2 , probs =c (0.025 , 0.975) )
       2.5%       97.5% 
-0.02410652  1.38840964 
>  quantile ( coef3 , probs =c (0.025 , 0.975) )
      2.5%      97.5% 
-1.1318769  0.0458944 
>  quantile ( coef4 , probs =c (0.025 , 0.975) )
      2.5%      97.5% 
-0.6697405  0.3939357 
>  quantile ( coef5 , probs =c (0.025 , 0.975) )
      2.5%      97.5% 
-0.3224407  0.4067906 
```

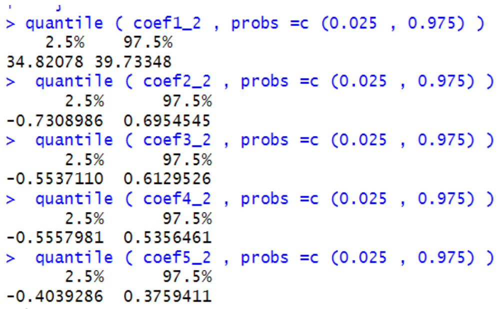
Figure 1: 95% confidence intervals computed using the five computed coefficients



```
> quantile ( cor1 , probs = c (0.025 , 0.975) )
       2.5%       97.5% 
-0.02706099  0.09938093 
>  quantile ( cor2 , probs = c (0.025 , 0.975) )
       2.5%       97.5% 
-0.06124900  0.05802978 
>  quantile ( cor3 , probs = c (0.025 , 0.975) )
       2.5%       97.5% 
-0.09782252  0.02314994 
>  quantile ( cor4 , probs = c (0.025 , 0.975) )
       2.5%       97.5% 
-0.02185381  0.10317754 
```

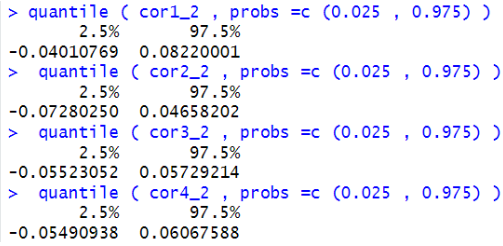
Figure 2: 95% confidence intervals computed using the four computed correlations

Using Non-Parametric Bootstrapping we have obtained similar 95% confidence intervals:



```
> quantile ( coef1_2 , probs =c (0.025 , 0.975) )
    2.5%    97.5% 
34.82078 39.73348 
>  quantile ( coef2_2 , probs =c (0.025 , 0.975) )
      2.5%      97.5% 
-0.7308986  0.6954545 
>  quantile ( coef3_2 , probs =c (0.025 , 0.975) )
      2.5%      97.5% 
-0.5537110  0.6129526 
>  quantile ( coef4_2 , probs =c (0.025 , 0.975) )
      2.5%      97.5% 
-0.5557981  0.5356461 
>  quantile ( coef5_2 , probs =c (0.025 , 0.975) )
      2.5%      97.5% 
-0.4039286  0.3759411 
```

Figure 3: 95% confidence intervals computed using the five computed coefficients



```
> quantile ( cor1_2 , probs =c (0.025 , 0.975) )
       2.5%       97.5% 
-0.04010769  0.08220001 
>  quantile ( cor2_2 , probs =c (0.025 , 0.975) )
       2.5%       97.5% 
-0.07280250  0.04658202 
>  quantile ( cor3_2 , probs =c (0.025 , 0.975) )
       2.5%       97.5% 
-0.05523052  0.05729214 
>  quantile ( cor4_2 , probs =c (0.025 , 0.975) )
       2.5%       97.5% 
-0.05490938  0.06067588 
```

Figure 4: 95% confidence intervals computed using the four computed correlations

In order to assess the quality and accuracy of the coefficients, we will generate histograms for each of them. A valid histogram should exhibit a resemblance to a normal distribution. Although, the correlation histograms are not of primary importance, we have also included them for reference. Let's observe the following five graphics that show the accuracy of each coefficient in relation to the patient's age at the time of lung cancer detection:

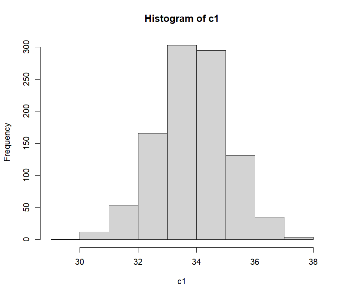


Figure 5: Histogram depicting the accuracy of coefficient 1, associated with the intercept, with respect to patient's Age.

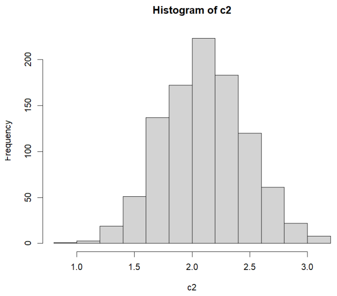
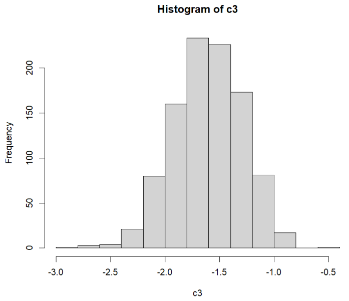
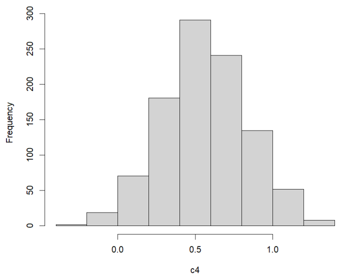


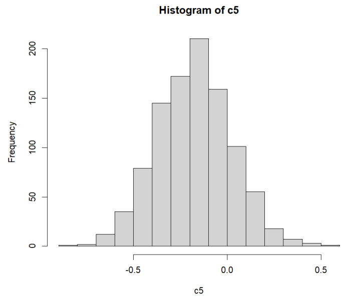
Figure 6: Histogram depicting the accuracy of coefficient 2, associated with the genetic risk with respect to patient's Age.



(a) Histogram depicting the accuracy of coefficient 3, associated with the chest pain with respect to patient's Age.



(b) Histogram depicting the accuracy of coefficient 4, associated with the snoring with respect to patient's Age.



(c) Histogram depicting the accuracy of coefficient 5, associated with the shortness of breath, with respect to patient's Age.

Upon analyzing the generated graphics, we can confidently confirm that all coefficients indeed exhibit a resemblance to a normal distribution.

The results obtained using Non-parametric Bootstrapping are quite similar. We can see some examples hereunder:

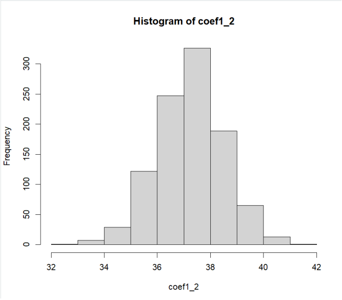


Figure 8: Histogram depicting the accuracy of coefficient 1 (calculated using Non-Parametric Bootstrapping), associated with the intercept.

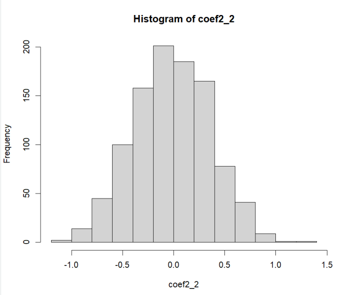
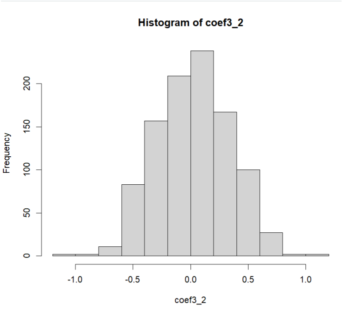
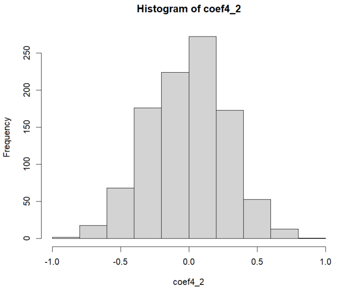


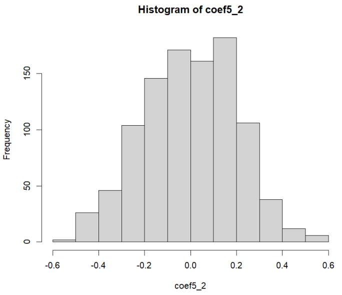
Figure 9: Histogram depicting the accuracy of coefficient 2 (calculated using Non-Parametric Bootstrapping), associated with the genetic risk, with respect to patient's Age



(a) Histogram depicting the accuracy of coefficient 3 (calculated using Non-Parametric Bootstrapping), associated with the chest pain, with respect to patient's Age



(b) Histogram depicting the accuracy of coefficient 4 (calculated using Non-Parametric Bootstrapping), associated with the snoring, with respect to patient's Age



(c) Histogram depicting the accuracy of coefficient 5 (calculated using Non-Parametric Bootstrapping), associated with the shortness of breath, with respect to patient's Age

Let us examine some of the correlation histograms provided for reference:

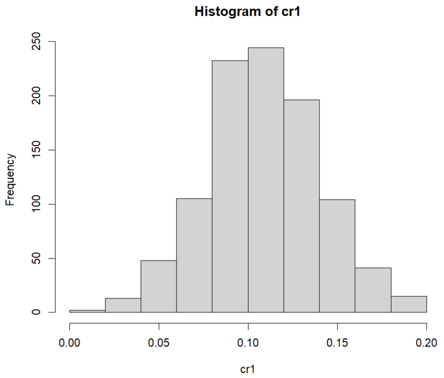


Figure 11: Correlation histogram between patient's Age and calculated coefficient 1.

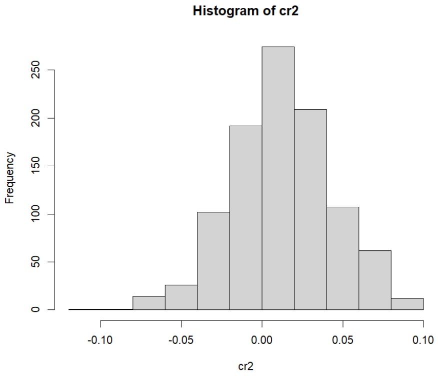


Figure 12: Correlation histogram between patient's Age and calculated coefficient 2.

Let's see an example of how similar the results using Non-parametric Bootstrapping have been:

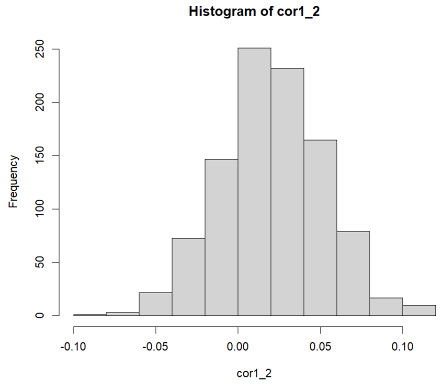


Figure 13: Correlation histogram between patient's Age and calculated coefficient 1 using Non-Parametric Bootstrap.

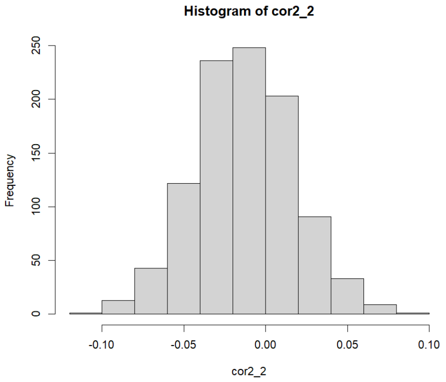
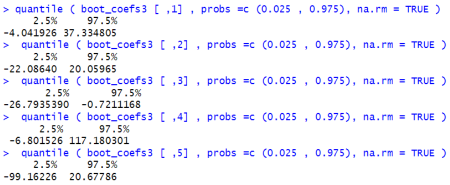


Figure 14: Correlation histogram between patient's Age and calculated coefficient 2 using Non-Parametric Bootstrap.

When comparing the results obtained using parametric and non-parametric bootstrap techniques, we observed consistency in the estimates, validating the robustness of the models used.

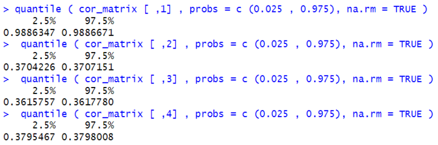
4.0.b Lung cancer stage regression model

Firstly, the obtained 95% confidence intervals using Parametric Bootstrapping can be observed hereunder:



```
> quantile ( boot_coefs3 [ ,1] , probs =c (0.025 , 0.975), na.rm = TRUE )
     2.5%     97.5%
-4.041926 37.334805
>  quantile ( boot_coefs3 [ ,2] , probs =c (0.025 , 0.975), na.rm = TRUE )
     2.5%     97.5%
-22.08640  20.05965
>  quantile ( boot_coefs3 [ ,3] , probs =c (0.025 , 0.975), na.rm = TRUE )
       2.5%       97.5%
-26.7935390  -0.7211168
>  quantile ( boot_coefs3 [ ,4] , probs =c (0.025 , 0.975), na.rm = TRUE )
      2.5%      97.5%
 -6.801526 117.180301
>  quantile ( boot_coefs3 [ ,5] , probs =c (0.025 , 0.975), na.rm = TRUE )
     2.5%     97.5%
-99.16226  20.67786
```

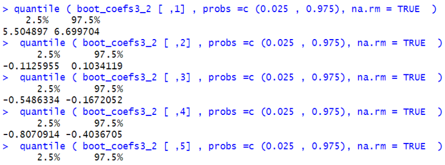
Figure 15: 95% confidence intervals computed using the five computed coefficients



```
> quantile ( cor_matrix [ ,1] , probs = c (0.025 , 0.975), na.rm = TRUE )
     2.5%     97.5%
0.9886347 0.9886671
>  quantile ( cor_matrix [ ,2] , probs = c (0.025 , 0.975), na.rm = TRUE )
     2.5%     97.5%
0.3704226 0.3707151
>  quantile ( cor_matrix [ ,3] , probs = c (0.025 , 0.975), na.rm = TRUE )
     2.5%     97.5%
0.3615757 0.3617780
>  quantile ( cor_matrix [ ,4] , probs = c (0.025 , 0.975), na.rm = TRUE )
     2.5%     97.5%
0.3795467 0.3798008
```

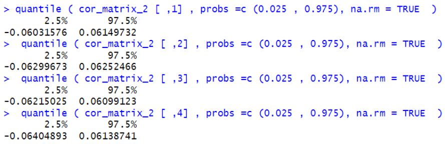
Figure 16: 95% confidence intervals computed using the four computed correlations

Using Non-Parametric Bootstrapping we have obtained similar 95% confidence intervals:



```
> quantile ( boot_coefs3_2 [ ,1] , probs =c (0.025 , 0.975), na.rm = TRUE  )
    2.5%    97.5%
5.504897 6.699704
>  quantile ( boot_coefs3_2 [ ,2] , probs =c (0.025 , 0.975), na.rm = TRUE  )
      2.5%      97.5%
-0.1125955  0.1034119
>  quantile ( boot_coefs3_2 [ ,3] , probs =c (0.025 , 0.975), na.rm = TRUE  )
      2.5%      97.5%
-0.5486334 -0.1672052
>  quantile ( boot_coefs3_2 [ ,4] , probs =c (0.025 , 0.975), na.rm = TRUE  )
      2.5%      97.5%
-0.8070914 -0.4036705
>  quantile ( boot_coefs3_2 [ ,5] , probs =c (0.025 , 0.975), na.rm = TRUE  )
      2.5%      97.5%
```

Figure 17: 95% confidence intervals computed using (using Non-Parametric Bootstrap) the five computed coefficients



```
> quantile ( cor_matrix_2 [ ,1] , probs =c (0.025 , 0.975), na.rm = TRUE  )
       2.5%       97.5%
-0.06031576  0.06149732
>  quantile ( cor_matrix_2 [ ,2] , probs =c (0.025 , 0.975), na.rm = TRUE  )
       2.5%       97.5%
-0.06299673  0.06252466
>  quantile ( cor_matrix_2 [ ,3] , probs =c (0.025 , 0.975), na.rm = TRUE  )
       2.5%       97.5%
-0.06215025  0.06099123
>  quantile ( cor_matrix_2 [ ,4] , probs =c (0.025 , 0.975), na.rm = TRUE  )
       2.5%       97.5%
-0.06404893  0.06138741
```

Figure 18: 95% confidence intervals computed using (using Non-Parametric Bootstrap) the four computed correlations

To evaluate the goodness and accuracy of our coefficients, we will generate histograms for each of them. However, we have also included the correlation histograms for reference, even though they do not need to resemble a specific distribution.

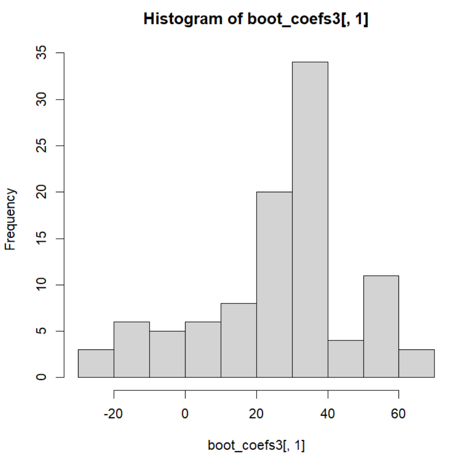


Figure 19: Histogram depicting the accuracy of coefficient 1, associated with the intercept, with respect to patient's level of cancer.

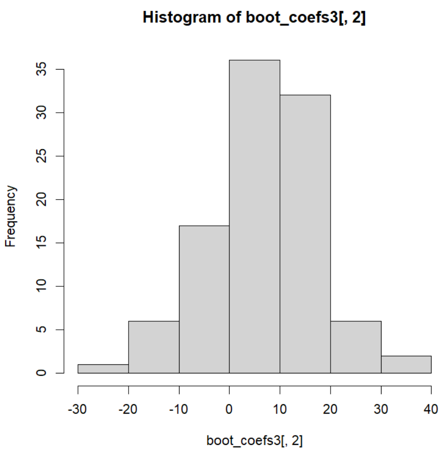
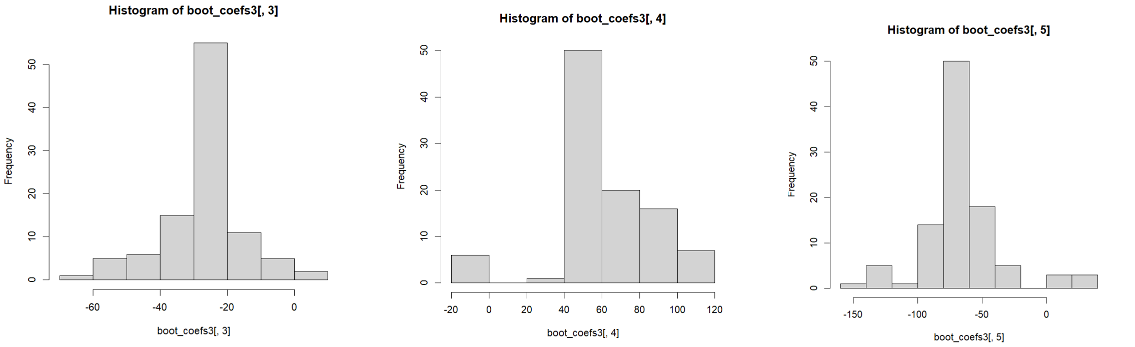


Figure 20: Histogram depicting the accuracy of coefficient 2, associated with the level of smoking, with respect to patient's level of cancer.



(a) Histogram depicting the accuracy of coefficient 3, associated with the level of air pollution, with respect to patient's level of cancer.

(b) Histogram depicting the accuracy of coefficient 4, associated with the genetic risk, with respect to patient's level of cancer.

(c) Histogram depicting the accuracy of coefficient 5, associated with the alcohol use, with respect to patient's level of cancer.

The results obtained using Non-parametric Bootstrapping are quite similar. We can see some examples hereunder:

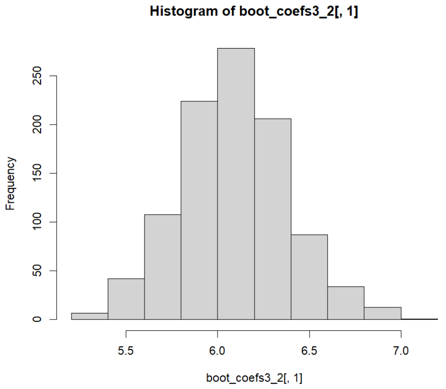


Figure 22: Histogram depicting the accuracy of coefficient 1 (with Non-Parametric Bootstrap), associated with the intercept, with respect to patient's level of cancer.

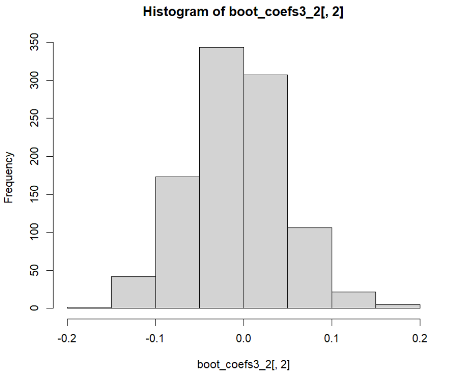
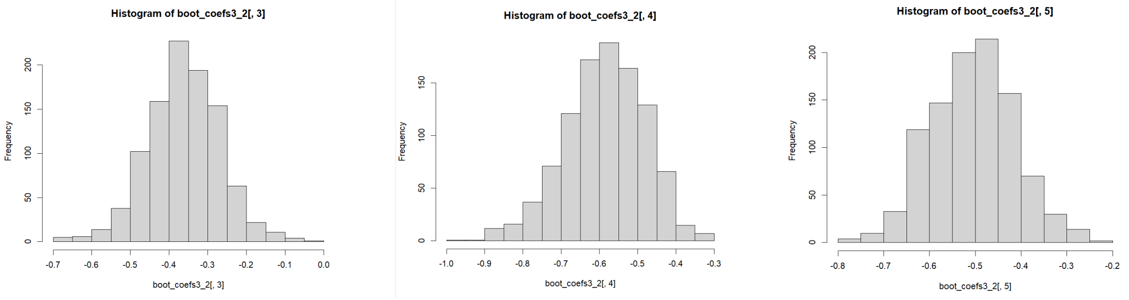


Figure 23: Histogram depicting the accuracy of coefficient 2 (with Non-Parametric Bootstrap), associated with the level of smoking, with respect to patient's level of cancer.



(a) Histogram depicting the accuracy of coefficient 3 (with Non-Parametric Bootstrap), associated with the level of air pollution, with respect to patient's level of cancer.

(b) Histogram depicting the accuracy of coefficient 4 (with Non-Parametric Bootstrap), associated with the genetic risk, with respect to patient's level of cancer.

(c) Histogram depicting the accuracy of coefficient 5 (with Non-Parametric Bootstrap), associated with the alcohol use, with respect to patient's level of cancer.

After analysing the obtained graphics, we can say more or less that all the coefficients resemble a multinomial distribution. Let's observe the correlation histograms included for reference:

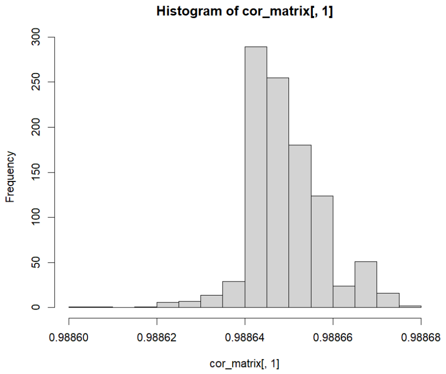


Figure 25: Correlation histogram between patient's level of cancer and calculated coefficient 1

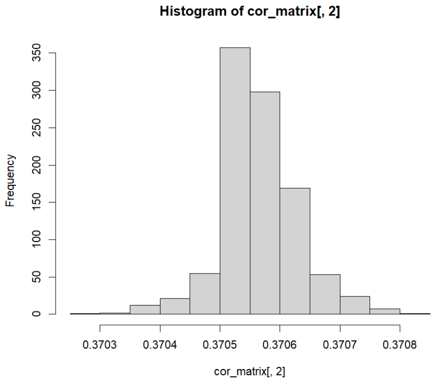


Figure 26: Correlation histogram between patient's level of cancer and calculated coefficient 2

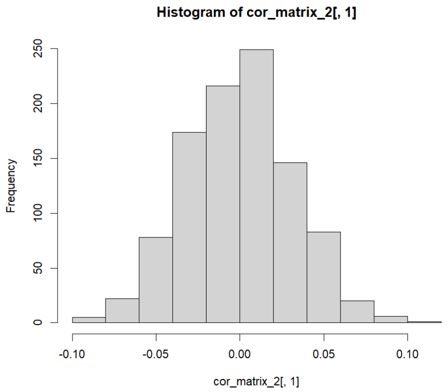


Figure 27: Correlation histogram between patient's level of cancer and calculated coefficient 1 with Non-Parametric Bootstrap

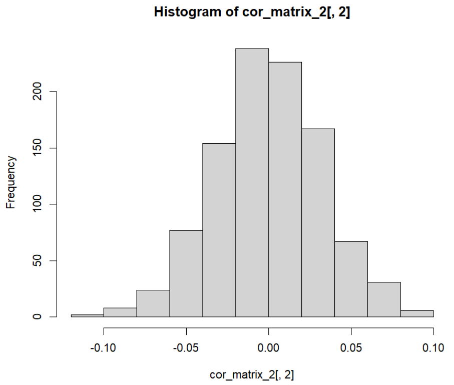


Figure 28: Correlation histogram between patient's level of cancer and calculated coefficient 2 with Non-Parametric Bootstrap

5 Conclusions

The objective of this project was to develop a predictive framework for lung cancer patients by exploring the relationship between medical history, genetic predisposition, lifestyle factors, and lung disease characteristics. We implemented two models: one to predict the stage of cancer and another to determine the age of cancer detection. The cancer stage regression model was based on variables such as smoking, air pollution, genetic risk, and alcohol consumption, while the age regression model considered genetic risk and respiratory problems.

Using parametric and non-parametric bootstrap techniques, we generated data simulations to evaluate the accuracy and reliability of the models. The analyses showed significant correlations between the predictor variables and the response variables. Specifically, we found that smoking and air pollution are strong predictors of cancer stage, while genetic risk and respiratory problems significantly affect the age of cancer detection. These findings provide valuable insights that can improve lung cancer prediction and early detection. They suggest that targeting these factors through specific interventions could significantly enhance disease management.

6 Bibliography

- Lung Cancer dataset. Kaggle. https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link
- Lung cancer-NHS. https://www.nhs.uk/conditions/lung-cancer/symptoms/
- Outdoor air pollution and cancer. https://pubmed.ncbi.nlm.nih.gov/32964460/
- Air pollution's role in the promotion of lung cancer. https://www.nature.com/articles/d41586-023-00929-x

7 Appendix

7.1 R Code

```
#---------------------READ THE FILE--------------------------------
setwd("C:/Users/USER/Desktop/2n␣curs/2n␣semestre/An lisi␣Dades␣Complexes/Practica/Final␣HW")
data <- read.csv("cancer␣patient␣data␣sets.csv")

#-----------------------PARAMETER ANALYSIS---------------------------
selected_data <- data [ , c ("Age", "Smoking", "Air_Pollution", "Genetic_Risk", "Alcohol_use", "Shortness_of_Breath", "Gender", "Chest_Pain", "Obesity", "Level", "Snoring")]
data_sample <- selected_data [ sample ( nrow ( selected_data ) , 1000 , replace = TRUE ) , ]

#-----------------AGE REGRESSION MODEL -------------------------------
AGE_GENETIC_RISK <- lm ( Genetic_Risk ~ Age , data_sample )
AGE_CHEST_PLAIN <- lm ( Chest_Pain ~ Age , data_sample )
AGE_SNORING <- lm ( Snoring ~ Age , data_sample )
AGE_SHORTNESS <- lm ( Shortness_of_Breath ~ Age , data_sample )

#Plot the regression models
plot ( data_sample$Snoring ~ data_sample$Age , type ='p', col ='Blue', xlab ='Age'
           , ylab ='Snoring')
 abline ( AGE_SNORING, col='red' )

 plot ( data_sample$Shortness_of_Breath ~ data_sample$Age , type ='p', col ='Blue', xlab ='Age'
        , ylab ='Shortness␣of␣breath')
 abline ( AGE_SHORTNESS, col='red' )

 plot ( data_sample$Genetic_Risk ~ data_sample$Age , type ='p', col ='Blue', xlab ='Age'
        , ylab ='Genetic␣risk')
 abline ( AGE_GENETIC_RISK, col='red'  )

 plot ( data_sample$Chest_Pain ~ data_sample$Age , type ='p', col ='Blue', xlab ='Age'
        , ylab ='Chest␣Pain')
 abline ( AGE_CHEST_PLAIN, col='red'  )


#PARAMETRIC BOOTSTRAP------------------------------------------------

 regressio <- lm (Age ~Genetic_Risk+Chest_Pain+Snoring+Shortness_of_Breath, data = data_sample )
 summary ( regressio )
 resum <- summary ( regressio )
 sigma <- resum$sigma
 n_sim <- 1000
 len <- length ( data_sample$Age )
 c1 <- numeric ( n_sim )
 c2 <- numeric ( n_sim )
 c3 <- numeric ( n_sim )
 c4 <- numeric ( n_sim )
 c5 <- numeric ( n_sim )
 cr1 <- numeric ( n_sim )
 cr2 <- numeric ( n_sim )
 cr3 <- numeric ( n_sim )
 cr4 <- numeric ( n_sim )


 evaluate <- function ( error ){
   resum$coefficient [1] + resum$coefficient [2] * data_sample$Genetic_Risk +
     resum$coefficient [3] * data_sample$Chest_Pain +
     resum$coefficient [4]* data_sample$Snoring + resum$coefficient [5] *
     data_sample$Shortness_of_Breath + error
 }
  for (i in 1: n_sim ){
   error <- rnorm ( len , 0 , sigma )
   new_type <- evaluate ( error )
   new_regressio <- lm ( new_type ~ data_sample$Genetic_Risk +
                               data_sample$Chest_Pain + data_sample$Snoring +
                               data_sample$Shortness_of_Breath)
   c1[i] <- new_regressio$coefficient [1]
   c2[i] <- new_regressio$coefficient [2]
   c3[i] <- new_regressio$coefficient [3]
   c4[i] <- new_regressio$coefficient [4]
   c5[i] <- new_regressio$coefficient [5]

   cr1[i] <- cor (new_type,data_sample$Genetic_Risk )
   cr2[i] <- cor (new_type,data_sample$Chest_Pain )
   cr3[i] <- cor (new_type,data_sample$Snoring )
   cr4[i] <- cor (new_type,data_sample$Shortness_of_Breath )
}

# 95% CONFIDENCE INTERVALS
 quantile (c1 , probs =c(0.025, 0.975))
 quantile (c2 , probs =c(0.025, 0.975))
 quantile (c3 , probs =c(0.025, 0.975))
 quantile (c4 , probs =c(0.025, 0.975))
 quantile (c5 , probs =c(0.025, 0.975))
 quantile (cr1 , probs = c(0.025, 0.975))
 quantile (cr2 , probs = c(0.025, 0.975))
 quantile (cr3 , probs = c(0.025, 0.975))
 quantile (cr4 , probs = c(0.025, 0.975))

# HISTOGRAMS
 hist ( c1 )
 hist ( c2 )
 hist ( c3 )
 hist ( c4 )
 hist ( c5 )
 hist ( cr1 )
 hist ( cr2 )
 hist ( cr3 )
 hist ( cr4 )

 #NON-PARAMETRIC BOOTSTRAP---------------------------------------
 coef1_2 <- numeric ( n_sim )
 coef2_2 <- numeric ( n_sim )
```

```
coef3_2 <- numeric ( n_sim )
coef4_2 <- numeric ( n_sim )
coef5_2 <- numeric ( n_sim )
cor1_2 <- numeric ( n_sim )
cor2_2 <- numeric ( n_sim )
cor3_2 <- numeric ( n_sim )
cor4_2 <- numeric ( n_sim )

for (j in 1: n_sim ){
  sample2 <- data_sample [ sample ( nrow ( data_sample ) , 1000 , replace
      = TRUE ) , ]
  error <- rnorm ( len , 0 , sigma )
  new_type <- evaluate ( error )
  new_regressio_bootstrap <- lm ( new_type ~ sample2$Genetic_Risk +
                                              sample2$Chest_Pain + sample2$
                                                  Snoring + sample2$Shortness_of
                                                  _Breath)
  coef1_2 [j] <- summary ( new_regressio_bootstrap ) $coefficient [1]
  coef2_2 [j] <- summary ( new_regressio_bootstrap ) $coefficient [2]
  coef3_2 [j] <- summary ( new_regressio_bootstrap ) $coefficient [3]
  coef4_2 [j] <- summary ( new_regressio_bootstrap ) $coefficient [4]
  coef5_2 [j] <- summary ( new_regressio_bootstrap ) $coefficient [5]

  cor1_2 [ j] <- cor ( sample2$Age , sample2$Genetic_Risk )
  cor2_2 [ j] <- cor ( sample2$Age , sample2$Chest_Pain )
  cor3_2 [ j] <- cor ( sample2$Age , sample2$Snoring )
  cor4_2 [ j] <- cor ( sample2$Age , sample2$Shortness_of_Breath )
  }

# 95% CONFIDENCE INTERVALS
quantile ( coef1_2 , probs =c (0.025, 0.975))
quantile ( coef2_2 , probs =c (0.025, 0.975))
quantile ( coef3_2 , probs =c (0.025, 0.975))
quantile ( coef4_2 , probs =c (0.025, 0.975))
quantile ( coef5_2 , probs =c (0.025, 0.975))
quantile ( cor1_2 , probs =c (0.025, 0.975))
quantile ( cor2_2 , probs =c (0.025, 0.975))
quantile ( cor3_2 , probs =c (0.025, 0.975))
quantile ( cor4_2 , probs =c (0.025, 0.975))

# HISTOGRAMS
hist ( coef1_2 )
hist ( coef2_2 )
hist ( coef3_2 )
hist ( coef4_2 )
hist ( coef5_2 )
hist ( cor1_2 )
hist ( cor2_2 )
hist ( cor3_2 )
hist ( cor4_2 )

#------------LUNG CANCER STAGE REGRESSION MODEL--------------------

data_sample$Level <- factor(data_sample$Level, levels = c("Low", "Medium",
    "High"))
data_sample$LevelNumeric <- as.numeric(data_sample$Level)
par(mar = c(5, 4, 4, 2) + 0.1)

```

```
LEVEL_SMOKING <- lm(Smoking ~ LevelNumeric, data = data_sample)
LEVEL_AIR_POLLUTION <- lm(Air_Pollution ~ LevelNumeric, data = data_sample)
LEVEL_GENETIC_RISK <- lm(Genetic_Risk ~ LevelNumeric, data = data_sample)
LEVEL_ALCOHOL_USE <- lm(Alcohol_use ~ LevelNumeric, data = data_sample)

#Plot the regression models
plot(data_sample$Smoking ~ data_sample$LevelNumeric, type = 'p', col = 'Blue', xlab = 'Level', ylab = 'Smoking')
abline(LEVEL_SMOKING, col = "red")

plot(data_sample$Air_Pollution ~ data_sample$LevelNumeric, type = 'p', col = 'Blue', xlab = 'Level', ylab = 'Air␣Pollution')
abline(LEVEL_AIR_POLLUTION, col = "red")

plot(data_sample$Genetic_Risk ~ data_sample$LevelNumeric, type = 'p', col = 'Blue', xlab = 'Level', ylab = 'Genetic␣Risk')
abline(LEVEL_GENETIC_RISK, col = "red")

plot(data_sample$Alcohol_use ~ data_sample$LevelNumeric, type = 'p', col = 'Blue', xlab = 'Level', ylab = 'Alcohol␣use')
abline(LEVEL_ALCOHOL_USE, col = "red")

#PARAMETRIC BOOTSTRAP--------------------------------------------------
data_sample$Level <- as.factor( data_sample$Level) #
grade_regression <- multinom( Level ~ Smoking + Air_Pollution + Genetic_Risk
                               + Alcohol_use , data = data_sample )
summary <- summary ( grade_regression )
pp <- fitted ( grade_regression )
prob <- c ( pp )
n_sim <- 100
len <- length ( data_sample$Level )

boot_coefs3 <- matrix (NA , nrow = n_sim , ncol = ncol ( coef ( grade_regression )))
boot_coefs4 <- matrix (NA , nrow = n_sim , ncol = ncol ( coef ( grade_regression )))
cor_matrix <- matrix ( NA , nrow = n_sim , ncol = 5)
evaluate <- function ( error ){
  summary$coefficient [1] + summary$coefficient [2] * data_sample$Smoking +
    summary$coefficient [3] * data_sample$Air_Pollution + summary$coefficient [4] *
    data_sample$Genetic_Risk + summary$coefficient [5] * data_sample$Alcohol_use+ error
}

for (i in 1: n_sim ) {
   error <- rmultinom (len ,3 , prob )
   new_type <- evaluate ( error [1: nrow ( data_sample ) ])
   new_regressio <- multinom ( new_type ~ Smoking + Air_Pollution +
                                 Genetic_Risk + Alcohol_use, data = data_sample , trace = FALSE )
   boot_coefs3 [i , ] <- coef ( new_regressio ) [1 , ]
   boot_coefs4 [i , ] <- coef ( new_regressio ) [2 , ]

  cor_matrix [i , 1] <- cor ( new_type , data_sample$Smoking )
  cor_matrix [i , 2] <- cor ( new_type , data_sample$Air_Pollution )
  cor_matrix [i , 3] <- cor ( new_type , data_sample$Genetic_Risk )
  cor_matrix [i , 4] <- cor ( new_type , data_sample$Alcohol_use )
}

# 95% CONFIDENCE INTERVALS

quantile ( boot_coefs3 [ ,1] , probs =c (0.025 , 0.975) , na.rm = TRUE )
quantile ( boot_coefs3 [ ,2] , probs =c (0.025 , 0.975) , na.rm = TRUE )
quantile ( boot_coefs3 [ ,3] , probs =c (0.025 , 0.975) , na.rm = TRUE )
quantile ( boot_coefs3 [ ,4] , probs =c (0.025 , 0.975) , na.rm = TRUE )
quantile ( boot_coefs3 [ ,5] , probs =c (0.025 , 0.975) , na.rm = TRUE )
quantile ( boot_coefs4 [ ,1] , probs =c (0.025 , 0.975) , na.rm = TRUE )
quantile ( boot_coefs4 [ ,2] , probs =c (0.025 , 0.975) , na.rm = TRUE )
quantile ( boot_coefs4 [ ,3] , probs =c (0.025 , 0.975) , na.rm = TRUE )
quantile ( boot_coefs4 [ ,4] , probs =c (0.025 , 0.975) , na.rm = TRUE )
quantile ( boot_coefs4 [ ,5] , probs =c (0.025 , 0.975) , na.rm = TRUE )
quantile ( cor_matrix [ ,1] , probs = c (0.025 , 0.975) , na.rm = TRUE )
quantile ( cor_matrix [ ,2] , probs = c (0.025 , 0.975) , na.rm = TRUE )
quantile ( cor_matrix [ ,3] , probs = c (0.025 , 0.975) , na.rm = TRUE )
quantile ( cor_matrix [ ,4] , probs = c (0.025 , 0.975) , na.rm = TRUE )

#HISTOGRAMS
hist ( boot_coefs3 [ ,1])
hist ( boot_coefs3 [ ,2])
hist ( boot_coefs3 [ ,3])
hist ( boot_coefs3 [ ,4])
hist ( boot_coefs3 [ ,5])
hist ( boot_coefs4 [ ,1])
hist ( boot_coefs4 [ ,2])
hist ( boot_coefs4 [ ,3])
hist ( boot_coefs4 [ ,4])
hist ( boot_coefs4 [ ,5])
hist ( cor_matrix [ ,1])
hist ( cor_matrix [ ,2])
hist ( cor_matrix [ ,3])
hist ( cor_matrix [ ,4])

#NON-PARAMETRIC BOOTSTRAP---------------------------------------
boot_coefs3_2 <- matrix (NA , nrow = n_sim , ncol = ncol ( coef ( grade_regression )))
boot_coefs4_2 <- matrix (NA , nrow = n_sim , ncol = ncol ( coef ( grade_regression )))
cor_matrix_2 <- matrix ( NA , nrow = n_sim , ncol = 5)

for (i in 1: n_sim ) {
  bootstrap_sample <- data_sample [ sample ( nrow ( data_sample ) , replace = TRUE ) , ]
  error <- rmultinom (len ,3 , prob )
  new_type <- evaluate ( error [1: nrow ( bootstrap_sample ) ])
  new_regressio <- multinom ( Level ~ Smoking + Air_Pollution + Genetic_Risk +
                                              Alcohol_use , data = bootstrap_sample , trace = FALSE )
  boot_coefs3_2 [i , ] <- coef ( new_regressio ) [1 , ]
  boot_coefs4_2 [i , ] <- coef ( new_regressio ) [2 , ]

  cor_matrix_2 [i , 1] <- cor ( new_type , bootstrap_sample$Smoking )
```

```
    cor_matrix_2 [i , 2] <- cor ( new_type , bootstrap_sample$Air_Pollution
       )
    cor_matrix_2 [i , 3] <- cor ( new_type , bootstrap_sample$Genetic_Risk )
    cor_matrix_2 [i , 4] <- cor ( new_type , bootstrap_sample$Alcohol_use )
    }

# 95% CONFIDENCE INTERVALS
quantile ( boot_coefs3_2 [ ,1] , probs =c (0.025, 0.975), na.rm = TRUE  )
quantile ( boot_coefs3_2 [ ,2] , probs =c (0.025, 0.975), na.rm = TRUE  )
quantile ( boot_coefs3_2 [ ,3] , probs =c (0.025, 0.975), na.rm = TRUE  )
quantile ( boot_coefs3_2 [ ,4] , probs =c (0.025, 0.975), na.rm = TRUE  )
quantile ( boot_coefs3_2 [ ,5] , probs =c (0.025, 0.975), na.rm = TRUE  )
quantile ( boot_coefs4_2 [ ,1] , probs =c (0.025 , 0.975), na.rm = TRUE  )
quantile ( boot_coefs4_2 [ ,2] , probs =c (0.025 , 0.975), na.rm = TRUE  )
quantile ( boot_coefs4_2 [ ,3] , probs =c (0.025 , 0.975), na.rm = TRUE  )
quantile ( boot_coefs4_2 [ ,4] , probs =c (0.025 , 0.975), na.rm = TRUE  )
quantile ( boot_coefs4_2 [ ,5] , probs =c (0.025 , 0.975), na.rm = TRUE  )
quantile ( cor_matrix_2 [ ,1] , probs =c (0.025 , 0.975), na.rm = TRUE  )
quantile ( cor_matrix_2 [ ,2] , probs =c (0.025 , 0.975), na.rm = TRUE  )
quantile ( cor_matrix_2 [ ,3] , probs =c (0.025 , 0.975), na.rm = TRUE  )
quantile ( cor_matrix_2 [ ,4] , probs =c (0.025 , 0.975), na.rm = TRUE  )

 # HISTOGRAMS
hist ( boot_coefs3_2 [ ,1])
hist ( boot_coefs3_2 [ ,2])
hist ( boot_coefs3_2 [ ,3])
hist ( boot_coefs3_2 [ ,4])
hist ( boot_coefs3_2 [ ,5])
hist ( boot_coefs4_2 [ ,1])
hist ( boot_coefs4_2 [ ,2])
hist ( boot_coefs4_2 [ ,3])
hist ( boot_coefs4_2 [ ,4])
hist ( boot_coefs4_2 [ ,5])
hist ( cor_matrix_2 [ ,1])
hist ( cor_matrix_2 [ ,2])
hist ( cor_matrix_2 [ ,3])
hist ( cor_matrix_2 [ ,4])
```