VIDYASAGAR UNIVERSITY

PROJECT SUBMITTED FOR PARTIAL FULFILMENT OF BACHELOR'S DEGREE IN STATISTICS HONOURS

STATISTICAL ANALYSIS ON IMPACT OF SEVERAL FACTOR ON LUNG CANCER

REGISTRATION NO: 1160461

SESSION: 2020-2021

DEPARTMENT OF STATISTICS
HALDIA GOVERNMENT COLLEGE

ACKNOWLEDGMENT

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along with the completion of my project. All that I have done is only due to their supervision and assistance and I would not forget to thank them. I respect and thank, for providing me an opportunity to do the project work under the department of Statistics, Haldia Government College and giving you all the support and guidance that I required, which made me complete the project duly on time. I owe my deep gratitude to our project guides Dr. Shyamsundar Sahoo, Mr. Sibsankar Karan, Mr. Tanmay Kumar Maity and Mr. Bijitesh Halder, who took keen interest on my project work and guided me all along, till the completion of our project work by providing all the necessary information for developing a good system. I would not forget to remember my parents, and also my friends for their encouragement and more over for their timely support and guidance till the completion of our project work**.**

RAJ SHEKHAR DAS

CONTENTS

INTRODUCTION

Lung cancer is the leading cause of cancer death worldwide, accounting for 1.59 million deaths in 2018. Lung cancer is the most dangerous and deadliest type of cancer. The majority of lung cancer cases are attributed to smoking, and it accounts for 85 out of 100 people dying every year, but exposure to air pollution is also a risk factor. Uranium is a metallic chemical element, which breaks down, with time, to form radon gas, which spreads in the air and water causing pollution and great harm to the lungs. A new study has found that air pollution may be linked to an increased risk of lung cancer, even in non-smokers. The researchers found that the people in the high-pollution group were more likely to develop lung cancer than those in the low-pollution group. They also found that the risk was higher in non-smokers than smokers, and that the risk increased with age. While this study does not prove that air pollution causes lung cancer, it does suggest that there may be a link between the two. More research is needed to confirm these findings and to determine what effect different types and levels of air pollution may have on lung cancer risk**.**

Our project will entail collecting data on the relevant predictors and applying a chi-square test to identify the most significant factors to include in the logistic regression model. We will then fit the logistic regression model to the data to estimate the effect of each predictor on the likelihood of heart disease while controlling for others factor.

OBJECTIVE

Objective of this study is shown as given below-

- Identifying risk factors for lung cancer
- Predict the probability of having lung cancer of patients

DATA SOURCE & DATA COLLECTION

The study, which was published in the journal "Nature Medicine", looked at data from over 462,000 people in China who were followed for an average of six years.

Link :: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7044659/

From above population, we choose 1000 random samples and work on it.

This dataset contains information on patients with lung cancer, including their age, gender, air pollution exposure, alcohol use, dust allergy, occupational hazards, genetic risk, chronic lung disease, balanced diet, obesity, smoking, passive smoker, chest pain, coughing of blood, fatigue, weight loss ,shortness of breath ,wheezing ,swallowing difficulty ,clubbing of finger nails and snoring.

DATA DESCRIPTION

The given dataset contains information on patients with lung cancer, including many casual factors. Details about the factors explained in brief is given below -

Causal Factor –

1. Air pollution exposure
2. Alcohol use
3. Dust allergy
4. Occupational hazards
5. Genetic risk
6. Chronic lung disease
7. Balanced diet
8. Obesity
9. Smoking
10. Passive smoker
11. Chest pain
12. Coughing of blood
13. Fatigue
14. Weight loss
15. Shortness of breath
16. Wheezing
17. Swallowing difficulty

18. Clubbing of finger nails

19. Frequent Cold

20. Dry Cough

21. Snoring

22. Age

23. Gender

Description of Factor –

* **AGE (x1)**

- The age of the patient. (Numeric)

* **Gender (x2)**

- The gender of the patient. (Categorical)

* **Air Pollution (x3)**

- The level of air pollution exposure of the patient. (Categorical)

* **Alcohol use (X4)**

- The level of alcohol use of the patient.(Categorical)

* **Dust Allergy (x5)**

- The level of dust allergy of the patient.(Categorical)

* **Occupational Hazards (x6)**

- The level of occupational hazards of the patient. (Categorical)

* **Genetic Risk (x7)**

- The level of genetic risk of the patient. (Categorical)

* **Chronic Lung Disease (x8)**

- The level of chronic lung disease of the patient. (Categorical)

* **Balanced Diet (x9)**

- The level of balanced diet of the patient. (Categorical)

* **Obesity (x10)**

- The level of obesity of the patient. (Categorical)

* **Smoking (x11)**

- The level of smoking of the patient. (Categorical)

* **Passive Smoker (x12)**

- The level of passive smoker of the patient. (Categorical)

* **Chest Pain (x13)**

- The level of chest pain of the patient. (Categorical)

* **Coughing of Blood (x14)**

- The level of coughing of blood of the patient. (Categorical)

* **Frequent Cold (x21)**

- The level of fatigue of the patient. (Categorical)

* **Dry Cough (x22)**

- The level of fatigue of the patient. (Categorical)

* **Snoring (x23)**

- The level of fatigue of the patient. (Categorical)

Index of Factor

* **Gender**

  :: Male = 1(598), Female= 2(402)

* **Age**

  :: under 30="1", 30-50="2", above 50="3"

* **Genetic risk, Chronic lung disease, Balanced diet, Obesity, Frequent Cold, Dry Cough, Snoring**

  :: Low(1,2),Medium(3,4,5), High(6,7) [1 to 7 {ascending order (lower to higher)}]

* **Air pollution exposure, Alcohol use, Dust allergy, Occupational hazards, Passive smoker, Fatigue, Weight loss, Shortness of breath, Wheezing, Clubbing of finger nails**

  :: Low(1,2,3),Medium(4,5,6),High(7,8) [1 to 8 {ascending order (lower to higher)}]

* **Smoking, Chest pain, Coughing of blood, Swallowing difficulty**

  :: Low(1,2,3),Medium(4,5,6), Medium(7,8,9) [1 to 9 {ascending order (lower to higher)}]

* **Level (y)**

  :: level of lung cancer (high="2", medium="1", low="0")

❖ **Frist 16 observation of dataset are given below –**

index	Patient Id	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
0	P1	2	1	1	2	2	2	2	1	1	2	1
1	P10	1	1	1	1	2	1	2	1	1	1	1
2	P100	2	1	2	2	2	2	2	2	3	3	1
3	P1000	2	1	3	3	3	3	3	3	3	3	3
4	P101	2	1	2	3	3	3	3	3	3	3	3

5	P102	2	1	2	2	2	2	2	2	3	3	1
6	P103	3	2	1	2	2	2	2	1	1	2	1
7	P104	1	2	1	1	2	1	1	2	2	2	1
8	P105	2	2	2	2	2	2	3	2	2	2	2
9	P106	2	1	1	1	2	1	2	2	2	2	1
10	P107	2	1	2	3	3	3	3	3	3	3	3
11	P108	3	2	2	3	3	3	3	3	3	3	3
12	P109	2	2	2	2	2	2	2	2	3	3	2
13	P11	2	1	2	3	3	3	3	3	3	3	3
14	P110	1	2	1	1	2	1	2	1	2	2	1
15	P111	3	1	2	2	2	2	3	2	3	2	3

x11	x12	x13	x14	x15	x16	x17	x18	x19	x20	x21	x22	x23	y
1	1	1	2	1	2	1	1	1	1	1	2	2	0
1	2	1	1	1	1	3	3	2	1	1	3	1	1
1	1	2	3	3	3	3	1	1	2	3	3	1	2
3	3	3	3	2	1	1	1	2	2	3	3	2	2
3	3	3	3	1	1	2	1	2	1	2	1	2	2
1	1	2	3	3	3	3	1	1	2	3	3	1	2
1	1	1	2	1	2	1	1	1	1	1	2	2	0
1	2	1	1	1	1	1	2	1	1	2	2	2	0
2	2	2	2	1	2	1	1	2	2	1	2	1	1
1	1	2	2	1	1	2	2	2	2	1	1	2	1
3	3	3	3	2	1	1	3	3	1	2	2	2	2
3	3	3	3	3	2	2	3	1	2	2	1	2	2
2	2	2	2	2	1	1	2	1	1	3	2	3	1
3	3	3	3	2	1	1	1	2	2	3	3	2	2
1	1	2	1	1	1	1	2	1	2	1	3	1	0
3	2	2	2	2	1	2	1	1	1	1	3	1	`1

METHODOLOGY

To analysis the data we use some statistical methodology which are given below -

- Identifying risk factors for lung cancer
  - Pearson's chi-square test(test of independence)

- Predict the probability of having lung cancer of patients
  - Multinomial logistic regression

Chi-square test for independence of two attributes -

The Chi-square test of independence checks whether two variables are likely to be related or not i.e. to check independency between two attributes. It is commonly used in research to test the hypothesis that there is no significant difference between observed frequency and expected frequency of the attributes being studied.

Now, to conduct the test, follows the following step which given below –

1. **Define null hypothesis –**

   H0: There is no association between two attributes

2. **Data collect –**

   Collect data on the two attributes being studied in the form of a contingency table. The contingency table shows the frequency of each category of the two attributes.

3. **Calculate the expected frequencies -**

   Calculate the expected frequencies for each cell in the contingency table using the formula- E= (Row total*Column total)/Grand total.

4. **Calculate the test statistic –**

   Calculate the chi-square statistic using given formula

$$\chi 2 = \sum_{i=1}^{rc} \frac{(o_i - E_i)^2}{E_i} \quad \sim \chi^2_{(r-1)(c-1)}, \text{ under null hypothesis.}$$

Where, o_i = the observed frequency , Ei = the expected frequency , r= no. of rows, c = no. of columns.

5. **Determine the degrees of freedom -**

   Calculate degrees of freedom (df) as (no. of rows – 1)*(no. of columns – 1).

6. **Determine the critical value –**

   Determine the critical value of chi-square from a chi-square distribution table using the degrees of freedom and the desired level of significance (usually o.o5).

7. **Compare the calculated chi-square statistic with the critical value –**

   If the calculated chi-square statistic is greater than the critical value, reject the null hypothesis and conclude that there is a significant association between the two attributes. If the calculated chi-square statistic is less than the critical value, fail to reject the null hypothesis and conclude that there is no significant association between the two attributes.

   In summary, the Pearson's chi-square test can be used to determine whether a factor is significant or not by analysis the association between two categorical variables.

- **Pearson's chi-Square Test in R-**

To perform a chi-square test in R we can use the

"chisq.test(.)" function.

The output includes the chi-square statistic, degrees of freedom (df), and p-value.

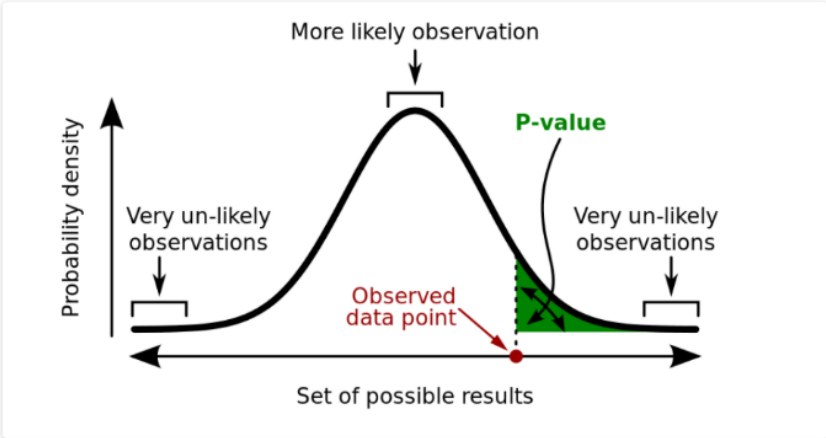
- **P-Value -** In statistics, the p-value refers to the probability of obtaining a test statistic as extreme or more extreme then the observed value, assuming that the null hypothesis is true.

The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favour of the alternative hypothesis.

We Know that P-value is a statistical measure, that helps to determine whether the hypothesis is correct or not. P-value is a number that lies between 0 and 1. The level of significance (α) is a predefined threshold that should be set by the researcher. It is usually fixed as 0.05. The formula for the calculation for P-value is –

If the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables.

If the p-value is greater than or equal to the significance level, we fail to reject the null hypothesis and conclude that there is not enough evidence to suggest a significant association between the variables.



logistic regression –

The logistic model (logit model) is a statistical model that models probability of an event taking place by having the log odds for the event be a linear combination variable.

Logistic regression using the logit function theory is a statistical method used to model the relationship between a binary response variable and predictor variables. The logit function is a mathematical function used to model the relationship between the probability of the response variable being 1 and the predictor variables.

The logit function is defined as the natural logarithm of the odds of the response variable being 1, given the values of the predictor variables. Mathematically, the logit function can be expressed as:

$\text{logit}(p) = \ln(p / (1 - p))$,

where p is the probability of the response variable being 1.

In multiple logistic regression using the logit function theory, the log odds of the response variable being 1 is model as a linear combination of the predictor variables. Mathematically, the model can be expressed as:

$\text{logit}(p) = \alpha + \beta1*x1 + \beta2*x2 + ... + \beta k*xk$,

where α is the intercept term, $\beta1$, $\beta2$, ..., βk are the coefficients of the predictor variables x1, x2, ..., xk, respectively.

The coefficients of the model are estimated using maximum likelihood estimation, which involves finding the set of coefficients that maximize the likelihood of the observed data given the model. The significance of each predictor variable is determined using hypothesis testing, which involves testing whether the coefficient of each predictor variable is significantly different from zero.

Multiple logistic regression using the logit function theory is a powerful tool for analysing data with a binary response variable and multiple predictor variables. It allows researchers to model the relationship between the response variable and the predictor variables while controlling for other relevant factors.

- **Types of Logistic Regression –**

- **Binary logistic regression**: The dependent variable has only two possible outcome/classes.
  Ex.: Yes/No.

- **Multinomial logistic regression:** The dependent variable has only three or more possible outcome/classes without ordering.
  Ex.: Red, Green, Blue

- **Ordinal logistic regression:** The dependent variable has only three or more possible outcome/classes with ordering.
  Ex.: Movie rating 1 to 5.

Multinomial logistic regression

Multinomial logistic regression is a statistical method used for predicting the probability of different categories of a categorical dependent variable based on one or more independent variables. It's an extension of binary logistic regression, which is used for binary classification problems.

Here's a brief overview of how multinomial logistic regression works:

- **Data Preparation:** We need a dataset with a categorical dependent variable and one or more independent variables (continuous or categorical)

- **Model Setup:** The multinomial logistic regression model estimates the relationship between the independent variables and the probabilities of each category of the dependent variable. It uses multiple sets of coefficients, one for each category.

- **Model Fitting:** The algorithm fits the coefficients of the model using a method like maximum likelihood estimation.

- **Interpretation:** After fitting the model, you can interpret the coefficients to understand the effect of each independent variable on the probabilities of the different categories.

- **Prediction:** You can use the fitted model to predict the probabilities or categories for new observations.

Multinomial logistic regression in R

In R, you can perform multinomial logistic regression using the "multinom" function from the "nnet" package.

RESULTS & ANALYSIS

Output of chi square independence test --

Here, we want to test the above data. So we conduct the test between lung cancer and causal factors.

❖ Age & Level –

The following 3*3 contingency table (Age*Level) given below -

Lung cancer Age	Low	Medium	High	Total
Under 30	128	69	104	301
30-50	132	212	221	565
Above 50	43	51	40	134
Total	303	332	365	1000

❑ Result –

Pearson's Chi-squared test

Data: age_ table

X-squared = 40.663, df = 4, p-value = 3.156e-08

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Age" and "Level"). So we conclude that Age is significant on Level.

❖ Gender & Level –

The following 2*3 contingency table (Gender*Level) given below -

Level Gender	Low	Medium	High	Total
Male	149	197	252	598
Female	154	135	113	402
Total	303	332	365	1000

- **Result –**

Pearson's Chi-squared test

Data: Gender_ table

X-squared = 27.225, df = 2, p-value = 1.225e-06

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Gender" and "Level"). So we conclude that Gender is significant on Level.

- **Air Pollution & Level –**

The following 3*3 contingency table (Air Pollution*Level) given below –

Lung Cancer Air Pollution	Low	Medium	High	Total
Low	263	232	20	515
Medium	40	100	296	436
High	0	0	49	49
Total	303	332	365	1000

- **Result –**

Pearson's Chi-squared test

Data: AirPollution_ table

X-squared = 526.19, df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Air Pollution" and "Level"). So we conclude that Air Pollution is significant on Level.

- ❖ **Alcohol use & Level –**

The following 3*3 contingency table (Alcohol Use*Level) given below –

Lung Cancer Alcohol Use	Low	Medium	High	Total
Low	272	162	0	434
Medium	21	100	90	211
High	10	70	275	355
Total	303	332	365	1000

- ❑ **Result –**

Pearson's Chi-squared test

Data: AlcoholUse_ table

X-squared = 625.71, df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Alcohol use" and "Level"). So we conclude that Alcohol use is significant on Level.

- ❖ **Dust Allergy & Level –**

The following 3*3 contingency table (Dust Allergy*Level) given below –

Lung Cancer Dust Allergy	Low	Medium	High	Total
Low	180	41	10	231
Medium	113	161	80	354
High	10	130	275	365
Total	303	332	365	1000

- **Result –**

Pearson's Chi-squared test

Data: DustAllergy_ table

X-squared = 497.85,df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Dust Allergy" and "Level"). So we conclude that Dust Allergy is significant on Level.

- **Occupational Hazards & Level –**

The following 3*3 contingency table (Occupational Hazards *Level) given below –

Lung Cancer Occupational Hazards	Low	Medium	High	Total
Low	202	121	10	333
Medium	81	111	80	272
High	20	100	275	395
Total	303	332	365	1000

❑ **Result –**

Pearson's Chi-squared test

Data: OccupationalHazards _ table

X-squared = 422.3, df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Occupational Hazards" and "Level"). So we conclude the Occupational Hazards t is significant on Level.

❖ **Genetic Risk & Level –**

The following 3*3 contingency table (Genetic Risk*Level) given below –

Lung Cancer Genetic Risk	Low	Medium	High	Total
Low	161	61	0	222
Medium	112	131	80	323
High	30	140	285	455
Total	303	332	365	1000

❑ **Result –**

Pearson's Chi-squared test

Data: GeneticRisk _ table

X-squared = 385.74, df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Genetic Risk" and "Level"). So we conclude that Genetic Risk is significant on Level.

❖ **Chronic Lung Disease & Level –**

The following 3*3 contingency table (Chronic Lung Disease*Level) given below –

Lung Cancer Chronic Disease	Low	Medium	High	Total
Low	132	81	0	213
Medium	141	131	80	352
High	30	120	285	408
Total	303	332	365	1000

- **Result –**

Pearson's Chi-squared test

Data: ChronicLungDisease_ table

X-squared = 380.8, df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Chronic Lung Disease" and "Level"). So we conclude that Chronic Lung Disease is significant on Level.

- **Balanced Diet & Level –**

The following 3*3 contingency table (Balanced Diet*Level) given below –

Lung Cancer Balanced Diet	Low	Medium	High	Total
Low	141	120	0	261
Medium	142	142	10	294
High	20	70	355	445
Total	303	332	365	1000

❑ **Result –**

Pearson's Chi-squared test

Data: BalancedDiet _ table

X-squared = 641.56 , df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Balanced Diet" and "Level"). So we conclude that Balanced Diet is significant on Level.

❖ **Obesity & Level –**

The following 3*3 contingency table (Obesity*Level) given below –

Lung Cancer Obesity	Low	Medium	High	Total
Low	170	40	0	210
Medium	133	242	29	404
High	0	50	326	376
Total	303	332	365	1000

❑ **Result –**

Pearson's Chi-squared test

Data: Obesity _ table

X-squared = 884.79 , df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Obesity" and "Level"). So we conclude that Obesity is significant on Level.

❖ **Smoking & Level –**

The following 3*3 contingency table (Smoking*Level) given below –

Lung Cancer Smoking	Low	Medium	High	Total
Low	213	292	70	575
Medium	70	30	29	129
High	20	10	266	296
Total	303	332	365	1000

- **Result –**

Pearson's Chi-squared test

Data: Smoking _ table

X-squared = 555.01, df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Smoking" and "Level"). So we conclude that Smoking is significant on Level.

- **Passive Smoker & Level –**

The following 3*3 contingency table (Passive Smoker*Level) given below –

Lung Cancer Passive Smoker	Low	Medium	High	Total
Low	212	202	70	484
Medium	91	130	0	221
High	0	0	295	295
Total	303	332	365	1000

❑ **Result –**

Pearson's Chi-squared test

Data: PassiveSmoker _ table

X-squared = 750.35, df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Passive Smoker" and "Level"). So we conclude that is significant on Level.

❖ **Chest Pain & Level –**

The following 3*3 contingency table (Chest pain*Level) given below –

Lung Cancer Chest Pain	Low	Medium	High	Total
Low	222	182	10	414
Medium	51	120	70	241
High	30	30	285	345
Total	303	332	365	1000

❑ **Result –**

Pearson's Chi-squared test

Data: Chest Pain _ table

X-squared = 567.33,df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("*Chest pain*" and "Level"). So we conclude that *Chest pain* is significant on Level.

❖ **Coughing of Blood & Level –**

The following 3*3 contingency table (Coughing of Blood*Level) given below –

Lung Cancer Coughing of Blood	Low	Medium	High	Total
Low	192	161	10	363
Medium	101	141	29	271
High	10	30	326	366
Total	303	332	365	1000

- **Result –**

Pearson's Chi-squared test

Data: CoughingofBlood _ table

X-squared = 708.55,df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Coughing of Blood" and "Level"). So we conclude that Coughing of Blood is significant on Level.

- **Fatigue & Level –**

The following 3*3 contingency table (Fatigue*Level) given below –

Lung Cancer Fatigue	Low	Medium	High	Total
Low	283	160	90	533
Medium	20	172	127	319
High	0	0	148	148
Total	303	332	365	1000

❑ **Result –**

Pearson's Chi-squared test

Data: Fatigue _ table

X-squared =509.69, df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Fatigue" and "Level"). So we conclude that Fatigue is significant on Level.

❖ **Weight Loss & Level –**

*The following 3*3 contingency table (Weight Loss*Level) given below*

Lung Cancer Weight Loss	Low	Medium	High	Total
Low	242	151	158	551
Medium	51	61	97	209
High	10	120	110	240
Total	303	332	365	1000

❑ **Result –**

Pearson's Chi-squared test

Data: WeightLoss _ table

X-squared = 137.82, df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("*Weight Loss*" and "Level"). So we conclude that *Weight Loss* is significant on Level.

❖ **Shortness of Breath & Level –**

The following 3*3 contingency table (Shortness of Breath*Level) given below –

Lung Cancer Shortness of Breath	Low	Medium	High	Total
Low	263	121	79	463
Medium	30	171	177	378
High	10	40	109	159
Total	303	332	365	1000

- **Result –**

Pearson's Chi-squared test

Data: Shortness of Breath _ table

X-squared = 330.26, df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Shortness of Breath" and "Level"). So we conclude that Shortness of Breath is significant on Level.

- **Wheezing & Level –**

The following 3*3 contingency table (Wheezing*Level) given below –

Lung Cancer Wheezing	Low	Medium	High	Total
Low	221	30	198	449
Medium	82	262	58	402
High		0	109	149
Total	303	332	365	1000

❑ **Result –**

Pearson's Chi-squared test

Data: Wheezing _ table

X-squared = 447.52, df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Wheezing" and "Level"). So we conclude that Wheezing is significant on Level.

❖ **Swallowing Difficulty & Level –**

The following 3*3 contingency table (Swallowing Difficulty*Level) given below –

Lung Cancer Swallowing Difficulty	Low	Medium	High	Total
Low	223	120	128	471
Medium	70	162	158	390
High	10	50	79	139
Total	303	332	365	1ooo

❑ **Result –**

Pearson's Chi-squared test

Data: Swallowing Difficulty _ table

X-squared = 134.93, df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Swallowing Difficulty" and "Level"). So we conclude that Swallowing Difficulty is significant on Level.

❖ **Clubbing of Finger Nails**

The following 3*3 contingency table(Clubbing of Finger Nails *Level) given below –

Lung Cancer Clubbing of Finger Nails	Low	Medium	High	Total
Low	242	110	119	471
Medium	61	112	177	350
High	0	110	69	179
Total	303	332	365	1000

- **Result –**

Pearson's Chi-squared test

Data: ClubbingofFingerNails _ table

X-squared = 234.87, df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Clubbing of Finger Nails" and "Level"). So we conclude that Clubbing of Finger Nails is significant on Level.

- **Frequent Cold & Level –**

The following 3*3 contingency table (Frequent Cold*Level) given below –

Lung Cancer Frequent Cold	Low	Medium	High	Total
Low	192	80	59	331
Medium	111	161	158	430
High	0	91	148	239
Total	303	332	365	1000

- **Result –**

Pearson's Chi-squared test

Data: FrequentCold _ table

X-squared = 245.26, df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Frequent Cold" and "Level"). So we conclude that Frequent Cold is significant on Level.

- **Dry Cough & Level –**

The following 3*3 contingency table (Dry Cough*Level) given below –

Lung Cancer Dry Cough	Low	Medium	High	Total
Low	130	141	99	370
Medium	152	141	80	373
High	21	50	186	257
Total	303	332	365	1000

- **Result –**

Pearson's Chi-squared test

Data: DryCough _ table

X-squared = 200.91, df = 4, p-value = < 2.2e-16

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Dry Cough" and "Level"). So we conclude that Dry Cough is significant on Level.

- **Snoring & Level –**

The following 3*3 contingency table (Snoring*Level) given below –

Lung Cancer(y) Snoring(x23)	Low	Medium	High	Total
Low	201	130	139	470
Medium	102	182	197	481
High	0	20	29	49
Total	303	332	365	1000

- **Result –**

Pearson's Chi-squared test

data: Snoring_table

X-squared = 76.1, df = 4, p-value = 1.166e-15

Here, the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables ("Snoring" and "Level"). So we conclude that Snoring is significant on Level.

Output of multinomial logistic regression –

The causal factor, communication with lung cancer is not a ordered categorical variable so it should be convert into dummy variable .

The p value of each independent factor is given by

	(Intercept)	x1	x2	x3	x4	x5
1	0.001061244	1.955415e-05	0.001911885	0.0011860818	0.0045544797	1.187698e-04
2	0.001689130	4.891205e-05	0.001786964	0.0002707725	0.0006477881	7.771153e-05

	(Intercept)	x6	x7	x8	x9	x10	x11
1	0.0001225827	0.0010782939	7.671586e-06	0.0016150967	0.007576396	0.0001229986	
2	0.0013288355	0.0007008733	7.292348e-04	0.0002397678	0.002743252	0.0012177431	

	(Intercept)	x12	x13	x14	x15	x16	x17
1	0.0011962638	0.0006051735	0.0018550685	0.0008937841	0.0002494271	8.160686e-05	
2	0.0003213613	0.0010240603	0.0009219743	0.0001791994	0.0009723011	6.324883e-04	

	(Intercept)	x18	x19	x20	x21	x22	x23
1	0.0018375852	0.0002716485	0.0005322034	0.0008829789	0.0007056941	0.005027971	
2	0.0006752824	0.0005254020	0.0008356506	0.0009820695	0.0002362010	0.004647041	

CONCLUSION

From above study, we can state two conclusions.

Frist one is from chi-square independence test. Hence we can say all causal factor (like-smoking, air pollution, etc.) are significant for lung cancer.

And the last one is form multinomial logistic regression. Here all the p-values are less than 0.05. So, all the causal factors are significant with lung cancer.

REFERENCE

For this Project, I collected data from various sources and the modeling I have taken help from many resources. They are

1. https://www.kaggle.com/datasets.
2. ”SK sir” note
3. Wikipedia
4. An Introduction to Categorical Data Analysis by ALAN AGRESTI ,Department of Statistics University of Florida Gainesville, Florida, WILEY-INTERSCIENCE, A JOHN WILEY & SONS ,INC.,PUBLICATION

APPENDIX

For chi-square test ::

```
#chi-square test

airpollution_table<- matrix(c(263,40,0,232,100,0,20,296,49),nrow=3)

airpollution_table

# Conduct Pearson's chi-square test

chisq.test(airpollution_table)


AlcoholUse_table<- matrix(c(272,21,10,162,100,70,0,90,275),nrow=3)

AlcoholUse_table

# Conduct Pearson's chi-square test

chisq.test(AlcoholUse_table)


DustAllergy_table<- matrix(c(180,113,10,41,161,130,10,80,275),nrow=3)

DustAllergy_table

# Conduct Pearson's chi-square test

chisq.test(DustAllergy_table)


OccupationalHazards_table<- matrix(c(202,81,20,121,111,100,10,80,275),nrow=3)

OccupationalHazards_table

# Conduct Pearson's chi-square test

chisq.test(OccupationalHazards_table)
```

```
GeneticRisk_table<- matrix(c(161,112,30,61,121,120,0,80,285),nrow=3)

GeneticRisk_table

# Conduct Pearson's chi-square test

chisq.test(GeneticRisk_table)


ChronicLungDisease_table<- matrix(c(132,141,30,91,141,130,0,80,285),nrow=3)

ChronicLungDisease_table

# Conduct Pearson's chi-square test

chisq.test(ChronicLungDisease_table)


BalancedDiet_table<- matrix(c(141,142,20,130,152,80,0,10,355),nrow=3)

BalancedDiet_table

# Conduct Pearson's chi-square test

chisq.test(BalancedDiet_table)


Obesity_table<- matrix(c(170,133,0,40,242,50,0,29,326),nrow=3)

Obesity_table

# Conduct Pearson's chi-square test

chisq.test(Obesity_table)


Smoking_table<- matrix(c(213,70,20,292,30,10,70,29,266),nrow=3)

Smoking_table
```

```
# Conduct Pearson's chi-square test

chisq.test(Smoking_table)


PassiveSmoker_table<- matrix(c(212,91,0,202,130,0,70,0,295),nrow=3)

PassiveSmoker_table

# Conduct Pearson's chi-square test

chisq.test(PassiveSmoker_table)


ChestPain_table<- matrix(c(222,51,30,182,120,30,10,70,285),nrow=3)

ChestPain_table

# Conduct Pearson's chi-square test

chisq.test(ChestPain_table)


CoughingofBlood_table<- matrix(c(192,101,10,161,141,30,10,29,326),nrow=3)

CoughingofBlood_table

# Conduct Pearson's chi-square test

chisq.test(CoughingofBlood_table)


Fatigue_table<- matrix(c(283,20,0,160,172,0,90,127,148),nrow=3)

Fatigue_table

# Conduct Pearson's chi-square test

chisq.test(Fatigue_table)
```

```
WeightLoss_table<- matrix(c(242,51,10,151,61,120,158,97,110),nrow=3)

WeightLoss_table

# Conduct Pearson's chi-square test

chisq.test(WeightLoss_table)


ShortnessofBreath_table<- matrix(c(263,30,10,121,171,40,79,177,109),nrow=3)

ShortnessofBreath_table

# Conduct Pearson's chi-square test

chisq.test(ShortnessofBreath_table)


Wheezing_table<- matrix(c(221,82,0,30,262,40,198,58,109),nrow=3)

Wheezing_table

# Conduct Pearson's chi-square test

chisq.test(Wheezing_table)


SwallowingDifficulty_table<- matrix(c(223,70,10,120,162,50,128,158,79),nrow=3)

SwallowingDifficulty_table

# Conduct Pearson's chi-square test

chisq.test(SwallowingDifficulty_table)


ClubbingofFingerNails_table<- matrix(c(242,61,0,110,112,110,119,177,69),nrow=3)

ClubbingofFingerNails_table

# Conduct Pearson's chi-square test
```

```
chisq.test(ClubbingofFingerNails_table)


FrequentCold_table<- matrix(c(192,111,0,80,161,91,59,158,148),nrow=3)

FrequentCold_table

# Conduct Pearson's chi-square test

chisq.test(FrequentCold_table)


DryCough_table<- matrix(c(130,152,21,141,141,50,99,80,186),nrow=3)

DryCough_table

# Conduct Pearson's chi-square test

chisq.test(DryCough_table)


Snoring_table<- matrix(c(201,102,0,130,182,20,139,197,29),nrow=3)

Snoring_table

# Conduct Pearson's chi-square test

chisq.test(Snoring_table)
```

OR

```
#chi-square test

setwd('C:\\Users\\DELL\\Desktop\\project')

df <-read.csv('cancer patient data sets -final version.csv')

chisq.test(df$y,df$x1)

chisq.test(df$y,df$x2)
```

```
chisq.test(df$y,df$x3)

chisq.test(df$y,df$x4)

chisq.test(df$y,df$x5)

chisq.test(df$y,df$x6)

chisq.test(df$y,df$x7)

chisq.test(df$y,df$x8)

chisq.test(df$y,df$x9)

chisq.test(df$y,df$x10)

chisq.test(df$y,df$x11)

chisq.test(df$y,df$x12)

chisq.test(df$y,df$x13)

chisq.test(df$y,df$x14)

chisq.test(df$y,df$x15)

chisq.test(df$y,df$x16)

chisq.test(df$y,df$x17)

chisq.test(df$y,df$x18)

chisq.test(df$y,df$x19)

chisq.test(df$y,df$x20)

chisq.test(df$y,df$x21)

chisq.test(df$y,df$x22)

chisq.test(df$y,df$x23)
```

For multinomial logistic regression ::

```
#multinomial logit model
```

```
setwd('C:\\Users\\DELL\\Desktop\\project')

df <-read.csv('cancer patient data sets.csv')

head(df)

#logistic model

library (nnet)

model=multinom(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+x15+x16+x
17+x18+x19+x20+x21+x22+x23,order=TRUE,,data=df)

summary(model)

z<-summary(model)$coefficients/summary(model)$standard.errors

p<-(1-pnorm(abs(z),0,1))*2

ynew<-predict(model,newdata=df,type='class')

#confusion matrix

t<-table(df$y,ynew)

sum(diag(t))/sum(t)

z

p
```