WOMEN HEALTH DATA

**TABLE OF CONTENTS**

1. EXECUTIVE SUMMARY
2. FACTORS CONSIDERED FOR ANALYSIS
3. ANALYSIS CONDUCTED
4. ASSUMPTIONS
5. PAIRED SAMPLE TEST
6. CORRELATION ANALYSIS
7. MULTIPLE REGRESSION
8. LOGISTIC REGRESSION
9. TEST FOR INDEPENDENCE
10. MODERATION ANALYSIS
11. **EXECUTIVE SUMMARY**

**“IT COSTS MONEY TO BE HEALTHY BUT IT’S EVEN MORE EXPENSIVE TO GET SICK”**

Health is a fundamental aspect of quality of life, not only because being free from illness or injury directly affects our capacity to enjoy life, but also because health indirectly affects our capacity to produce and consume other valuable goods and services. Whilst the Global Burden of Disease (GBD) assessment assigns each death to one specific cause, we know that the risk of disease burden and health outcomes are closely linked to several risk factors. It includes risk factors across four broad categories: **behavioral, environmental, occupational, and metabolic risks**. High BMI, for example, may likely be present with other lifestyle factors such as low physical activity levels, high blood pressure, low fruit and vegetable intake, environmental factors including clean water and sanitation. As women represent 51% of total population and study reveals not only women suffer from gender specific disease but also they are diagnosed with disease which were previously considered to have affect only on men, and as a result we decided to explore some factors contributing to health of women. It is also true that BMI for people residing in rural areas differ from BMI for people residing in urban areas. There are many factors which in correlation can be responsible for degrading the quality of life and which can have impact on average life expectancy of women. Life Expectancy can help us determine the health of population of country which can help us to analyze various factors such as pollution, alcohol consumption, protein and calorie intake for that country. Life Expectancy is measure of premature death and differs all throughout the globe. Certain Factors if not directly but in indirect way can certainly hamper the quality of life such as physical activity, Depression and Anxiety. We also determine factors that can increase chances of non-communicable diseases and that how pollution is one of largest health and environmental problems.

In brief, all these factors related to behavioral, environmental, occupational and rate of metabolism are responsible for contributing towards major life crises. In this project we decide to explore some of factors to understand the relation between them and their impact on health. It will help us to analyze the areas which can be controlled for healthier and happy living, Since the greatest wealth is good health and been healthy is not a goal it is way of living.

1. **FACTORS CONSIDERED FOR ANALYSIS**

We have performed analysis on data pertaining to Women’s Health across 100 countries for the year 2015. This data is taken from https://ourworldindata.org/. Parameters considered for data set are:

1. Underweight: Percentage of women who have inability to acquire enough food defined as a level of food intake insufficient to meet dietary energy requirements.
2. Healthy: Percentage of women who are healthy according to Food and Agriculture organization based on caloric intake if sufficient to meet their daily energy requirements
3. Overweight or Obesity: Percentage of women who are obese with high body mass index
4. Total Fertility: Number of births per women in population, i.e live births per woman
5. Breast Cancer: Number of Women diagnosed with breast cancer for every 100,000 women in the year 2015.
6. Cervical Cancer: Number of Women diagnosed with cervical cancer for every 100,000 women.
7. Pollution: Air pollution in that country. This is a categorical variable with 3 categories ‘Low, Moderate and Severe’.
8. Prevalence of Alcohol use disorder: Percentage of women who are addicted to alcohol
9. Anemia: Percentage of women with hemoglobin level less than grams per liter at sea level
10. Vitamin A Deficiency: Percentage of women who have Vitamin A Deficiency whose semi retinol levels are below key indicator
11. Prevalence of Smoking Disorder: Percentage of Women who are addicted to smoking
12. Life expectancy in years: Number of years a woman is expected to live. This is a binary categorical variable with categories low and good.
13. Depression: Percentage of Women suffering from depression based on medical,  
    epidemiological data, surveys and meta-regression modeling.
14. Women Deaths During Childbirth: This is a categorical variable with high, low and moderate categories.
15. Prevalence of Low nutrition: Percentage of women who have prolonged low nutrition intake
16. Deaths due to Poor Sanitation: Women who die prematurely due to poor sanitation condition in every country
17. Low Physical Activity: Percentage of women who do not exercise often
18. High Blood Sugar (deaths): Share of women with high blood Sugar in 100,000 individuals
19. High Blood Pressure (deaths): Share of women with high blood Pressure in 100,000 individuals
20. Mental disorder: share of women with mental health or substance use disorder; this includes depression, anxiety, bipolar, eating disorders, alcohol or drug use disorders, and schizophrenia
21. Drug use disorder: Percentage of women with prevalence of usage of drugs.
22. Mean BMI for women in Urban areas based on body mass index
23. Mean BMI for Women in Rural areas based on body mass index
24. Death rate for non-communicable disease in women: Age-standardized DALY (Disability-Adjusted Life Year) rates per 100,000 individuals from non-communicable diseases (NCDs): cardiovascular disease, cancer, respiratory disease, digestive disease, liver disease, neurological disease,
25. Mean BMI
26. Daily Caloric Intake: Average daily per capita caloric supply in kilocalories per woman
27. Daily Protein Intake: Average daily per capita protein supply in grams per woman
28. **ANALYSIS CONDUCTED**

We performed various tests to analyze inter-relationships between variables and how some variables can help predict the outcome of other variables. We conducted following tests:

1. Paired Sample test: To determine if there is a significant difference between Mean BMI of Urban and Rural areas.
2. Correlation between:  
   Fertility and Obesity

Mental disorder, daily caloric with Depression,

Smoking, Low nutrition intake and non-communicable disease

1. Multiple Regression: To determine if Smoking, Obesity and prevalence of alcohol predict non communicable diseases
2. Logistic Regression: To determine if Anemia, Pollution and Prevalence of low nutrition predicts low life expectancy
3. Test of Independence:

Pollution and Life Expectancy

Pollution and Women deaths during child birth

1. Moderation Analysis:

Does Life expectancy and smoking impact depression,

Does BMI

1. **ASSUMPTIONS**

* Significance level of 0.05 level is considered for all the analysis. α = 0.05
* Life Expectancy: if below or equal to 65 years than it belongs to low category, if more than 65 years than category for life expectancy Is good.
* Women Death during Childbirth: <0.1 is low, 0.1-moderate, 1.0-5.0-high, >5.0-very high
* Pollution: Severe, low, moderate

1. **PAIRED SAMPLE TEST**

**Is there a significant difference between the Mean BMI of urban and rural areas?**

The data for this context corresponds to the Mean BMI of urban and rural areas that belongs to a particular country. It becomes paired data and hence Paired Sample test is conducted to perform the analysis.

The hypothesis for this context is set as follows:

: = 0

: ≠ 0

Here, the difference d is calculated as the difference between Mean BMI of urban and rural areas.

Here,

n = 98

= Average of differences = 0.6774

= Standard Dev for = 0.811

For Paired test, the ‘t’ value is calculated as follows:

t = ( – ) / (/) = 8.268

This is a two-tailed test,

Therefore, P = [1-T.Dist(t, n-1, TRUE)]\*2

As the value of t is high, P-value will be very less.

Here, P < α which means that we REJECT THE NULL

Conclusion: Hence, there is supporting evidence to show that there is a significant difference between the Mean BMI of urban and rural areas.

1. **CORRELATION ANALYSIS**
2. **Correlation between Fertility rate and Obesity**

Hypothesis,

: ρ = 0

: ρ ≠ 0

r = Sample correlation coefficient = CORREL(fertility rate, obesity)

r = -0.2285

n = 98

t-value for correlation analysis is,

t = r

t = -2.2996

For a two-tailed test,

P = [T.DIST(t, n-1, TRUE)]\*2

= 0.02333

Here, P < α which means that we REJECT THE NULL

Conclusion: Hence, there is supporting evidence to prove that there exists a significant correlation between Fertility Rate and Obesity.

1. **Correlation of mental disorder, daily caloric intake with depression**

Hypothesis,

: ρ = 0

: ρ ≠ 0

r = Sample correlation coefficient = CORREL(depression, mental disorder),

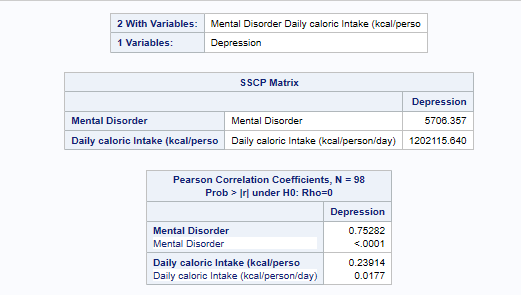
CORREL(Depression, daily caloric intake)

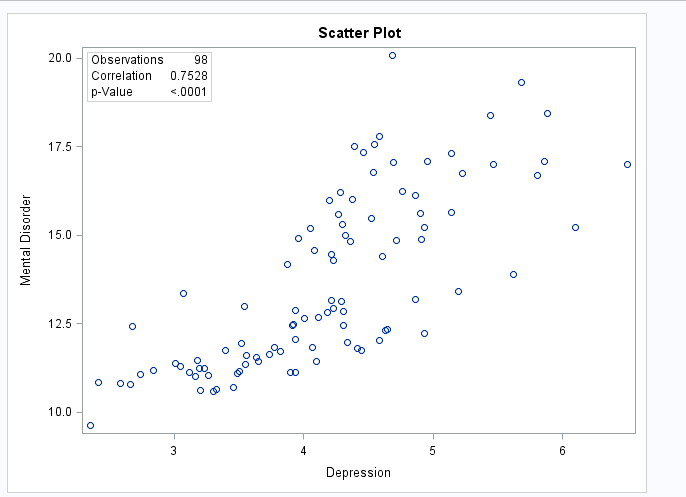
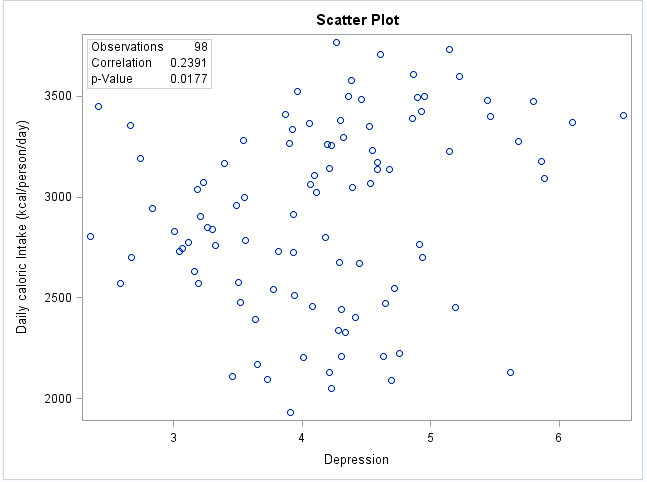
r = Mental disorder and depression = 0.7528

r = Depression with Low caloric intake: 0.239

n = 98  
This is a two-tailed test.

Analysis is conducted using SAS and below is the output:





Here, P < α which means that we REJECT THE NULL.

Conclusion: Hence, there is supporting evidence to prove that there exists correlation between Depression, Mental Disorder and Depression, Low Caloric Intake.

1. **Correlation of Smoking, Low nutrition intake with non-communicable disease**

Hypothesis,

: ρ = 0

: ρ ≠ 0

r = Sample correlation coefficient = CORREL(Non-Communicable Diseases, Smoking),

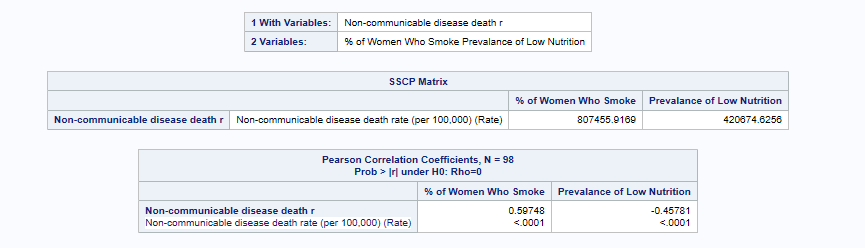
CORREL(Non-Communicable Disease, Low Nutrition intake)

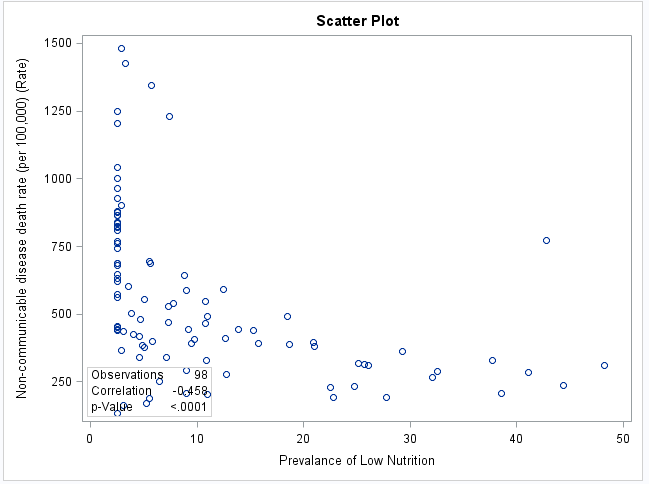
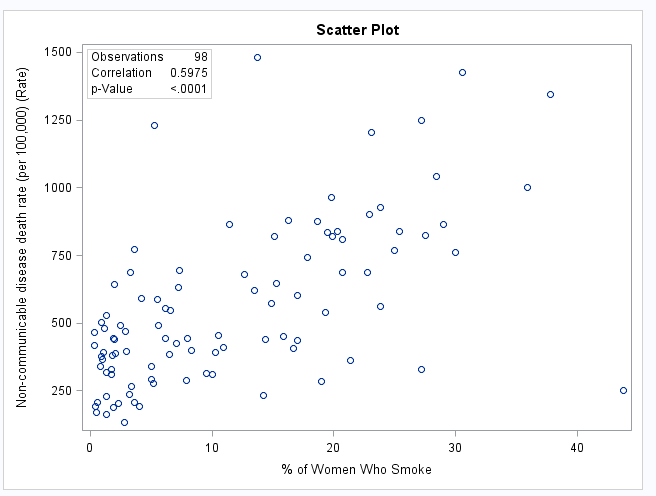
r = smoking and non-communicable disease: 0.597

r = non-communicable disease and low nutrition intake = -0.457

n = 98  
This is a two-tailed test.

Analysis is conducted using SAS and below is the output:





Here, P < α which means that we REJECT THE NULL.

Conclusion: Hence, there is supporting evidence to prove that there exists correlation between smoking, non-communicable diseases and  Low nutrition intake, non-communicable diseases.

1. **MULTIPLE REGRESSION**

**How Prevalence of smoking disorder, Obesity and Prevalence of Alcohol Use effects the non-communicable diseases?**

As all the variables are continuous in nature, we have conducted Multiple Regression analysis.

Attributes that have been considered**:**

Dependent Variable: Non-Communicable diseases

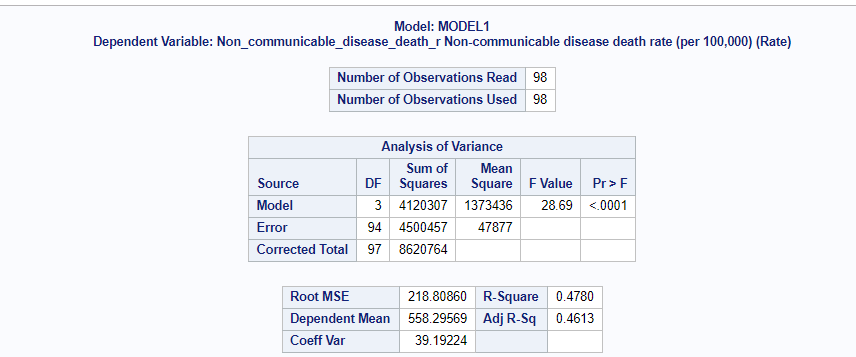
Independent Variables: Prevalence of Smoking disorder

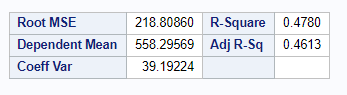
Obesity and

Prevalence of Alcohol use

From Analysis of Variance table, P-value of F-Statistic is < 0.0001.

As P < α we can conclude that the model is significant

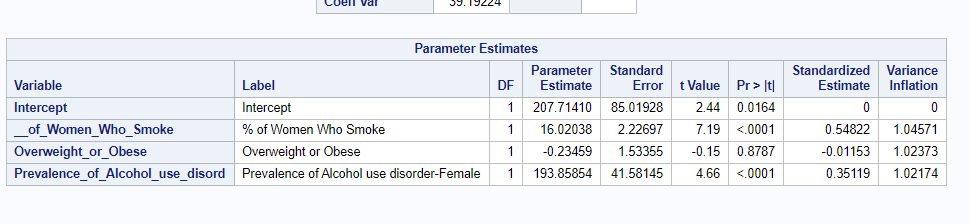




R- Square value infers the entire variation in the dependent variable I.e (Non communicable diseases)

Adj.R-Square=0.4613

This means that 46.13% of variance in Non-communicable diseases can be explained by Smoking Disorder, Obesity and Prevalence of Alcohol Use

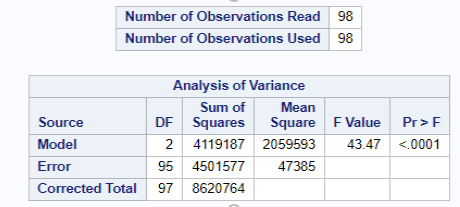


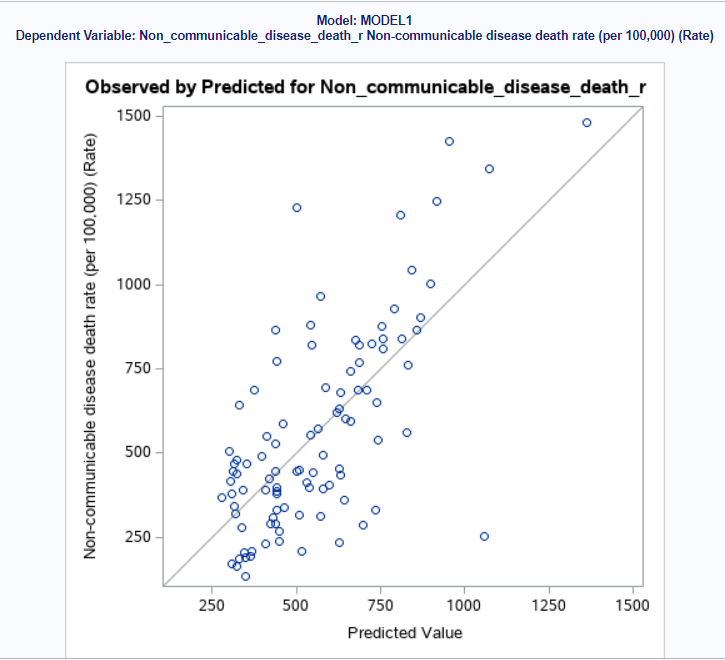
For Smoking disorder and prevalence of alcohol disorder, P < α which means that all these variables predict the dependent variable Non-communicable diseases.

For the factor obesity, P > α which means that this variable donot help in predicting the independent variable non-communicable diseases.

Hence, here we drop the independent variable ‘Obesity’ and rerun the analysis in SAS.

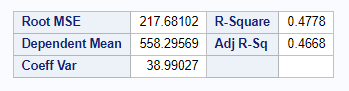
Below are the results:





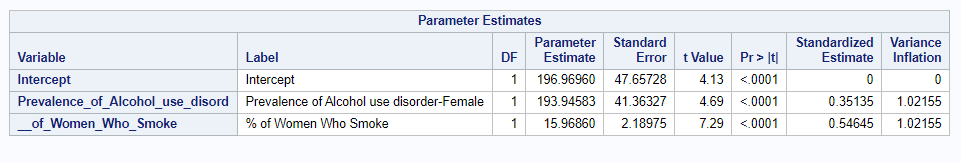
From ANOVA table it is clear that the P-value for the F-statistic is <0.0001

As P < α, we can clearly state that the model is significant.



Adjusted R^2 = 0.4668

This means that 46.68% of variance in Non-communicable diseases can be explained by Smoking Disorder and Prevalence of Alcohol Use



Here, P < α for all the independent variables and hence our model is significant. Also, the variance inflation factor shows that both the variables are not collinear.

Analysis of Intercept: Logically, Non-communicable diseases can have a value if all the independent variables become zero.

Hypothesis for intercept is,

= 4.13 and P < 0.0001.

As P < α, we reject the null and this means that the intercept and has a value.

From the table, = 196.97

Analysis of slopes: Hypothesis for slopes:

Here, n = {1,2} (all the independent variables)

From parameter estimates table, the value of P < α for all the independent variables and hence we reject the null. This means that each of the slopes for independent variables are not equal to zero.

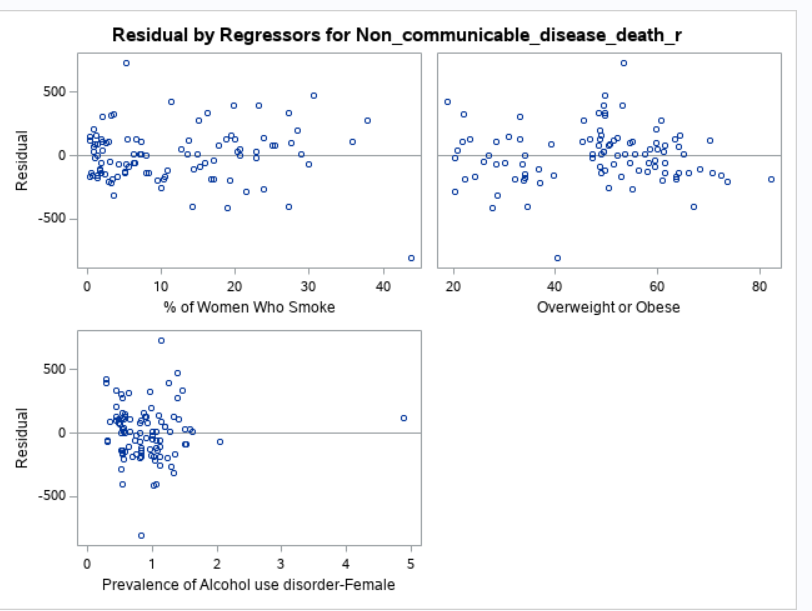
= 193.94

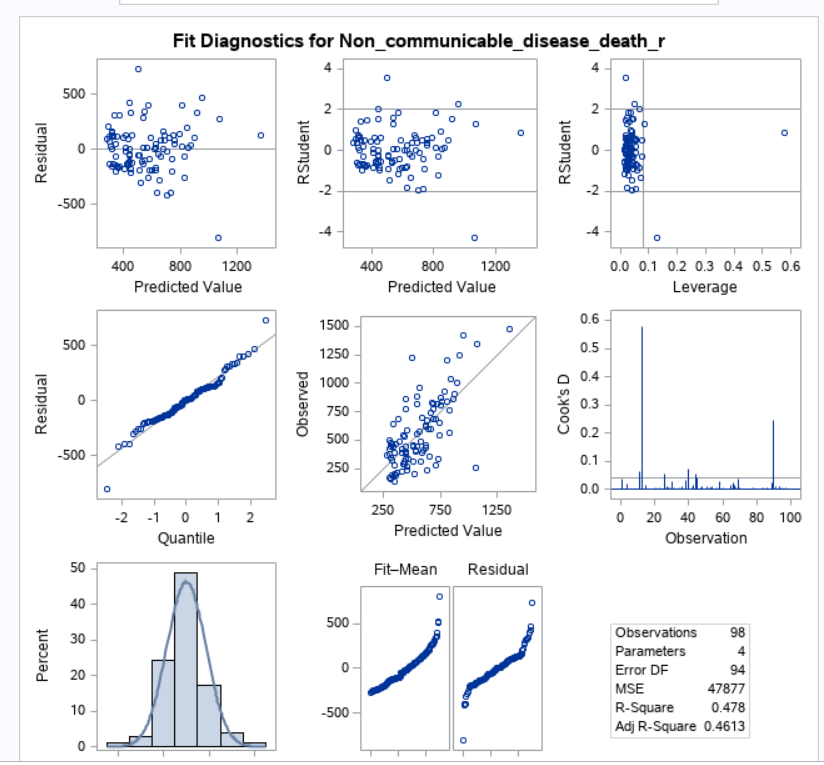
= 15.96

Hence, we obtain the relation between dependent and independent variables as follows:

Non-Communicable Diseases = 196.97 + 193.94 (Prevalence of Alcohol Use) + 15.96 (Smoking Disorder)

Conclusion**:** The model is significant and after careful observation of parameters this to conclude that all the parameters smoking disorder and Prevalence of Alcohol use have a greater impact on establishing the risk of non-communicable diseases.





1. **LOGISTIC REGRESSION**

**Determine whether Anemia, Pollution and Prevalence of low nutrition predicts low life expectancy?**

Hypothesis:

SAS INPUTS:

Response Variable: Life Expectancy

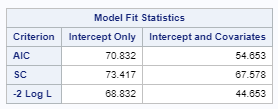
Event of Interest: low

Classification Variables: Pollution

Continuous Variables: Prevalence of low nutrition, Anemia



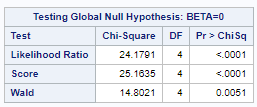
This indicates that 11 out of 98 countries has low life expectancy.





Max-rescaled R-Square = 0.4333

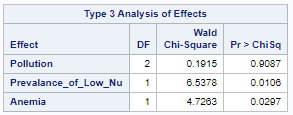
This means that 43.33 % of variance in low life expectancy is predicted by Pollution, Anemia and Prevalence of low nutrition.

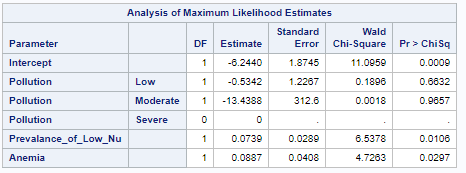


Here,

And P-value of statistic is; P < α

Hence, the model is significant.



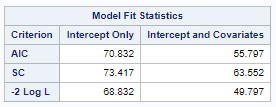


From the above tables,

P < α for ‘Prevalence of Low Nutrition’ and ‘Anemia’ which means that low nutrition and having Anemia predicts low life expectancy.

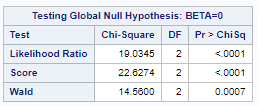
P > α for ‘Pollution’ which means that pollution doesn’t predict low life expectancy.

Hence, we re-run the model removing the pollution variable.

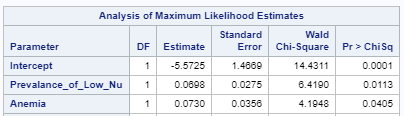




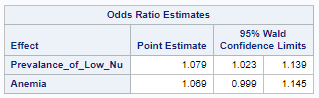
34.99% of variance in low life expectancy is explained by Prevalence of low nutrition and anemia.



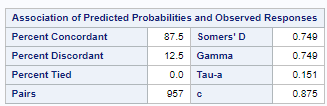
P < α for statistic and hence the model is significant.



P < α for prevalence of low nutrition and Anemia. Hence low nutrition and anemia predicts low life expectancy.



From above table, as prevalence of low nutrition score rises by 1, the log of odds increases by 7.9%. Similarly, if the score of Anemia increases by 1, the log of odds increases by 6.9%.



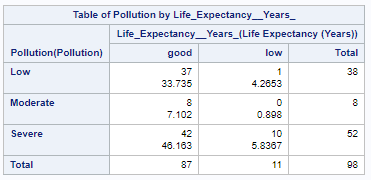
C = 0.875. This means that the model is able to predict whether prevalence of low nutrition and Anemia predicts low life expectancy/not from among 87.5% of rows.

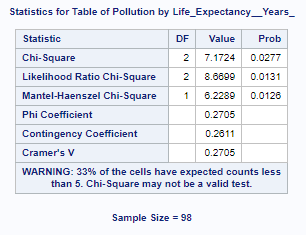
1. **TEST OF INDEPENDENCE:**

**Dependency of Pollution and Life Expectancy**

As both the variables are categorical in nature, test of independence is used.

Hypothesis:

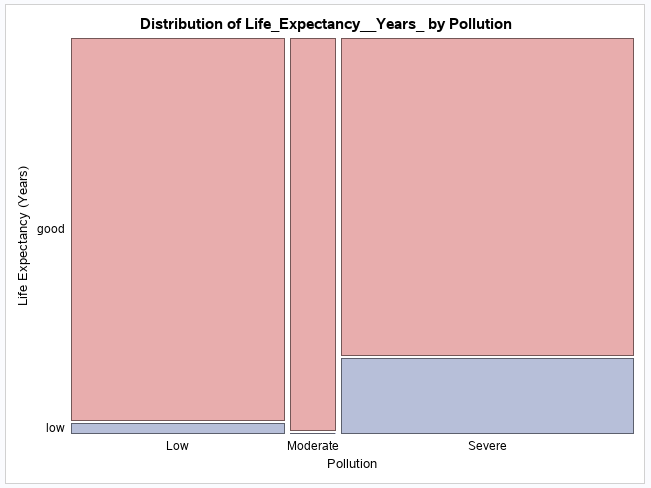




Here, P < α for statistic.

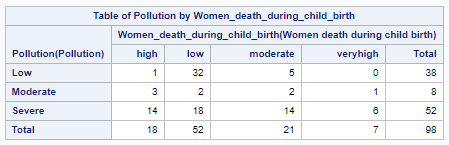
Hence, we REJECT THE NULL.

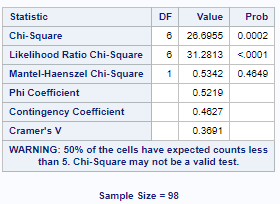
This means that the variables pollution and life expectancy are dependent.



1. **Pollution and Women deaths during child birth**

Hypothesis:

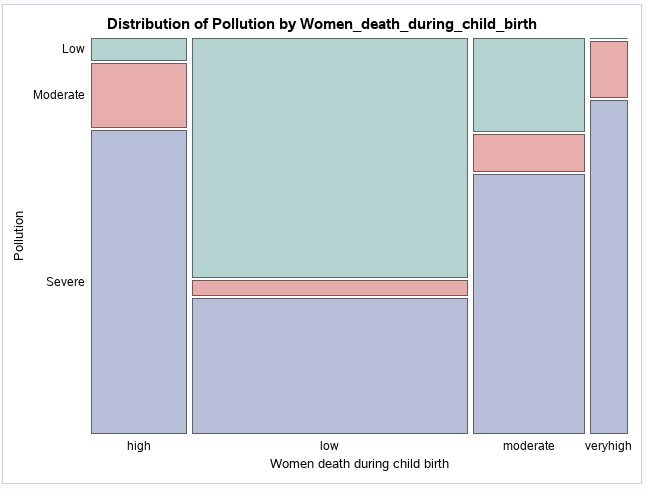




Here, P < α for statistic.

Hence, we REJECT THE NULL.

This means that the variables pollution and women death during child birth are dependent.



1. **MODERATION ANALYSIS:**
2. **To find if there is an interaction effect of Alcohol disorder and Depression on Low Life Expectancy.**

Test performed**:** As the dependent variable is binary categorical in nature, we perform Moderation Analysis with Binary Logistic Regression.

Attributes that have been considered:

Depression, Life Expectancy and Prevalence of Alcohol disorder.

Hypothesis:

From the obtained results we get P-value for statistic is < 0.0001.

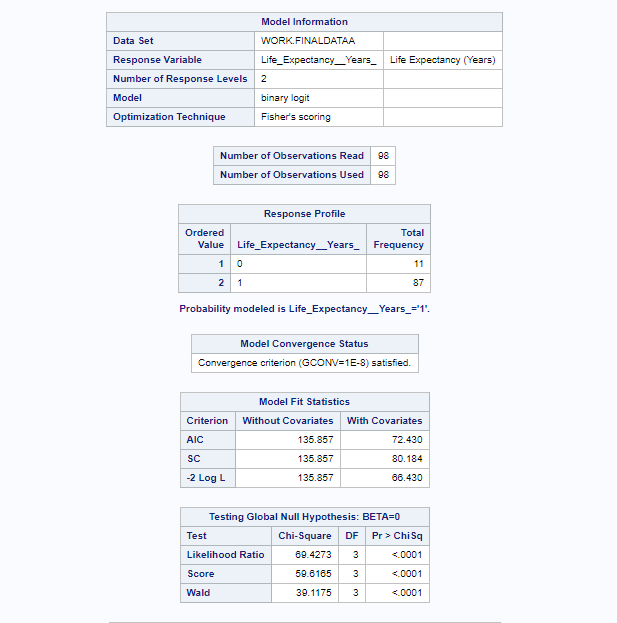
As P < α, the model is significant

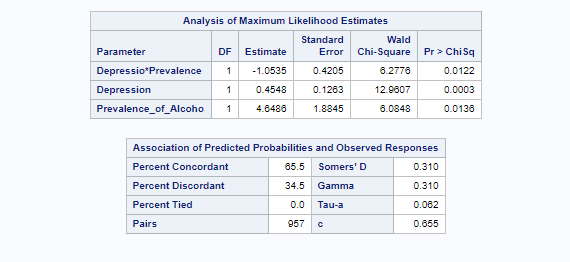
P = 0.0122 for the interaction effect.

As P < α, we can reject the null and say that there is interaction effect on depression and alcohol disorder on low life expectancy.

With c=0.65, This means that the model is able to predict whether the interaction effect of depression and alcohol disorder predicts low life expectancy/not from among 65% of rows.

Inference on hypothesis: Hence, there is supporting evidence to say that prevalence of Alcohol disorder is affecting the impact of Depression on Life Expectancy.





**b. To find if there is an interaction effect of Obesity and Calories intake on BMI.**

Analysis Performed: Moderation Analysis with Multiple Regression.

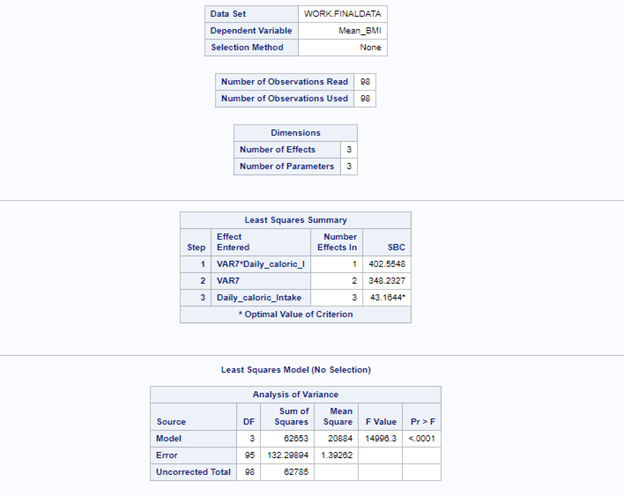
Attributes that have been considered: Obesity, Average BMI, Calories intake

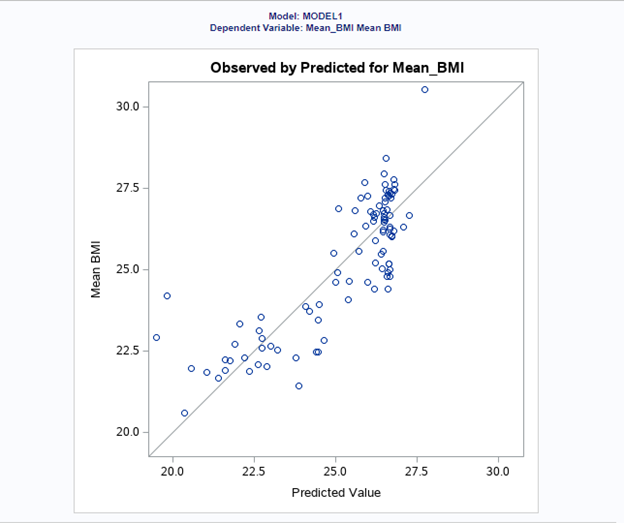
Hypothesis:

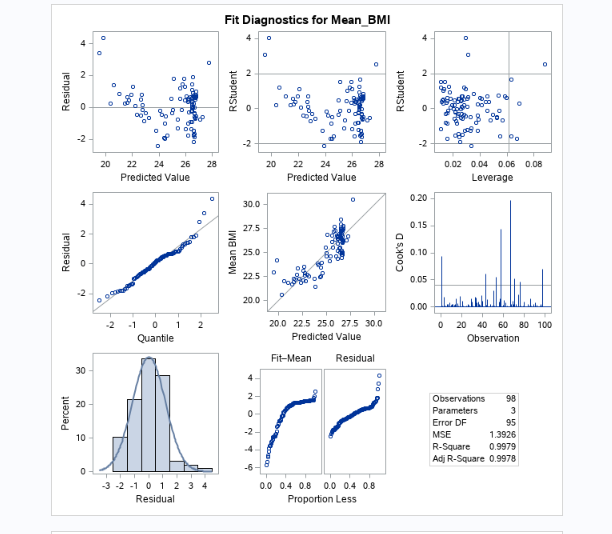
From the obtained results we get P-value<0.0001 i.e less than Alpha value. The Model and the variables considered are statistically significant.

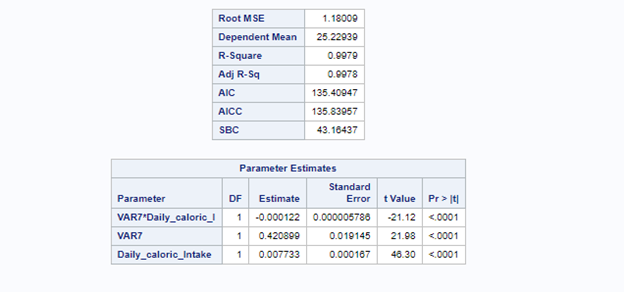
Adj R Square value 0.997 implies that 99.7% of the times, this model can tell us the moderation effect of Calories Intake and Obesity on Mean BMI.

**Inference on hypothesis:** Since P value is less than alpha value, we conclude that the model is significant. Hence the calories intake would affect the impact of obesity on the average BMI.









**CONCLUSION:**

By performing above tests we conclude that there is difference between Mean BMI urban and Mean BMI rural, also there correlation between Women’s Fertility Rate and obesity and between depression, mental disorder and daily caloric intake and there also exists a correlation between smoking, Low nutrition intake and non-communicable disease.

**REFERENCES:**

*1) Our world in data:* [*https://ourworldindata.org/search?q=*](https://ourworldindata.org/search?q=)

*2) Women Heath data:* [*https://ourworldindata.org/search?q=women+health*](https://ourworldindata.org/search?q=women+health)