Assignment 10.4

Term Project Summary

Sucharitha Puppala

Data Science, Bellevue University.

DSC530-T301 Data Exploration and Analysis.

Prof. Shankar Parajulee

August 10, 2022

In this paper I would like to give the summary of the Term Project which is about the analysis of the Lung Cancer dataset collected from the kaggle.com. for doing the exploratory data analysis.

1.Statistical/Hypothetical Question

I have selected the Lung Cancer data set from the Kaggle.com for the data analysis. Lung cancer is the second most common cancer in women and men, and it's the leading cause of cancer death, according to the American Cancer Society. Lung cancer begins in the lung and is most common in people who smoke. People who are 65 or older are at the greatest risk for lung cancer, but in rare cases it may affect people younger than 45.

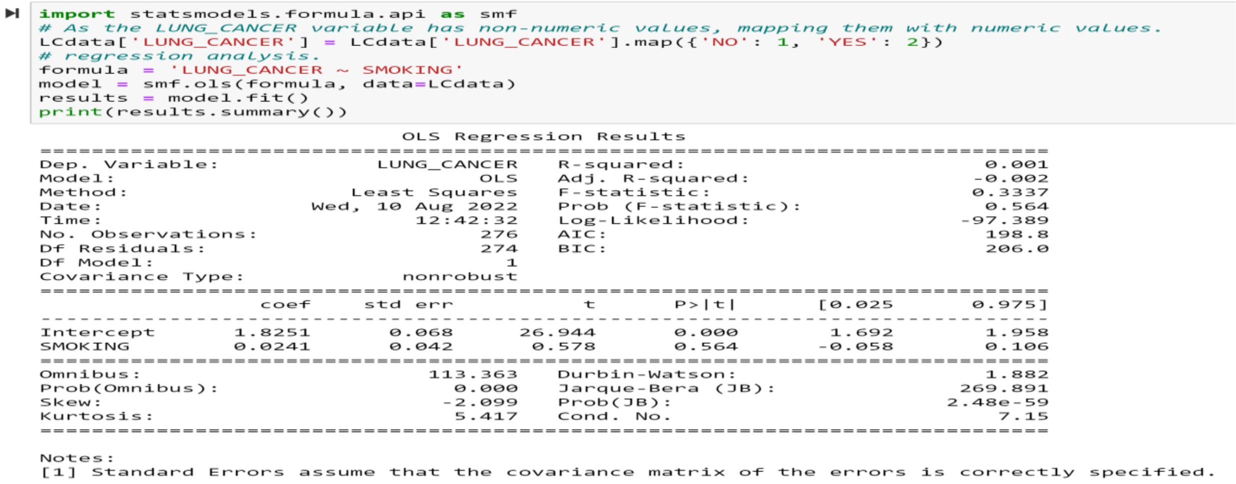
My Statistical/Hypothetical question is – Which age group of people are suffering with lung cancer and has the habit of smoking and alcohol consuming?

2. Outcome of your EDA

The following Figures are the output of the EDA of Lung cancer data set.

Figure 1 :

Regression analysis outcome on one dependent (Lung Cancer) and one explanatory (Smoking) variables.



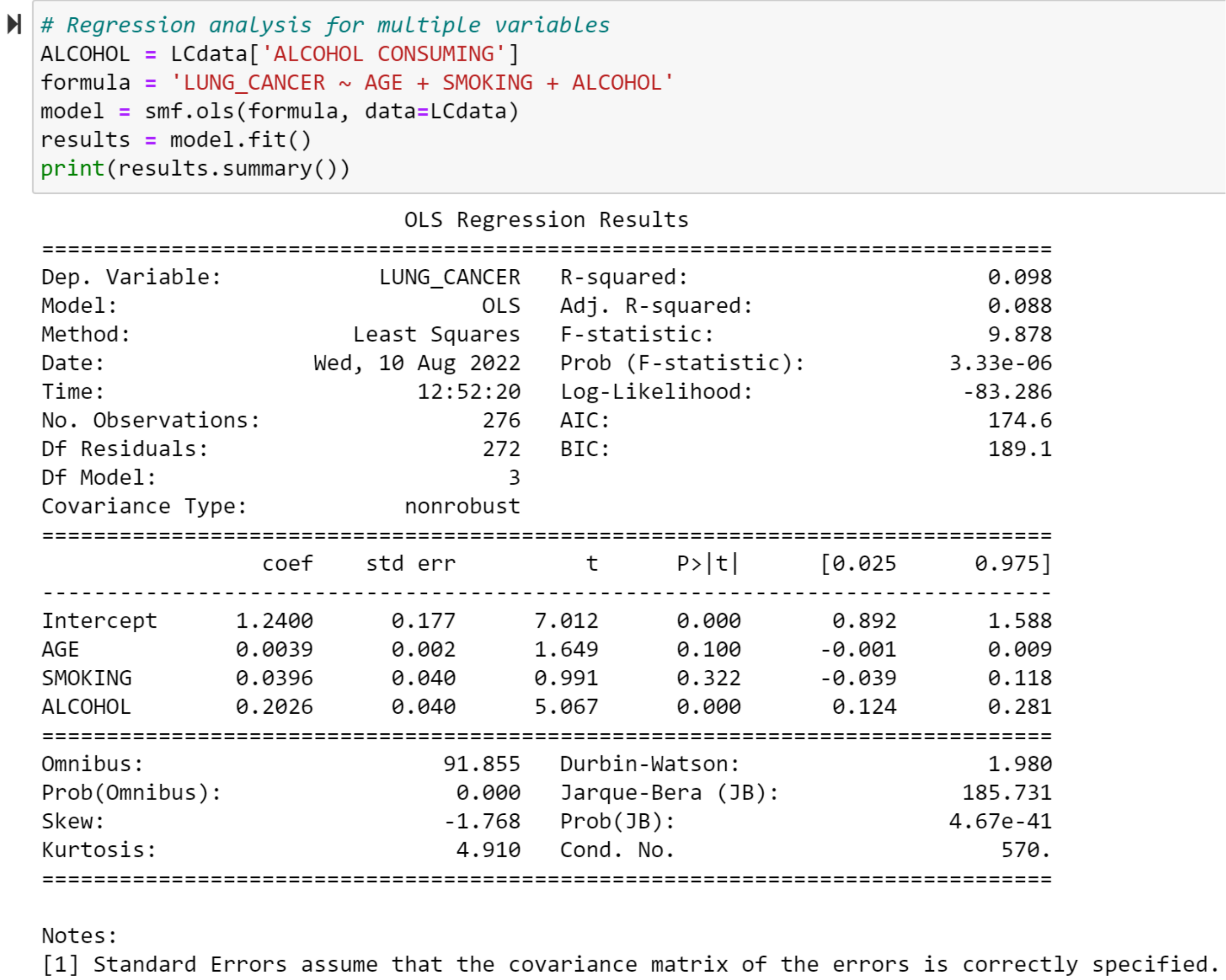
```python
import statsmodels.formula.api as smf
# As the LUNG_CANCER variable has non-numeric values, mapping them with numeric values.
LCdata['LUNG_CANCER'] = LCdata['LUNG_CANCER'].map({'NO': 1, 'YES': 2})
# regression analysis.
formula = 'LUNG_CANCER ~ SMOKING'
model = smf.ols(formula, data=LCdata)
results = model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            LUNG_CANCER   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                 -0.002
Method:                 Least Squares   F-statistic:                    0.3337
Date:                Wed, 10 Aug 2022   Prob (F-statistic):              0.564
Time:                        12:42:32   Log-Likelihood:                -97.389
No. Observations:                 276   AIC:                             198.8
Df Residuals:                     274   BIC:                             206.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.8251      0.068     26.944      0.000       1.692       1.958
SMOKING        0.0241      0.042      0.578      0.564      -0.058       0.106
==============================================================================
Omnibus:                      113.363   Durbin-Watson:                   1.882
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              269.891
Skew:                          -2.099   Prob(JB):                     2.48e-59
Kurtosis:                       5.417   Cond. No.                         7.15
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 2 :

Regression analysis outcome on one dependent (Lung Cancer) and multiple explanatory (Smoking) variables.



```
# Regression analysis for multiple variables
ALCOHOL = LCdata['ALCOHOL CONSUMING']
formula = 'LUNG_CANCER ~ AGE + SMOKING + ALCOHOL'
model = smf.ols(formula, data=LCdata)
results = model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            LUNG_CANCER   R-squared:                       0.098
Model:                            OLS   Adj. R-squared:                  0.088
Method:                 Least Squares   F-statistic:                     9.878
Date:                Wed, 10 Aug 2022   Prob (F-statistic):           3.33e-06
Time:                        12:52:20   Log-Likelihood:                -83.286
No. Observations:                 276   AIC:                             174.6
Df Residuals:                     272   BIC:                             189.1
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.2400      0.177      7.012      0.000       0.892       1.588
AGE            0.0039      0.002      1.649      0.100      -0.001       0.009
SMOKING        0.0396      0.040      0.991      0.322      -0.039       0.118
ALCOHOL        0.2026      0.040      5.067      0.000       0.124       0.281
==============================================================================
Omnibus:                       91.855   Durbin-Watson:                   1.980
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              185.731
Skew:                          -1.768   Prob(JB):                     4.67e-41
Kurtosis:                       4.910   Cond. No.                         570.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

From Figure 2, we can see that the independent variables AGE and SMOKING are not statistically significant and ALCOHOL CONSUMING variable is statistically significant.

But the main cause of LUNG CANCER is due to smoking, and the age of the patient is also to considered during analysis, we have to retain the variables in the analysis, whether they are statistically significant or not.

3. What do you feel was missed during the analysis?

After doing the analysis with the Lung cancer data set, I felt that the selection of data for analysis is very important. I have selected a data that has more Boolean values which for the analysis purpose has taken 1's and 2's instead of 0's and 1's as it is the survey of the Lung Cancer patients. It is more about whether they smoke or not, has the habit of alcohol consuming or not etc. mostly on YES or NO. I missed at giving importance to the measure of each variable, like how many days they are having coughing, how many years they are smoking etc. that might make the analysis more in depth.

4. Were there any variables you felt could have helped in the analysis?

There are few variables which when added to the 'Lung Cancer' data set can make the analysis more clear. The 'Lung Cancer' data set that I have taken has concentrated more on the symptoms and habits that a lung cancer patient is having. Some of the variables that I found would have been more helpful for the analysis are as below:

1. Type of Lung Cancer:

   There are two main types of lung cancer: Non-small cell lung cancer(NSCLC) and small cell lung (SCLC)cancer. In general NSCLC is easier to treat than SCLC and has a better prognosis.
2. Stage of the Lung Cancer:

   A cancer's stage is based on the size of the tumor and how far it has spread. Each type has its own staging system.

   a. Non-Small Cell Lung Cancer

   NSCLC is the most common type of lung cancer, and it grows more slowly than SCLC. This type of lung cancer may occur in people who have never smoked.

   Three Types of NSCLC:

i. Adenocarcinoma is the most common type of NSCLC, and it starts on the outer sections of the lungs.

ii. Squamous cell carcinoma begins in the middle of the lungs in the bronchi.

iii. Large cell carcinomas grow quickly and may begin in any part of the lungs.

And depending on the size of the tumor the stage of the cancer can be classified. There are '0 to 4' stages depending on the size of the tumor.

b. Small Cell Lung Cancer

SCLC is the most aggressive type of lung cancer and almost exclusively occurs in people who have smoked cigarettes. It has two types: small cell carcinoma and combined small cell carcinoma.

Staging for SCLC

SCLC has two stages: limited and extensive.

i. Limited stage cancer is on one side of the chest and might have spread to lymph nodes. It's rarer to find SCLC in the limited stage.

ii. Extensive stage cancer has spread widely through the lung or both lungs, through the lymph nodes on the other side of the chest and other organs or body parts. Two out of three people have extensive stage SCLC when it's first found.

3. Type of treatment:

Doctors will determine a patient's treatment plan based on the type and stage of a person's cancer. They will also take into account a person's general health and age. For example, some people with lung cancer might be too frail for surgery. In this case, the doctor will recommend radiation or chemotherapy.

4. Survival rate:

   Survival rates vary drastically depending on the type of lung cancer a person has. Each person is different and these rates don't take family history, genetics or risk factors into consideration.

5. Were there any assumptions made you felt were incorrect?

Before starting the analysis I felt that the SMOKING variable from the Lung cancer data set is the important variable that can give us a good model. But due to the data that is based on whether the patient smokes or not than the no. of cigars they smoke the analysis has changed. I followed the text book variable types and selected the dataset than looking depth into the variables, which might have given a better result of the analysis. The measurement of the variable plays an important role in the data analysis.

6. What challenges did you face, what did you not fully understand?

During the initial couple of weeks of the course, I felt Jupyter Notebook tough to work, but slowly started exploring the Jupyter Notebook with the weekly assignments and felt there is a lot more to explore in the Jupyter Notebook which helps in data analysis. I have worked on the R Markdown in the other course of this term, after which I felt Jupyter Notebook is more flexible in doing analysis and generating the reports.

I felt "Prediction" topic from Time Series analysis is more volatile and didn't fully understand the concept with the examples available in the text book. Referred many articles on Time Series analysis articles from different sources but ended up with lots of doubts. I did work on more examples to get it understand completely which took more time.

References

- *Google, drugwatch, Lung Cancer ,*

  https://www.drugwatch.com/health/cancer/lung-cancer/

- *Google, kaggle.com , Lung Cancer Dataset,*

  https://www.kaggle.com/datasets/nancyalaswad90/lung-cancer