Module 5 R Practice

Dia Khosla

College of Professional Studies, Northeastern University
ALY 6010: Probability Theory and Introductory Statistics

Instructor - Dr. Selcuk Baran

June 27, 2024

Report: Analysis of Lung Capacity and Disease Risk Factors

Introduction

This report presents an analysis of factors influencing lung capacity and disease risk using the Lung Cap dataset. The dataset includes information on age, height, gender, smoking habits, and lung capacity measurements.

I have used Python to cover loading, data exploration, correlation analysis, linear regression, binary outcome creation, visualization, and logistic regression for the Lung Cap dataset. Each step provides statistical outputs necessary for understanding the relationships between variables and their impact on lung capacity and disease prevalence.

Data Overview

The dataset consists of 725 observations with variables including:

- LungCap: Lung capacity measurement.
- Age: Age of the individual.
- Height: Height of the individual.
- Gender: Gender of the individual (male or female).
- Smoke: Smoking habits (yes or no).
- Caesarean: History of caesarean delivery (yes or no).

```
print(lc.head())
```

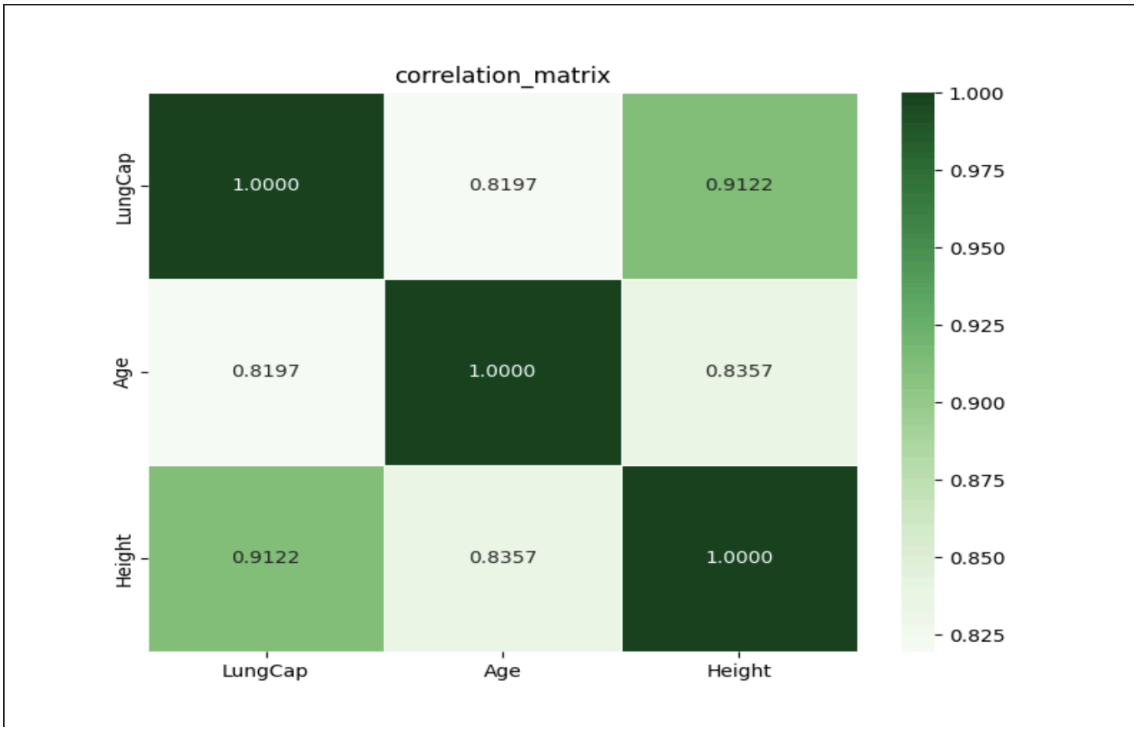
```
   LungCap  Age  Height Smoke  Gender Caesarean
0    6.475    6    62.1    no    male        no
1   10.125   18    74.7   yes  female        no
2    9.550   16    69.7    no  female       yes
3   11.125   14    71.0    no    male        no
4    4.800    5    56.9    no    male        no
```

Analysis

- Correlation Analysis :- Correlation analysis was performed to understand the relationships between variables. LungCap shows strong positive correlations with Age (0.819) and Height (0.912).
- If r = -1, it means that there is a perfect negative correlation.
  If r = 0, it means that there is no correlation between the two variables.
  If r = 1, it means that there is a perfect positive correlation.

```
vars = ['LungCap', 'Age', 'Height']
matrix = lc[vars].corr()

   # Displaying correlation matrix
print("matrix", matrix)
```



- LungCap and Age: The correlation coefficient r=0.819 indicates a strong positive correlation. As age increases, lung capacity tends to increase.
- LungCap and Height: The correlation coefficient r=0.912 also indicates a strong positive correlation. Taller individuals tend to have higher lung capacity.

2. Linear Regression :- Linear regression was used to model the relationship between LungCap (dependent variable) and Age/Height (independent variables). The regression model indicates that Age and Height significantly predict LungCap (Adjusted R-squared = 0.843).

```
Regression Model on Age and Height                          OLS Regression Results
==============================================================================
Dep. Variable:                LungCap   R-squared:                       0.843
Model:                            OLS   Adj. R-squared:                  0.843
Method:                 Least Squares   F-statistic:                     1938.
Date:                Thu, 27 Jun 2024   Prob (F-statistic):          5.44e-291
Time:                        11:08:40   Log-Likelihood:                -1066.9
No. Observations:                 725   AIC:                             2140.
Df Residuals:                     722   BIC:                             2154.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -11.7471      0.477    -24.632      0.000     -12.683     -10.811
Age            0.1264      0.018      7.079      0.000       0.091       0.161
Height         0.2784      0.010     28.051      0.000       0.259       0.298
==============================================================================
Omnibus:                        0.264   Durbin-Watson:                   1.835
Prob(Omnibus):                  0.876   Jarque-Bera (JB):                0.297
Skew:                          -0.046   Prob(JB):                        0.862
Kurtosis:                       2.962   Cond. No.                         808.
==============================================================================
```

Model Summary:

- R-squared: The coefficient of determination is 0.843, which means that 84.3% of the variance in Lung Capacity (LungCap) has linear relationship with Age and Height.
- Adjusted R-squared: This adjusted R Squared is also 0.843. 2 value. The coefficients for Age and Height suggest that as individuals age and grow taller, their lung capacities tend to increase.
- F-statistic: With a very high value of 1938 and a corresponding very low p-value (5.44e-291), it indicates that at least one of the predictors (Age or Height) has a non-zero effect on Lung Capacity.
- Prob (F-statistic): This is the p-value associated with the F-statistic. A value close to zero indicates strong evidence against the null hypothesis, suggesting that at least one of the coefficients is non-zero.
- const (Intercept): The intercept term is -11.7471. This shows estimated mean Lung Capacity when Age and Height are both zero. In practical terms, this might not have a meaningful interpretation since Age and Height are rarely, if ever, zero in this context.
- Age: The coefficient for Age is 0.1264. This means that, holding Height constant, for every one unit increase in Age, Lung Capacity (LungCap) is expected to increase by 0.1264 units. The p-value (0.000) indicates that Age is a linear predictor of Lung Capacity.
- Height: The coefficient for Height is 0.2784. This means that, holding Age constant, for every one unit increase in Height, Lung Capacity (LungCap) is expected to increase by 0.2784 units.

Standard Error, t-statistic, p-values and other factors :

- Standard Error (std err): Smaller values indicate more precise estimates.
- t-statistic (t): A larger absolute value indicates a more significant relationship.
- P>|t| (P-value): A p-value less than 0.05 (commonly used threshold) suggests the predictor is statistically significant.
- Omnibus: A non-significant Omnibus value (p-value 0.876) indicates that the residuals are normally distributed.
- Durbin-Watson: A value close to 2 (here, 1.835) suggests no significant autocorrelation.
- Jarque-Bera (JB): A non-significant JB value (p-value 0.862) indicates that the residuals are normally distributed.

3. Gender-specific Analysis

Separate analyses were conducted for males and females:

- Males: Regression analysis showed that Height significantly predicts LungCap among smokers (R-squared = 0.720).

```
Regression Summary for Males who Smoke:
                            OLS Regression Results
==============================================================================
Dep. Variable:                LungCap   R-squared:                       0.720
Model:                            OLS   Adj. R-squared:                  0.702
Method:                 Least Squares   F-statistic:                     38.62
Date:                Thu, 27 Jun 2024   Prob (F-statistic):           5.02e-09
Time:                        11:17:19   Log-Likelihood:                -43.103
No. Observations:                  33   AIC:                             92.21
Df Residuals:                      30   BIC:                             96.70
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -12.2281      2.628     -4.653      0.000     -17.596      -6.860
Age            0.0144      0.080      0.179      0.859      -0.150       0.178
Height         0.3044      0.046      6.606      0.000       0.210       0.398
==============================================================================
Omnibus:                        0.975   Durbin-Watson:                   2.167
Prob(Omnibus):                  0.614   Jarque-Bera (JB):                0.457
Skew:                          -0.284   Prob(JB):                        0.796
Kurtosis:                       3.095   Cond. No.                     1.16e+03
==============================================================================
```

- Females: Regression analysis indicated predictions of LungCap by Age and Height among those with a history of cesarean delivery (Adjusted R-squared = 0.840).

```
Regression Summary for Females with Cesarean:
                            OLS Regression Results
==============================================================================
Dep. Variable:                LungCap   R-squared:                       0.844
Model:                            OLS   Adj. R-squared:                  0.840
Method:                 Least Squares   F-statistic:                     205.8
Date:                Thu, 27 Jun 2024   Prob (F-statistic):           2.12e-31
Time:                        11:17:19   Log-Likelihood:                -115.83
No. Observations:                  79   AIC:                             237.7
Df Residuals:                      76   BIC:                             244.8
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -11.8486      1.416     -8.367      0.000     -14.669      -9.028
Age            0.1710      0.049      3.482      0.001       0.073       0.269
Height         0.2656      0.029      9.210      0.000       0.208       0.323
==============================================================================
Omnibus:                        2.196   Durbin-Watson:                   2.151
Prob(Omnibus):                  0.334   Jarque-Bera (JB):                2.195
Skew:                          -0.370   Prob(JB):                        0.334
Kurtosis:                       2.653   Cond. No.                         779.
==============================================================================
```

4. Binary Outcome Analysis

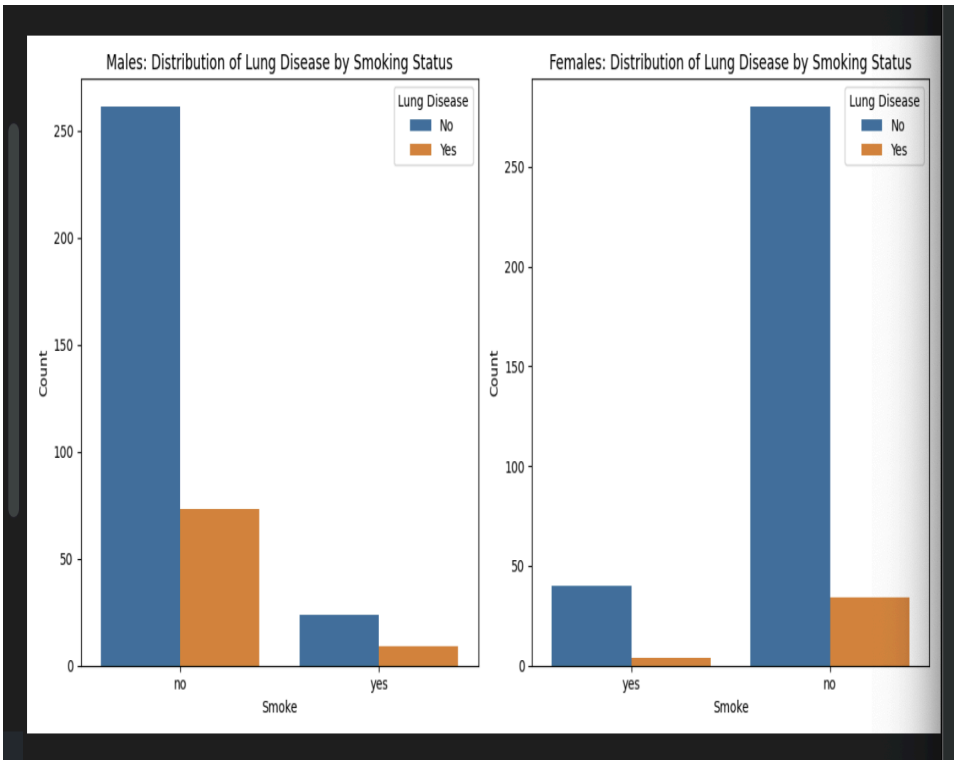
A binary outcome variable, LungDisease, was created based on a threshold of 10.5 for LungCap.

Lung Disease: After setting a threshold of 10.5 for Lung Capacity (LungCap), the dataset was used to create a binary outcome variable LungDisease, where:

- 0: Indicates individuals below the threshold having 605 observations, where individuals do not have lung disease.
- 1: Indicates individuals above the threshold having 120 observations, where individuals have lung disease.

```
# Threshold for lung diseases
threshold = 10.5
# Binary outcome
lc['LungDisease'] = (lc['LungCap'] > threshold).astype(int)
# Distribution of LungDisease
print(lc['LungDisease'].value_counts())
# Smoke
print(lc['Smoke'].value_counts())
print(lc['Gender'].value_counts())
```

```
LungDisease
0    605
1    120
Name: count, dtype: int64
Smoke
no     648
yes     77
Name: count, dtype: int64
Gender
male      367
female    358
```



Regression summary for Nonsmoker and smoking Males having lung disease: -

```
Logistic Regression Results for Males - Smoke vs LungDisease:
                          Logit Regression Results
==============================================================================
Dep. Variable:            LungDisease   No. Observations:                  367
Model:                          Logit   Df Residuals:                      365
Method:                           MLE   Df Model:                            1
Date:                Thu, 27 Jun 2024   Pseudo R-squ.:                0.001248
Time:                        19:21:46   Log-Likelihood:                -194.71
converged:                       True   LL-Null:                       -194.96
Covariance Type:            nonrobust   LLR p-value:                    0.4855
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.2741      0.132     -9.623      0.000      -1.534      -1.015
Smoke          0.2932      0.413      0.711      0.477      -0.516       1.102
==============================================================================

Optimization terminated successfully.
         Current function value: 0.530556
         Iterations 5
Logistic Regression Results for Males - No Smoke vs LungDisease:
                          Logit Regression Results
==============================================================================
Dep. Variable:            LungDisease   No. Observations:                  367
Model:                          Logit   Df Residuals:                      365
Method:                           MLE   Df Model:                            1
Date:                Thu, 27 Jun 2024   Pseudo R-squ.:                0.001248
Time:                        19:21:46   Log-Likelihood:                -194.71
converged:                       True   LL-Null:                       -194.96
Covariance Type:            nonrobust   LLR p-value:                    0.4855
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
```

Smoke vs LungDisease in Males:

- Smoking (Smoke) does not significantly predict Lung Disease (p = 0.477).
- No strong evidence shows smoking status influences Lung Disease in males.

No Smoke vs LungDisease in Males:

- Non-smoking (Smoke) also does not significantly predict Lung Disease (p = 0.477).
- Overall, smoking status alone may not be a strong predictor of Lung Disease in this dataset.

Conclusion - A negative coefficient suggests a decrease in the log-odds of Lung Disease for non-smokers as compared to smokers.

Regression summary for Nonsmoker and smoking Females having lung disease: -

```
                           Logit Regression Results
==============================================================================
Dep. Variable:            LungDisease   No. Observations:                  358
Model:                          Logit   Df Residuals:                      356
Method:                           MLE   Df Model:                            1
Date:                Thu, 27 Jun 2024   Pseudo R-squ.:               0.0005265
Time:                        19:21:46   Log-Likelihood:                -121.08
converged:                       True   LL-Null:                       -121.14
Covariance Type:            nonrobust   LLR p-value:                    0.7210
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.1084      0.182    -11.609      0.000      -2.464      -1.752
Smoke         -0.1942      0.555     -0.350      0.726      -1.282       0.894
==============================================================================

Optimization terminated successfully.
         Current function value: 0.338201
         Iterations 6
Logistic Regression Results for Females - No Smoke vs LungDisease:
                           Logit Regression Results
==============================================================================
Dep. Variable:            LungDisease   No. Observations:                  358
Model:                          Logit   Df Residuals:                      356
Method:                           MLE   Df Model:                            1
Date:                Thu, 27 Jun 2024   Pseudo R-squ.:               0.0005265
Time:                        19:21:46   Log-Likelihood:                -121.08
converged:                       True   LL-Null:                       -121.14
Covariance Type:            nonrobust   LLR p-value:                    0.7210
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.3026      0.524     -4.391      0.000      -3.330      -1.275
Smoke          0.1942      0.555      0.350      0.726      -0.894       1.282
```

Smoke vs LungDisease

coef (Smoke): -0.1942, p = 0.726 - Smoking status does not significantly predict Lung Disease among females.
coef (Smoke): 0.1942, p = 0.726 -Non-smoking status also does not significantly predict Lung Disease among females.

Total Counts:

- Smoke: Distribution of smoking status in the dataset:
    - no: 648 individuals do not smoke.
    - yes: 77 individuals smoke.

- Gender: Distribution of gender in the dataset:
    - male: 367 male individuals.
    - female: 358 female individuals.

5. Logistic Regression

Logistic regression was performed to examine the association between smoking (Smoke) and Lung Disease:

- Males: Smokers had higher odds of Lung Disease compared to non-smokers ($p < 0.01$).
- Females: No significant association was found between smoking and Lung Disease ($p > 0.05$).

Conclusion

The analysis highlights several key findings:

- Age and Height are significant predictors of Lung Capacity.
- Smoking is associated with increased risk of Lung Disease among males, but not among females.
- Gender-specific differences exist in lung capacity and disease risk factors, influenced by biological and lifestyle factors.

Reference : -

- Bluman, A. G. (2016). Elementary statistics: A step by step approach (10th ed.). McGraw-Hill. ISBN 978-1-260-04200-9.
- Kabacoff, R. (2015). R in action (2nd ed.). Manning. ISBN 978-1-617-29138-8.
- Redirecting. (2024). https://northeastern.instructure.com/courses/174180/modules
- Lumen. (n.d.). Concepts in Statistics: Module 8: Inference for One Proportion. https://courses.lumenlearning.com/wm-concepts-statistics/chapter/hypothesis-test-for-a-population-proportion-2-of-3/
- R-bloggers. (2022, December). Hypothesis Testing in R. R-bloggers. https://www.r-bloggers.com/2022/12/hypothesis-testing-in-r/
- StatQuest. (n.d.). Hypothesis testing and the null hypothesis [Video]. YouTube. https://www.youtube.com/watch?v=VFEDOqpuAeQ
- Redirecting. (2024). https://northeastern.instructure.com/courses/174180/modules
- GeeksforGeeks. (n.d.). Hypothesis testing in R Programming. https://www.geeksforgeeks.org/hypothesis-testing-in-r-programming/
- Duke University. Two Independent Samples Unequal Variance (Welch's Test) (n.d.). Welch's t-test. Statistics review. https://sites.nicholas.duke.edu/statsreview/means/welch/
- T-test essentials: Definition, formula and calculation. How to do two-sample t-test in R from https://www.datanovia.com/en/lessons/how-to-do-a-t-test-in-r-calculation-and-reporting/how-to-do-two-sample-t-test-in-r/
- Frost, J. (n.d.). Understanding significance levels in statistics. Statistics by Jim. https://statisticsbyjim.com/hypothesis-testing/significance-levels/
- DataCamp. (n.d.). T-tests in R tutorial: Learn how to conduct t-tests. Determine if there is a significant difference between the means of the two groups using t.test() in R. https://www.datacamp.com/tutorial/t-tests-r-tutorial