Statistics for Data Analytics CA1

Steps taken in SPSS to Model the Cancer Dataset using Multiple Linear Regression Analysis

Harshitha Poolakanda Somanna
Student ID: x22150366
Statistics for Data Analytics
National College of Ireland, Dublin
x22150366@student.ncirl.ie

A. Loading the Data into SPSS

1. You are required to download the "cancer.csv" file provided in Moodle by NCI to your local folder.
2. To load data into SPSS, these are the following steps
    a. Open SPSS and click on "File" in the top left corner
    b. Click “Import Data” and select “CSV Data”
    c. Browse to the location of the data file and select it.
    d. The file has variable names in the first row, make sure the box "Read variable names from the first row of data" is checked.
    e. To clean the data as a first step, check the below
        i. Remove leading spaces from string values
        ii. Remove trailing spaces from string values
3. Click "OK" to load the data into SPSS.
4. Once the data is loaded, you can view it by clicking on the "Data View" or "Variable View" tabs at the bottom of the screen. In "Data View", you can see the actual data and in "Variable View", you can see the variables and their properties.

B. Steps for Model 1: Predicting deathRate using all the Independent Variables

5. Go to the "Analyze" menu and select "Regression," then choose "Linear."
6. Select the method as “Enter”
7. In the "Linear Regression" dialog box, select the dependent variable as deathRate from the list on the left.
8. Select all the independent variables(except County) and Click the "Add" button to select one or more independent variables (predictors) from the list on the right.
9. Click on the “Statistics” dialog box to the extreme right and select the following from “Regression Coefficient” : Estimates , Model Fit , R Squared Change ,Descreptives, and Colineartity Diagnostics .From the “Residuals”, select “Durbin-Watson”
10. Click on Continue
11. Click the "Plots" button, then select "Zpred" in the X axis and "ZResid" (standardized residuals) in the Y axis.
12. Check “Histogram” and “Normal Probability Plot” from the “Standardized Residual Plots” section and click on continue.
13. Check the “Standardized" and “ Unstandardized” “Residuals” and “Predicted values” and also select “Include the Covariance Matrix” in the "Save" section.
14. Click "OK" to run the regression analysis
15. The output window will display various tables and charts, including the regression coefficients, standard errors, R-squared value, and scatter plot with standardized residuals.
16. To interpret the results, look at the regression coefficients to see the strength and direction of the relationship between each independent variable and the dependent variable. Use the scatter plot to check

for linearity, homoscedasticity, and outliers, and examine the standardized residuals to assess the model's fit and identify any influential data points.

C. Steps for Model 2: Predicting deathRate using 19 Independent Variables

17. Go to the "Analyze" menu and select "Regression," then choose "Linear."
18. Select the method as “Enter”
19. In the "Linear Regression" dialog box, select the dependent variable as deathRate from the list on the left.
20. Select the19 independent variables and Click the "Add" button to select one or more independent variables (predictors) from the list on the right.
21. Click on the “Statistics” dialog box to the extreme right and select the following from “Regression Coefficient” : Estimates , Model Fit , R Squared Change ,Descreptives, and Colineartity Diagnostics .From the “Residuals”, select “Durbin-Watson”
22. Click on Continue
23. Click the "Plots" button, then select "Zpred" in the X axis and "ZResid" (standardized residuals) in the Y axis.
24. Check “Histogram” and “Normal Probability Plot” from the “Standardized Residual Plots” section and click on continue.
25. Check the “Standardized" and “ Unstandardized” “Residuals” and “Predicted values” and also select “Include the Covariance Matrix” in the "Save" section.
26. Click "OK" to run the regression analysis
27. The output window will display various tables and charts, including the regression coefficients, standard errors, R-squared value, and scatter plot with standardized residuals.
28. To interpret the results, look at the regression coefficients to see the strength and direction of the relationship between each independent variable and the dependent variable. Use the scatter plot to check for linearity, homoscedasticity, and outliers, and examine the standardized residuals to assess the model's fit and identify any influential data points.

D. Steps for Model 3: Predicting deathRate using 15 Independent Variables

29. Go to the "Analyze" menu and select "Regression," then choose "Linear."
30. Select the method as “Enter”
31. In the "Linear Regression" dialog box, select the dependent variable as deathRate from the list on the left.
32. Select the15 independent variables and Click the "Add" button to select one or more independent variables (predictors) from the list on the right.
33. Click on the “Statistics” dialog box to the extreme right and select the following from “Regression Coefficient” : Estimates , Model Fit , R Squared Change ,Descreptives, and Colineartity Diagnostics .From the “Residuals”, select “Durbin-Watson”
34. Click on Continue
35. Click the "Plots" button, then select "Zpred" in the X axis and "ZResid" (standardized residuals) in the Y axis.
36. Check “Histogram” and “Normal Probability Plot” from the “Standardized Residual Plots” section and click on continue.
37. Check the “Standardized" and “ Unstandardized” “Residuals” and “Predicted values” and also select “Include the Covariance Matrix” in the "Save" section. From the “Distances” in the “Save” section, select “Cook’s” from the “Distance” section
38. Click "OK" to run the regression analysis
39. A new variable called “COO1” will be created with “Cook’s Distance” as the “Label” in the Dataset.
40. I have taken the threshold value as 0.00131 for “Cook’s Distance” and from the dataset, delete all the rows which has value greater than this threshold. This way, I have removed 220 rows as outliers.
41. Now follow the same steps as above to create the regression model.
42. The output window will display various tables and charts, including the regression coefficients, standard errors, R-squared value, and scatter plot with standardized residuals.
43. To interpret the results, look at the regression coefficients to see the strength and direction of the relationship between each independent variable and the dependent variable. Use the scatter plot to check

for linearity, homoscedasticity, and outliers, and examine the standardized residuals to assess the model's fit and identify any influential data points.