Steps in SPSS for Statistical Analysis of Time Series and Binary Logistic Regression

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

Part A: Time Series Analysis

Loading the data:

1. You are required to download the " nity18442004 and nitm18442004" 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.

A. Load the ‘nity18442004.csv’ to SPSS to perform Time series Analysis

Steps to perform ARIMA:

1. Open SPSS and click on "File" > "Open" to select the dataset you want to import.
2. In the "Open" dialog box, choose the file ‘nity18442004.csv’.
3. Select the file you want to import and click on "Open".
4. In the "Text Import Wizard" dialog box, select the appropriate options for delimiter and variable type, and then click on "Finish".
5. Data-> Define Date and Time-> Select years from “Cases are” and give year as 1844. Year and date columns will be created with Year and Date from 1844 to 2004
   Sequence Chart:
6. Go to "Analyse" in the top menu bar, then select "Forecasting" and then "Sequence Charts".
7. In the "Sequence Charts" dialog box, select the variable 'x' representing the time series data and click on the "OK" button to generate the sequence chart.
   Autocorrelations
8. Go to Analyze > Forecasting > Autocorrelation.

9. Select the variable ‘x’ that you want to analyze and move it to the "Variable" box.
10. Click on Autocorrelations and Partial Autocorrelations and click on the "OK" button to generate the Autocorrelation.
11. Now perform step 6 and 7 and set difference as 1
12. Now perform steps 8 and 9 and set difference as 1
13. Observe the P, D and Q value
14. Split the data into training and testing sets. The training set here is 1844 to 2003 which is used to fit the model, while the testing set 2004 is used to evaluate its performance.
15. Click on create traditional models -> dependent variable as x-> Method as Arima and select criteria
16. Fit the ARIMA model to the training data using the selected values of p, d, and q.in Non seasonal and click continue
17. Go to statistics-> select stationary r2, r2, RMSE, MSE, MAPE, MAE, normalized BIC in fit measures and select Goodness of fit and display fit measures, Ljung-Box statistic and number of outliers by model.
18. Go to plots and select, R2, ACF, PACF from plots for comparing models. Check the residuals of the fitted model to ensure that they are white noise
19. From plots for individual models, select series, observed values, forecasts, fit values, confidence intervals for forecasts, ACF and PACF
20. From options, select “First case after end of estimation period through a specific date” and under date, select as 2004 for forecast, else leave it unchecked.
21. Generate forecasts for the testing data using the fitted ARIMA model.
22. Evaluate the accuracy of the forecasts using appropriate metrics such as mean absolute error (MAE), root mean squared error (RMSE), or mean absolute percentage error (MAPE)

B. Load the ‘nitm18442004.csv’ to SPSS to perform Time series Analysis

Steps to perform SARIMA:

23. Open SPSS and click on "File" > "Open" to select the dataset you want to import.
24. In the "Open" dialog box, choose the file nitm18442004.csv’.
25. Select the file you want to import and click on "Open".
26. In the "Text Import Wizard" dialog box, select the appropriate options for delimiter and variable type, and then click on "Finish".
27. Data-> Define Date and Time-> Select “years , months” from “Cases are” and give year as 1844 and period as 12. Year ,month and date columns will be created with Year and Date from Jan 1844 to Dec 2004

    Sequence Chart:
28. Go to "Analyse" in the top menu bar, then select "Forecasting" and then "Sequence Charts".
29. In the "Sequence Charts" dialog box, select the variable 'x' representing the time series data and click on the "OK" button to generate the sequence chart.

    Autocorrelations
30. Go to Analyze > Forecasting > Autocorrelation.
31. Select the variable ‘x’ that you want to analyze and move it to the "Variable" box.
32. Click on Autocorrelations and Partial Autocorrelations and click on the "OK" button to generate the Autocorrelation.
33. Now perform step 6 and 7 and set difference as 1
34. Now perform steps 8 and 9 and set difference as 1
35. Observe the P, D and Q value

36. Split the data into training and testing sets. The training set here is 1844 Jan to 2003 dec which is used to fit the model, while the testing set 2004 jan to 2004 dec is used to evaluate its performance.
37. Click on create traditional models -> dependent variable as x-> Method as Arima and select criteria
38. Fit the ARIMA model to the training data using the selected values of p, d, and q.in Non seasonal and click continue
39. Go to statistics-> select stationary r2, r2, RMSE, MSE, MAPE, MAE, normalized BIC in fit measures and select Goodness of fit and display fit measures, Ljung-Box statistic and number of outliers by model.
40. Go to plots and select, R2, ACF, PACF from plots for comparing models. Check the residuals of the fitted model to ensure that they are white noise
41. From plots for individual models, select series, observed values, forecasts, fit values, confidence intervals for forecasts, ACF and PACF
42. From options, select “First case after end of estimation period through a specific date” and under date, select as 2004 as years and 12 as Month for forecast, else leave it unchecked.
43. Generate forecasts for the testing data using the fitted ARIMA model.
44. Evaluate the accuracy of the forecasts using appropriate metrics such as mean absolute error (MAE), root mean squared error (RMSE), or mean absolute percentage error (MAPE)

C. Steps for Exponential smoothing for monthly observation.

45. Select "Forecasting" from the top menu bar and then click on "Exponential Smoothing" in the drop-down menu.
46. In the "Exponential Smoothing" dialog box, select the variable ‘x’t to forecast from the list of available variables on the left-hand side.
47. Choose the type of exponential smoothing as Holt's Winter's Additive by clicking on the corresponding radio button.
48. Go to statistics-> select stationary r2, r2, RMSE, MSE, MAPE, MAE, normalized BIC in fit measures and select Goodness of fit and display fit measures, Ljung-Box statistic and number of outliers by model.
49. Go to plots and select, R2, ACF, PACF from plots for comparing models. Check the residuals of the fitted model to ensure that they are white noise
50. From plots for individual models, select series, observed values, forecasts, fit values, confidence intervals for forecasts, ACF and PACF
51. From options, select “First case after end of estimation period through a specific date” and under date, select as 2004 as years and 12 as Month for forecast, else leave it unchecked.
52. Specify the length of the forecast period by entering the number of periods you want to forecast in the "Forecasting period" box. Here it is 2004 and 12 to forecast, if not, leave it unchecked.
53. Click "OK" to run the analysis and produce the forecast.
54. Review the output, which will include the forecast values and any other statistics or graphs you requested. You can save the output for later use by clicking "Paste" or "Save" in the output window.

D. Steps for Exponential smoothing for yearly observation.

55. Select "Forecasting" from the top menu bar and then click on "Exponential Smoothing" in the drop-down menu.
56. In the "Exponential Smoothing" dialog box, select the variable ‘x’t to forecast from the list of available variables on the left-hand side.

57. Choose the type of exponential smoothing as Simple for 1st model and Holt's linear trend for 2nd model by clicking on the corresponding radio button.
58. Go to statistics-> select stationary r2, r2, RMSE, MSE, MAPE, MAE, normalized BIC in fit measures and select Goodness of fit and display fit measures, Ljung-Box statistic and number of outliers by model.
59. Go to plots and select, R2, ACF, PACF from plots for comparing models. Check the residuals of the fitted model to ensure that they are white noise
60. From plots for individual models, select series, observed values, forecasts, fit values, confidence intervals for forecasts, ACF and PACF
61. From options, select "First case after end of estimation period through a specific date" and under date, select as 2004 as years for forecast, else leave it unchecked.
62. Specify the length of the forecast period by entering the number of periods you want to forecast in the "Forecasting period" box. Here it is 2004 and 12 to forecast, if not, leave it unchecked.
63. Click "OK" to run the analysis and produce the forecast.
64. Review the output, which will include the forecast values and any other statistics or graphs you requested. You can save the output for later use by clicking "Paste" or "Save" in the output window.

Part B: Logistics Regression

Loading the data:

1. You are required to download the Diabetes Dataset.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.

A. Steps after loading 'Diabetes Dataset' to SPSS

1. Split the cases to N and Y for train dataset and P and Y for test data set by clicking on Data-> Split Cases
2. From the select dialog box, select, if condition is satisfied -> if-> select Class to the formula box-> and execute 'CLASS ~= 'P'' to extract N and Y class only and click continue.
3. In the output dialog box, select "copy selected cases to a new dataset" and give a name to the dataset.

4. From the select dialog box, select, if condition is satisfied -> if-> select Class to the formula box-> and execute ‘CLASS ~= 'N'’ to extract P and Y class only and click continue.
5. In the output dialog box, select “copy selected cases to a new dataset” and give a name to the dataset.

Steps for Model 1: Predicting diabetes using all the Independent Variables for train data

6. Go to the "Analyze" menu and select "Regression," then choose "Binary Logistics."
7. Select the method as “Enter”
8. In the "Logistics Regression" dialog box, select the dependent variable as Class from the list on the left.
9. Select all the independent variables(except ID and No_Pation) and Click the "Add" button to select one or more independent variables (predictors) from the list on the right.
10. Click on the “Options” dialog box to the extreme right and select the following from “Hosmer-Lemeshow goodness of fit”
11. Click on save and select “Cook’s”
12. Click on Continue
13. Once the analysis is complete, you can examine the results, including the coefficients and p-values for each independent variable, as well as the overall model fit and prediction accuracy. You can also use the results to make predictions for new cases based on the values of the independent variables.

Steps for Model 2: Predicting diabetes removing Cr and LDL Variables for train data

14. Go to the "Analyze" menu and select "Regression," then choose "Binary Logistics."
15. Select the method as “Enter”
16. In the "Logistics Regression" dialog box, select the dependent variable as Class from the list on the left.
17. Select all the independent variables(except ID, Cr ,LDL and No_Pation) and Click the "Add" button to select one or more independent variables (predictors) from the list on the right.
18. Click on the “Options” dialog box to the extreme right and select the following from “Hosmer-Lemeshow goodness of fit”
19. Click on save and select “Cook’s”
20. Click on Continue
21. Once the analysis is complete, you can examine the results, including the coefficients and p-values for each independent variable, as well as the overall model fit and prediction accuracy. You can also use the results to make predictions for new cases based on the values of the independent variables.

Steps for Model 3: Predicting diabetes removing age and Urea additionally for train data

22. Go to the "Analyze" menu and select "Regression," then choose "Binary Logistics."
23. Select the method as “Enter”

24. In the "Logistics Regression" dialog box, select the dependent variable as Class from the list on the left.
25. Select all the independent variables(except ID, Cr ,Urea, LDL and age and No_Pation) and Click the "Add" button to select one or more independent variables (predictors) from the list on the right.
26. Click on the “Options” dialog box to the extreme right and select the following from “Hosmer-Lemeshow goodness of fit”
27. Click on save and select “Cook’s”
28. Click on Continue
29. Once the analysis is complete, you can examine the results, including the coefficients and p-values for each independent variable, as well as the overall model fit and prediction accuracy. You can also use the results to make predictions for new cases based on the values of the independent variables.

Steps for Model 4: Predicting diabetes using outcome variables from model3 for test data:

30. Go to the "Analyze" menu and select "Regression," then choose "Binary Logistics."
31. Select the method as “Enter”
32. In the "Logistics Regression" dialog box, select the dependent variable as Class from the list on the left.
33. Select all the independent variables(except ID, Cr ,Urea, LDL and age and No_Pation) and Click the "Add" button to select one or more independent variables (predictors) from the list on the right.
34. Click on the “Options” dialog box to the extreme right and select the following from “Hosmer-Lemeshow goodness of fit”
35. Click on save and select “Cook’s”
36. Click on Continue
37. Once the analysis is complete, you can examine the results, including the coefficients and p-values for each independent variable, as well as the overall model fit and prediction accuracy.

Steps to perform a factor analysis with an eigenvalue of 1 in SPSS, follow these steps:

38. Open the dataset in SPSS and click on "Analyze" from the menu bar.
39. From the drop-down menu, select "Dimension Reduction" and then click on "Factor."
40. In the "Factor" dialog box, select the variables you want to include in the factor analysis from the left-hand column and move them to the right-hand column using the arrow button.
41. Click on the "Extraction" button and select "Principal Components" as the extraction method.
42. Check the "Unrotated factor solution" box and set the "Eigenvalue" to 1.
43. Click on the "Options" button and select "Varimax" as the rotation method.
44. Select scree plot.
45. Click "Continue" and then "OK" to run the factor analysis.

46. The output will show the factor loadings for each variable on each factor, as well as the eigenvalues and the percentage of variance explained by each factor.

Steps to suppress data in rotated component matrix

47. Go to "Analyze" in the top menu, then select "Dimension Reduction" and click on "Factor..."
48. In the "Factor" dialog box, select the variables you want to include in the analysis and move them to the "Variables" box using the arrow button.
49. In the "Extraction" tab, select the method you want to use to extract the factors. You can choose between Principal Components, Principal Axis Factoring, and Maximum Likelihood.
50. In the "Rotation" tab, select the rotation method you want to use. The most commonly used rotation methods are Varimax and Promax.
51. In the "Options" tab, check the box that says "Suppress small coefficients" and enter a value for the suppression threshold. This will suppress any factor loadings that are smaller than the specified threshold.
52. Click "OK" to run the analysis and view the rotated component matrix with suppressed data.

Steps for factor analysis set as regression

53. Open your dataset in SPSS and go to Analyze > Dimension Reduction > Factor.
54. In the Factor Analysis dialog box, select the variables you want to include in the analysis and move them to the Variables list on the right.
55. Under the Extraction Method tab, select "Principal Components" as the method and check the box for "Scores."
56. Under the Scores tab, select "Regression" as the method and choose the variables for which you want to calculate factor scores.
57. Under the Rotation tab, select the rotation method you want to use (e.g., Varimax).
58. In the Factor Analysis dialog box, click the "Options" button.
59. Under the "Output" tab, select "Factor Scores" and "Rotated Factor Matrix."
60. Under the "Extraction" tab, set the eigenvalue to 1.
61. Click "OK" to run the analysis.

Steps for new: Predicting diabetes using 3 factors after the component transformation matrix

62. Go to the "Analyze" menu and select "Regression," then choose "Binary Logistics."
63. Select the method as “Enter”
64. In the "Logistics Regression" dialog box, select the dependent variable as Class from the list on the left.
65. Select the independent variables as FAC1_1,FAC2_1 and FAC3_1 and Click the "Add" button to select one or more independent variables (predictors) from the list on the right.
66. Click on the “Options” dialog box to the extreme right and select the following from “Hosmer-Lemeshow goodness of fit”

67. Click on save and select “Cook’s”
68. Click on Continue
69. Once the analysis is complete, you can examine the results, including the coefficients and p-values for each independent variable, as well as the overall model fit and prediction accuracy.