BUSN:5760 Applied Business Statistics

Assignment 2: Regression Analysis

Pick any cross-sectional data that you’re interested in from Kaggle and download it as a CSV. It must include at least 4 independent variables and one that you’ll use as your dependent variable. Pick a dataset with an actual logical / causal relationship for your analysis. If your model is nonsensical, I won’t accept it.

Create a Jupyter Notebook and perform the following tasks in separate cells. After each cell, the results of that cell’s operation need to be displayed. If you send me a file with lots of errors or that lacks something resembling the requested output, I won’t accept it.

1) Import Pandas, Matplotlib, Numpy, Statsmodels, Seaborn, and Scipy. Do not use any other libraries. No output is necessary here.

2) Import your CSV file into Python as a Pandas dataframe. Print the top five rows and data types for each variable.

3) Use Pandas to calculate summary statistics.

4) Use Seaborn’s pairplot to create the big graphic containing all the different variables plotted against each other.

5) Create a list of your independent variables and use a loop to create a plot of each one against your dependent variable.

6) Create three different linear regression models using different subsets of independent variables. Print the regression results for each.

7) Compare your models using the four different selection statistics we discussed. (R^2, AIC…) and select one as your best model.

8) Calculate the variance inflation factor for your independent variables to check for multicollinearity. Print the results and tell me if you think multicollinearity was detected. If it was, make the appropriate modifications to your model before proceeding.

9) Using your best model, create a variable containing the predicted values of Y, called y-hat, and a variable containing the residuals called e. Make a scatterplot with the residuals on the y-axis and the predicted values on the x-axis.

10) Plot a histogram of your residuals to check that they’re normally distributed. Tell me if you think they are normally distributed or not.

11) Use a statistical test to test for heteroskedasticity (BP test). Tell me if you think heteroskedasticity is present or not.

12) Create an influence plot to check for outliers. Display the plot and tell me if you think any of your data points are problematic.

13) Create a scatterplot of Y vs Y-Hat. Include a trendline.

14) Create a loop that loops through your x-variables and plots them against your residuals. Display the plots. Tell me if you see any problematic patterns.

15) Create a new dataframe with five hypothetical observations’ worth of x-values and use the predict function to calculate the estimated y values for each set of x values.

Make this look professional and correct. Don’t send me empty charts, tons of error codes, etc.

If the work you send me appears to have come from AI or is significantly beyond your actual knowledge of the subject, you will get an F on this assignment. If you send me work that appears to be someone else’s, you and the other person will get an F on the assignment.

I teach multiple courses at multiple universities on data analysis and data science. I know lots of tricks that you likely don’t know and have spent a good amount of time coming up with some pretty crafty ways to catch cheaters. Don’t cheat. Just do the work. Use the notebooks and lecture slides from class, as well as the Python and specific library documentation. Don’t use anything else. It is fine to work together, but it is not fine one bit to just send me someone else’s work.

Library Documentation:

Numpy: <https://numpy.org/>

Pandas: <https://pandas.pydata.org/>

Statsmodels: <https://www.statsmodels.org/stable/index.html>

Scipy: <https://scipy.org/>

Matplotlib: <https://matplotlib.org/>

Seaborn: <https://seaborn.pydata.org/>

Jupyter Notebooks:

Pandas Notebook: <https://github.com/rjwrobel86/Python4Statistics/blob/main/Notebooks/Pandas101.ipynb>

Intro to Regression: <https://github.com/rjwrobel86/Python4Statistics/blob/main/Notebooks/Regression.ipynb>

Housing Prices Regression: <https://github.com/rjwrobel86/Python4Statistics/blob/main/Notebooks/HousingRegression.ipynb>

Regression - Dummy and Interaction Variables <https://github.com/rjwrobel86/Python4Statistics/blob/main/Notebooks/DummyAndInteraction.ipynb>