DSL201: Statistical Programming

Dr. Anil Kumar Sao

Assignment 5

Deadline: 17/04/2024 11:59pm

Instructions for Submission: You can submit your solution as a Jupyter Notebook/ Matlab file with comments and discussions on the results obtained in each step

1. Follow Standard Report Format: Include sections like Introduction, Data, Methodology, Results, Discussion, and Conclusion.
2. **File Naming Convention:** Adhere to the specified naming convention for each file you submit (e.g. RollNumber_FirstName_Asg2).
3. Refrain from using zip files. If necessary, submit multiple files.
4. Include comments in the code explaining the logic and any assumptions made.
5. Include References: Cite any external sources or references used in your assignment.
6. Code Quality: Ensure your code follows best practices and is well-organized and **avoid plagiarism** as a plagiarism check will be conducted.
7. Be aware that late submissions are not permitted; ensure timely submission.

1. The management of a biotechnology company specializing in enzyme production is concerned about the impact of fluctuating humidity levels on the moisture content of the substrates used in their fermentation process. Over the course of 15 days, they collected data on the relative humidity and moisture content of the substrates stored in their fermentation facility.

   Relative Humidity (%):

   `46, 53, 29, 61, 36, 39, 47, 49, 52, 38, 55, 32, 57, 54, 44`

   Moisture Content (%):

   `12, 15, 7, 17, 10, 11, 11, 12, 14, 9, 16, 8, 18, 14, 12`

   1. Create a Python function named `calculate_least_squares_estimator` that takes two arrays `x` and `y` (representing relative humidity and moisture content, respectively) as input and returns the least squares estimator (slope and intercept) for the given dataset.
   2. Define a function called `interpret_regression_line` that takes the slope and intercept of the regression line as input and provides an interpretation of their significance in the context of the problem.

   3. Write a function named `predict_moisture_content` that accepts the relative humidity, slope, and intercept of the regression line as arguments and returns the predicted moisture content for the given relative humidity.

   4. Implement a function named `plot_regression_line` that creates a scatter plot of the data points along with the estimated regression line. This function should take relative humidity, moisture content, slope, and intercept as inputs, and ensure that the plot is properly labeled and titled.

   5. Develop a function called `assess_goodness_of_fit` that evaluates the goodness of fit of the regression model. This function should take the actual moisture content and the predicted moisture content as inputs and return a metric (e.g., mean squared error) indicating the quality of the fit.

   6. The name of the function should be `compute_confidence_interval` to generate a confidence interval for moisture content prediction based on a certain relative humidity value. The uncertainty in the forecast must be taken into account and a range within which it is expected that the moisture levels will be with a given level of assurance for example 95%.

2. An individual claims that the amount of time spent studying does not affect exam scores. To test this hypothesis, the study time (in hours) and corresponding exam scores (out of 100) of several students were recorded. The following data was collected:

Study Time (hours)	Exam Score
2	70
4	75
6	80
8	85
10	90
12	92
14	94
16	96
18	98
20	100
22	101
24	102
26	103
28	104
30	105

   1. Develop a Python function named `calculate_slope_intercept` to compute the least squares estimator (slope and intercept) for the provided dataset of study time and exam scores. Ensure that your function takes the study time and exam score data as inputs and returns the slope and intercept of the regression line.

   2. Write a Python program to visualize the relationship between study time and exam scores using a scatter plot. Ensure appropriate labeling of axes and a title for the plot.

   3. Implement a Python function called `predict_exam_score` to predict the exam score for a given amount of study time using the estimated regression line. Ensure that your function takes the study time, slope, and intercept as inputs and returns the predicted exam score.

   4. Calculate the residual error for each data point and plot the residuals against the study time. Discuss any patterns observed in the residual plot and their implications for the regression

analysis.

5. Evaluate the significance of the regression model. Write code to perform hypothesis testing for the slope parameter. Discuss the results of the hypothesis test and their implications for the relationship between study time and exam scores.

6. Discuss whether the data support or refute the claim that the amount of time spent studying does not affect exam scores based on the results of your analysis in Python.