DS203: Exercise - 2

Submission Deadline

Feb 6, 2025, 11:55 PM.

Penalties

(-1) for late submissions, (-10) for copied/fraudulent work.

Instructions

1 Question 1:

(a) Load the dataset (provided as `E2.csv`) into a spreadsheet and visualize it using a scatter plot and histograms. Describe your observations and conclusions.

(b) On the basis of this initial analysis, discuss the correctness and possibility of fitting an SLR model to this data set.

(c) Calculate the descriptive statistics for x and y, including central tendency, spread, and shape (mean, median, mode, variance, standard deviation, range, skewness, and kurtosis). Analyze these statistics and state your observations and conclusions. Can you comment on the expected quality of the SLR model based on these statistics?

2 Question 2:

(a) Calculate the Pearson Correlation Coefficient (`r`) between `x` and `y`. What does this value convey about the relationship between `x` and `y`?

(b) Data Transformation and Correlation Analysis:

- Transform the variables `x` and `y` using the following transformations:

$$z = \log(y), \quad w = \sqrt{x}$$

- Calculate the Pearson Correlation Coefficient (`r`) between the transformed variables `z` and `w`.
- Compare the correlation between `z` and `w` with the original correlation between `x` and `y`.
- Comment / explain:
  - Whether these transformations have strengthened or weakened the relationship.
  - Why the correlation might change due to the transformations.
  - Can you think of situation(s) where applying the 'log' transformation to 'y' can strengthen the linear relationship between 'x' and 'y'?

3 Question 3:

(a) Fit a linear regression model ($y = a + bx$) using a spreadsheet to create the regression coefficients, R^2, `p-values`, F-statistic, and residual standard error in the output.

(b) Comment on the goodness-of-fit of the linear model using R^2 and residual diagnostics.

(c) Predict the average `y` value for the entire dataset and compare it with the actual average of `y`. State your observations and conclusions, with reasons,

(d) Plot the residuals against the fitted values and analyze it. What does this tell you about the model?

4 Question 4:

(a) Modify the dataset to create 5 variants of `'y'`, as explained below:

- Two datasets where the standard deviation of `y` is reduced to 5 and 10, respectively.
- The original dataset
- Two datasets where the standard deviation of `y` is increased to 20 and 25, respectively.

Use the following formulas to modify `y`:

- To reduce the spread:

$$y_{\text{new}} = y \times \left(\frac{\text{desired std}}{\text{current std}} \right)$$

- To increase the spread:

$$y_{\text{new}} = y + \text{noise}, \quad \text{where noise} \sim N(0, \sigma_{\text{noise}})$$

$$\sigma_{\text{noise}} = \sqrt{\text{desired std}^2 - \text{current std}^2}$$

For each variant, fit a linear regression model ($y = a + bx$) and report the following in a Table:

- R^2, `p-values`, F-statistic, and Root Mean Squared Error (RMSE).

(b) How does the standard deviation affect R^2, F-statistic, and RMSE. Analyze and explain using appropriate plots.

(c) In each noisy variant, explore the implications of the standard deviation on the confidence intervals associated with the regression coefficients. Specifically, interpret the width of the confidence intervals and what it suggests about parameter uncertainty.

Submission Guidelines

- Submit a concise, well-structured report with answers to all questions, including tables, plots, and analyses.
- Attach the spreadsheet or Python notebook used for calculations. - File naming: `E2-YourRollNo`.