Statistical Learning and Data Analysis - Exercise 1 (Basic Database Analysis)

In this exercise, we will conduct basic qualitative checks on data and perform regression analysis. The dataset used in this exercise (prostate.csv) comes from a study by Stamey et al. (1989), which examined the relationship between prostate-specific antigen (PSA) levels and various clinical measurements in 97 men scheduled for radical prostatectomy.

PSA is a protein produced by the prostate gland, and elevated PSA levels are often associated with a higher likelihood of prostate cancer. This dataset provides valuable insights into how PSA levels correlate with other clinical factors in the context of prostate cancer.

Description of Variables

- **lcavol**: Log of cancer volume
- **lweight**: Log of prostate weight.
- **age**: Age of the patient.
- **lbph**: Log of the amount of benign prostatic hyperplasia (BPH).
- **svi**: Indicates the presence of seminal vesicle invasion.
- **lcp**: Log of capsular penetration, indicating cancer spread outside the prostate. Very small values of capsular penetration cannot be reliably measured. In such cases, the measurement is arbitrarily set to 0.25.
- **gleason**: Gleason score, a grading system for prostate cancer aggressiveness.
- **pgg45**: Measures the percentage of Gleason scores of 4 or 5 recorded in the patient's visit history before their final Gleason score. Since a higher Gleason score indicates more severe cancer, **pgg45** provides insights into the patient's past cancer severity.
- **lpsa**: The logarithm of the PSA score.

This dataset allows us to explore the relationships between these clinical variables and PSA levels, aiding in a better understanding of prostate cancer progression.

1. **Describe the Dataset:**

- What are the explanatory variables in the problem? What are the dependent (explained) variables?
- What are the types of variables (e.g., categorical, continuous, etc.)? How do they behave?

2. **Dependency Between Explanatory Variables:**

   - Do you expect dependence among your explanatory variables?
   - Use pairplot to visualize relationships:

     ```
     import seaborn as sns
     sns.pairplot(dataset)
     ```

     where `dataset` is the matrix of explanatory variables. Consider the variables you want to display and how (explore the parameters' hue' and 'kind' of the pairplot function).
   - Explain what the two-dimensional and one-dimensional panels describe. Discuss whether they help understand variable dependencies.
   - Describe dependencies between ordinal/categorical variables and the dependent variable.

3. **Linear Regression Using Statsmodels:**

   - Use Python's `statsmodels` library to perform linear regression. The 'lpsa' column is the dependent variable, and the remaining are explanatory variables.
   - Extract the p-values for all slope coefficients (β values).
   - Explain the significance of these p-values.

4. **Backward Elimination Using Adjusted R^2:**

   - Apply the Backward Elimination method based on the adjusted R^2 metric.
   - Identify which columns are retained and explain why.
   - Generate a **single plot** displaying all eliminated variables against the dependent variable and interpret it.

   Reminder: In this method:

   (a) Compute adjusted R^2 for all X variables.

   (b) Remove one column at a time and fit a new linear model.

   (c) Recalculate adjusted R^2 and check whether it improves.

   (d) If it improves, the removed column is unnecessary for the model.

Submission Guidelines

- Work may be submitted in pairs (only one submission per pair, including both student IDs and names).
- Submissions can be in a **compiled** Jupyther notebook or PDF (if PDF, include source code file).
- The Aesthetic quality of the report matters and will influence grading. Ensure:
  - Clear and detailed explanations.
  - Well-labeled graphs with appropriate font sizes.
  - Logical grouping of related graphs into single figures where possible.

Good luck!