Homework 1:
Part I: Regression Modeling with the Prostate Cancer Data

(Modeling, Diagnostics, Collinearity)

DATA 210P

Overview

In this assignment you will analyze a classic dataset used in *The Elements of Statistical Learning* (Hastie, Tibshirani, Friedman), commonly referred to as `prostate.data`. The goal is to build, diagnose, and compare regression models with **log PSA** as the outcome.

Data access and citation. The canonical file is hosted by the authors of ESL at `https://hastie.su.domains/ElemStatLearn/datasets/prostate.data`. Some course repositories and Kaggle users mirror this dataset for convenience; if you obtain it from Kaggle, cite the Kaggle dataset page *and* the ESL source as the original reference.

Outcome.

$$\texttt{lpsa} = \log(\text{PSA})$$

Your primary modeling goal is to explain/predict `lpsa` using the clinical predictors.

Variables (feature dictionary)

The dataset contains $n = 97$ patients and the following columns:

- `lcavol`: log(cancer volume)
- `lweight`: log(prostate weight)
- `age`: age (years)
- `lbph`: log(benign prostatic hyperplasia amount)
- `svi`: seminal vesicle invasion indicator (0=no, 1=yes)
- `lcp`: log(capsular penetration)
- `gleason`: Gleason score
- `pgg45`: percent Gleason pattern 4 or 5
- `lpsa`: log(PSA) (response)
- `train`: indicator of training/test split used in ESL examples (TRUE/FALSE)

Software, reproducibility, and deliverables

- You may use **R** or **Python**, but your submission must include:
  - a clean script or notebook (`.R`/`.py`/`.ipynb`) that runs end-to-end,
  - a write-up (PDF) with explaining your work and interpreting results.
- Every figure must include axis labels, units (when relevant), and a caption.
- Report uncertainty where appropriate (standard errors, confidence intervals, cross-validated error bars, etc.).
- Submit your PDF write up and your R code.

Questions (answer all; show code + interpretation)

1. **Data ingestion and basic checks.**

   (a) Load the dataset and verify dimensions, column names, missing values, and data types.

   (b) Run a full exploratory data analysis. Visualize the data and the relationship between the variables.

   (c) Explain why many variables appear log-transformed and what that implies for interpretation.

   (d) Use the `train` indicator to create a training set and a test set. Report n_{train} and n_{test}.

2. **Exploratory data analysis (EDA).**

   (a) Create histograms/density plots for `lpsa` and at least four predictors.

   (b) Create a scatterplot matrix (or a set of faceted scatterplots) showing `lpsa` versus each predictor.

   (c) Identify at least two predictors that appear strongly associated with `lpsa` and justify visually.

3. **Simple linear regression (SLR) with a fitted regression line.**

   (a) Fit the model on the training set:

   $$\texttt{lpsa} = \beta_0 + \beta_1 \, \texttt{lcavol} + \varepsilon.$$

   (b) Plot the training data (`lpsa` vs `lcavol`) with the fitted regression line and a 95% confidence band for the mean response.

   (c) Interpret $\widehat{\beta}_1$ in words (in the *log–log* scale). What does a one-unit increase in `lcavol` represent?

   (d) Evaluate performance on the test set (RMSE and R^2). Comment on generalization.

4. **Multiple linear regression (MLR) and interpretation.**

   (a) Fit a full model on the training set using all predictors except `lpsa` and `train`.

   (b) Report estimated coefficients, standard errors, and 95% confidence intervals for all predictors.

   (c) Provide a careful interpretation of at least two coefficients, including `svi` (binary) and one continuous predictor.

   (d) Compare training vs test performance (RMSE and R^2). Does adding predictors improve test performance?

5. **Assumption checking and diagnostics (must include visuals).** For your chosen MLR model (full model or a selected model), assess the linear regression assumptions:

   (a) **Linearity:** residuals vs fitted plot; residuals vs each key predictor. What patterns would worry you?

   (b) **Homoscedasticity:** comment on whether residual spread is constant; optionally perform a Breusch–Pagan test.

   (c) **Normality of errors:** Q–Q plot of residuals; comment on tail behavior.

   (d) **Outliers and influence:** identify influential points using leverage and Cook's distance; report the top 3 influential observations and discuss whether they are clinically plausible.

Homework 1: Part II
Kaggle's Global AI and Data Science Job Market

Questions (answer all; show code + interpretation)

1. **Data ingestion and basic checks.**

   (a) Load the dataset from
   https://www.kaggle.com/datasets/mann14/global-ai-and-data-science-job-market-20202026
   and verify dimensions, column names, missing values, and data types.

   (b) Run a full exploratory data analysis. Visualize the data and the relationship between the variables.

   (c) Explain why many variables appear log-transformed and what that implies for interpretation.

   (d) Use the `train` indicator to create a training set and a test set. Report n_{train} and n_{test}.

2. **Exploratory data analysis (EDA).**

   (a) Create a response or outcome variable based on minimum and maximum salary in dollars.

   (b) Visualize data features, and use data summaries, to demonstrate their internal relationships and their associations with the outcome variable you created in the previous part.

   (c) Comment on missing values, outliers, anomalies in data, and also patterns of data across important feature categories.

   (d) provide a summary of your EDA.