DATA 200BP: Intermediate Probability & Statistics II

Homework #2: Linear Modeling, Selection, Validation, Ridge & Lasso

Assigned Dataset

For this homework, you will work with a real-world dataset from the **UCI Machine Learning Repository**:

Bike Sharing Dataset
`https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset`

The dataset contains observations from the Washington, DC Capital Bikeshare system during 2011–2012 and includes information on bike rental usage, weather conditions, seasonal patterns, and calendar effects. Two versions of the data are provided:

- `day.csv`: daily aggregated bike rental counts (recommended for this assignment),
- `hour.csv`: hourly bike rental counts (larger sample size with stronger time-of-day effects).

You may use **either** dataset, but you must clearly state which one you choose and briefly justify your decision.

Outcome (Response) Variable

The primary response variable is the total number of bike rentals, denoted by `cnt`.

- In `day.csv`, `cnt` represents the total number of bike rentals on a given day.
- In `hour.csv`, `cnt` represents the total number of bike rentals during a specific hour.

Important note before you start modeling data!. Although `cnt` is a count variable, this assignment intentionally focuses on **linear regression methods** rather than Poisson or negative binomial models. In this dataset, rental counts are typically large and exhibit substantial overdispersion due to strong seasonal and weather-related effects, violating key Poisson assumptions. You are expected to evaluate whether a transformation of the response variable (e.g., $\log(\texttt{cnt} + 1)$) is appropriate to improve linear model assumptions.

Part I. Baseline Linear Model and Interpretation

1. Fit an initial multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon,$$

   using a sensible subset of predictors. Do not include all available predictors indiscriminately; justify your choices based on subject-matter considerations.

2. Report:

   - a table of coefficient estimates, standard errors, and p-values,
   - an interpretation of at least five coefficients (including at least one categorical predictor),
   - a short discussion distinguishing statistical significance from practical relevance.

Part II. Transformations and Diagnostics

3. Assess linear model assumptions using appropriate diagnostics:

   - residuals vs fitted values,
   - residuals vs key predictors,
   - normal Q–Q plot of residuals,
   - assessment of constant variance,
   - leverage and influence diagnostics (e.g., Cook's distance).

4. If assumptions are violated, propose and implement at least **one** transformation strategy, such as:

   - transforming the response (e.g., $\log(\texttt{cnt} + 1)$),
   - transforming a skewed predictor,
   - adding polynomial terms,
   - including scientifically justified interaction terms.

5. Compare the baseline and transformed models using:

   - an in-sample measure (e.g., R^2 or adjusted R^2),
   - a cross-validated predictive metric (see Part IV),
   - a clear discussion of how the transformation affects interpretation.

Part III. Collinearity

6. Evaluate collinearity among predictors by:

   - computing and displaying a correlation matrix for numeric predictors,
   - computing variance inflation factors (VIFs).

7. Identify at least two predictors that are strongly collinear (or justify why none are). Discuss:

- how collinearity affects standard errors and interpretation,
- whether it harms predictive performance,
- strategies to address collinearity.

Part IV. Model Selection and Validation

8. Use K-fold cross-validation ($K = 5$ or 10) to estimate out-of-sample predictive performance. Report:
   - the chosen performance metric (e.g., RMSE or MAE),
   - mean performance and variability across folds.
9. Perform model selection using **one** of the following:
   - forward selection,
   - backward elimination,
   - stepwise selection.

   Clearly state the selection criterion (AIC, BIC, or CV-based).
10. Compare the selected model to the baseline model using both in-sample summaries and cross-validated performance.

Part V. Ridge and Lasso Regression

11. Fit ridge regression and lasso regression using `glmnet`. Use cross-validation to select the tuning parameter λ.
12. Report:
    - the selected λ value(s),
    - cross-validated predictive performance.
13. Compare ridge, lasso, and your selected OLS model. Discuss:
    - predictive accuracy,
    - interpretability,
    - the effect of regularization on collinearity.

Deliverables

- A concise written report (PDF) addressing all questions.
- Python or R code.