Multiple Linear Regression Project

Objective

In this assignment, you will build a multivariate Ordinary Least Squares (OLS) regression model to predict a target variable using various predictor variables. You will apply your knowledge of multiple linear regression, model diagnostics, and interpretation to develop a robust and reliable model.

Dataset

You will choose your own dataset for this assignment. The dataset must meet the following criteria:

- It should contain at least 9 variables (1 target variable and 8 predictor variables).
- It should have a minimum of 600 observations. Prefer datasets with more than 1000 observations.
- The dataset can include numerical, categorical, or a combination of both types of variables.

Tasks

1. Data Exploration and Preprocessing
   - Explore the dataset, understand the variables, and identify any missing values or outliers.
   - Analyze the distribution of the data, including numerical summaries and visualizations.
   - Identify and handle outliers using appropriate techniques (e.g., trimming, winsorization, or transformation).
   - Handle missing values using appropriate techniques (e.g., imputation, deletion).
   - Perform any necessary data transformations or feature engineering.
   - Encode categorical variables using suitable methods (e.g., one-hot encoding, dummy coding).
   - Include at least 2 visualizations (e.g., histograms, scatter plots, box plots) to explore the data.
2. Model Building
   - Split the dataset into training and testing sets (e.g., 80% training, 20% testing).
   - Fit a multivariate OLS regression model using the training set and the selected predictor variables.
   - Use another algorithm to predict the outcome and compare results.

3. Model Evaluation
   - Compute the adjusted R-squared and Root Mean Squared Error (RMSE) for the model on the testing set.
   - Interpret the adjusted R-squared and RMSE values and assess the model's goodness of fit.

4. Model Diagnostics

- Test 3 out of the following 5 assumptions for your regression model:
  a) Linearity of the model parameters
  b) Serial independence of errors
  c) Homoscedasticity (constant variance of errors)
  d) Normality of the residual distribution
  e) Absence of multicollinearity
- Provide appropriate statistics and visualizations for the 3 assumptions you choose to test.

5. Model Interpretation
   - Interpret the model coefficients and their statistical significance.
   - Discuss the practical implications of the model and its potential applications.
   - Identify any limitations or assumptions of the model.

6. Additional Considerations
- Identify and discuss any potential outliers in the dataset and their impact on the model.
- Discuss strategies for handling missing values, if applicable.
- Explain your approach to handling categorical variables and the rationale behind your chosen method.

Deliverables

a. Your final model equation.
b. At least 2 visualizations (e.g., histograms, scatter plots, box plots) to explore the data.
c. The statistical software output, including (adjusted) R-squared and Root Mean Squared Error (RMSE).
d. Your code file with annotations.
e. Model diagnostics, including statistics and visualizations for the 3 assumptions you tested.
f. Your interpretation of the model, discussing the coefficients, significance, practical implications, and limitations.
g. Project Final Report

Submission

1. Create a new repository on GitHub and upload your project files.
2. Submit the link to your GitHub repository containing all the required deliverables.

Evaluation

Your assignment will be evaluated based on the following criteria:

- Correctness of the model building and evaluation process
- Thoroughness of the model diagnostics for the 3 chosen assumptions
- Clarity and insight in the model interpretation

- Proficiency in handling data preprocessing tasks, including outlier detection, missing value treatment, and variable transformations.
- Appropriate use of visualizations for data exploration and assumption testing
- Overall organization and presentation of your work
- Successful submission of the project to GitHub

Classification Model Project

Objective

In this project, you will build a classification model to predict a categorical target variable using various predictor variables. You will apply your knowledge of classification algorithms, model evaluation, and interpretation to develop an accurate and reliable model.

Dataset

- You will choose your own dataset for this assignment. The dataset must meet the following criteria:
- It should contain at least 9 variables (1 categorical target variable and at least 8 predictor variables).
- It should have a minimum of 600 observations. Prefer datasets with more than 1000 observations.

The dataset can include numerical, categorical, or a combination of both types of variables.

Tasks

1. Data Exploration and Preprocessing
   - Explore the dataset, understand the variables, and identify any missing values or outliers.
   - Analyze the distribution of the data, including numerical summaries and visualizations.
   - Identify and handle outliers using appropriate techniques (e.g., trimming, winsorization, or transformation).
   - Handle missing values using appropriate techniques (e.g., imputation, deletion).
   - Perform any necessary data transformations or feature engineering.
   - Encode categorical variables using suitable methods (e.g., one-hot encoding, dummy coding).
   - Include at least 2 visualizations (e.g., histograms, scatter plots, box plots) to explore the data.

2. Model Building
   - Split the dataset into training and testing sets (e.g., 80% training, 20% testing).
   - Choose an appropriate classification algorithm (e.g., logistic regression, decision trees, k-means clustering, or others).
   - Use at least two algorithms to predict the outcome and compare results.
   - Train the classification model using the training set and the selected predictor variables.

3. Model Evaluation
   - Compute appropriate evaluation metrics for classification tasks (e.g., accuracy, precision, recall, F1-score, confusion matrix) on the testing set.
   - Interpret the evaluation metrics and assess the model's performance.

4. Model Diagnostics
    - Analyze the model's performance for different classes or subgroups, if applicable.
    - Investigate the impact of class imbalance, if present, and discuss potential techniques to address it.
    - Evaluate the importance or contribution of different predictor variables to the model's predictions.

5. Model Interpretation
    - Interpret the model's coefficients, feature importances, or decision rules, depending on the algorithm used.
    - Discuss the practical implications of the model and its potential applications.
    - Identify any limitations or assumptions of the model.

6. Additional Considerations
    - Identify and discuss any potential outliers in the dataset and their impact on the model.
    - Discuss strategies for handling missing values, if applicable.
    - Explain your approach to handling categorical variables and the rationale behind your chosen method.

Deliverables

a) A description of the classification algorithm used and the final model.
b) The evaluation metrics (e.g., accuracy, precision, recall, F1-score, confusion matrix) for the model on the testing set.
c) Your code file with annotations.
d) Model diagnostics, including analysis of performance for different classes or subgroups, class imbalance investigation, and feature importance evaluation.
e) Your interpretation of the model, discussing the coefficients, feature importances, practical implications, and limitations.
f) Project Final Report.
g) At least 2 visualizations (e.g., histograms, scatter plots, box plots) to explore the data.

Submission

1. Create a new repository on GitHub and upload your project files.
2. Submit the link to your GitHub repository containing all the required deliverables.

Evaluation

Your assignment will be evaluated based on the following criteria:

- Correctness of the model building and evaluation process
- Appropriate choice and understanding of the classification algorithm
- Thoroughness of the model diagnostics and performance analysis
- Clarity and insight in the model interpretation
- Proficiency in handling data preprocessing tasks, including outlier detection, missing value treatment, and variable transformations.
- Appropriate use of visualizations for data exploration and model evaluation
- Overall organization and presentation of your work
- Successful submission of the project to GitHub

Final Report

The final report for the project should encompass the following characteristics, restricted to a maximum of 10 pages:

Description of the Project:

- Provide a concise overview of the project, including its objectives, scope, and relevance within the field of study.
- Clearly state the problem or question being addressed by the project.

Literature Review:

- Summarize previous research related to the project's topic (include at least 3 references).
- Discuss key findings, methodologies, and theories from existing studies that inform the current project.

Methodology:

- Detail the research design and methods employed in the project, including data collection procedures, variables measured, and any tools or techniques utilized.
- Provide rationale for the chosen methodology and explain how it aligns with the project's objectives.
- Describe any limitations or constraints faced during the research process.

Decision for Modeling:

- Justify the choice of modeling techniques or algorithms used in the project.
- Explain how the selected modeling approach addresses the research question and is appropriate for the dataset and analysis.

Interpretations of Results:

- Present and interpret the findings of the analysis, including any statistical results or model outputs.
- Discuss the implications of the results in relation to the research question and objectives.
- Address any unexpected findings or limitations encountered during the analysis.
- Offer recommendations for future research or practical applications based on the results.

Data Repositories:

Google Dataset Search: https://toolbox.google.com/datasetsearchLinks to an external site.
data.worldLinks to an external site.
statista.comLinks to an external site.
Kaggle, https://www.kaggle.com/datasetsLinks to an external site.
COVID-19 data: https://www.statista.com/topics/6084/coronavirus-covid-19-in-the-us/Links to an external site.
All CDC Data: https://data.cdc.gov/browseLinks to an external site. and https://wonder.cdc.gov/Links to an external site.
Harvard dataverse: https://dataverse.harvard.edu/

This list is not exhaustive.