STAT 650 - Project 2
Data Requirements for Regression Analysis

Yoonsung Jung

October 2024

Introduction

This guideline outlines the data requirements necessary to perform a comprehensive regression analysis, ensuring the validity and reliability of the results.

1. Data Quality

1.1 Completeness

- **No Missing Values**: The dataset should have minimal or no missing values. If there are missing values, they should be handled appropriately through imputation or removal.

1.2 Accuracy

- **Correct Data Entries**: Ensure that the data entries are accurate and free from errors or inconsistencies.
- **Validation**: Validate the data against reliable sources to ensure its accuracy.

1.3 Consistency

- **Uniform Data Format**: The data should be in a consistent format, especially categorical variables which should be encoded uniformly.

2. Data Types and Variable Selection

2.1 Types of Variables

- **Numerical Variables**: These include continuous variables (e.g., age, salary) and discrete variables (e.g., number of children).

- **Categorical Variables**: These include ordinal variables (e.g., education level) and nominal variables (e.g., gender, marital status).

2.2 Dependent and Independent Variables

- **Dependent Variable (Target)**: The variable that you are trying to predict or explain (e.g., house price, employee salary).
- **Independent Variables (Predictors)**: The variables that are used to predict the dependent variable (e.g., square footage, number of bedrooms).

3. Data Pre-Processing

3.1 Handling Missing Values

- **Imputation**: Use appropriate imputation methods to fill in missing values (e.g., mean, median, mode for numerical variables; most frequent category for categorical variables).
- **Removal**: If imputation is not feasible, consider removing rows or columns with a high percentage of missing values.

3.2 Encoding Categorical Variables

- **Label Encoding**: Convert categorical variables into numerical values using label encoding.
- **One-Hot Encoding**: Use one-hot encoding for nominal variables to avoid ordinal relationships.

3.3 Scaling and Normalization

- **Feature Scaling**: Apply feature scaling techniques such as standardization (z-score normalization) or min-max scaling to normalize the data.

4. Exploratory Data Analysis (EDA)

4.1 Descriptive Statistics

- **Summary Statistics**: Calculate mean, median, mode, standard deviation, and other summary statistics for numerical variables.
- **Frequency Distribution**: Analyze the frequency distribution of categorical variables.

4.2 Data Visualization

- **Histograms**: Plot histograms for numerical variables to understand their distributions.
- **Box Plots**: Use box plots to detect outliers and understand the spread of numerical data.
- **Scatter Plots**: Create scatter plots to visualize relationships between pairs of numerical variables.
- **Correlation Matrix**: Generate a correlation matrix to identify correlations between numerical variables.

5. Checking Assumptions of Regression Models

5.1 Linearity

- **Linear Relationship**: Ensure that there is a linear relationship between the dependent and independent variables for linear regression models.

5.2 Independence

- **Independent Observations**: The observations should be independent of each other.

5.3 Homoscedasticity

- **Constant Variance of Errors**: The residuals (errors) should have constant variance across all levels of the independent variables.

5.4 Normality

- **Normal Distribution of Errors**: The residuals should be approximately normally distributed.

5.5 Multicollinearity

- **Low Multicollinearity**: The independent variables should not be highly correlated with each other.

6. Advanced Data Requirements for Specific Regression Models

6.1 Polynomial Regression

- **Non-Linearity**: Ensure that the relationship between the dependent and independent variables is non-linear.

6.2 Logistic Regression

- **Binary Outcome**: The dependent variable should be binary (0 or 1).

6.3 Regularization Techniques (LASSO, Ridge, Elastic Net)

- **Multicollinearity**: Regularization techniques are beneficial when multicollinearity is present among the independent variables.

6.4 Quantile Regression

- **Quantiles**: The dataset should allow for the prediction of conditional quantiles.

6.5 Poisson and Negative Binomial Regression

- **Count Data**: The dependent variable should be count data (non-negative integers).
- **Overdispersion**: For Negative Binomial Regression, the count data should exhibit overdispersion.

6.6 Cox Regression

- **Survival Data**: The data should be suitable for survival analysis, with time-to-event and event occurrence information.

6.7 Partial Least Squares Regression and Principal Component Regression

- **High-Dimensional Data**: These techniques are useful for high-dimensional datasets with more predictors than observations.

7. Selected Model Evaluation and Validation

7.1 Train-Test Split

- **Data Splitting**: Split the dataset into training and testing sets to evaluate model performance.

7.2 Cross-Validation

- **K-Fold Cross-Validation**: Use k-fold cross-validation to assess the model's performance more robustly.

7.3 Performance Metrics

- **Regression Metrics**: Use metrics such as R^2, RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) for regression models.
- **Classification Metrics**: Use metrics such as accuracy, precision, recall, F1-score, and ROC-AUC for logistic regression.