Regression Analysis

(You are expected to give a descriptive title)

Contents

Part I: EDA and Feature Engineering (for week #5)

1 Data Set

Choose a data set that has at least four categorical variables and four numerical variables. The sample size should be at least 200. You can find a data set either from my teaching data repository or other data sources. The data set should be cross-sectional (i.e., each of the data points must be observed/collected/generated at the same time).

2 Description of Data

The following information about the data should be provided in the report:

- A brief description of the data source.
- How the data set is generated or collected.
- Number of variables and their type (categorical or numerical), and size of the data set.
- List the variable names and their description/definitions.

3 EDA and Feature Engineering

Perform the standard EDA, such as distribution for categorical and numerical variables, respectively, the relationship between two variables (combinations of categorical and numerical variables), and the pairwise relationship. Keep in mind that the pairwise scatter plot is only meaningful for numerical variables.

For each EDA and associated representation, you should

- interpret what you observed and the implications of potential feature engineering
- Perform feature engineering based on EDA by writing an R/Python function.
- Write a main function to wrap individual feature engineering functions.
- Test the main function with different patterns in the components and ensure it produces the expected result.

Part II: Regression Analysis (for week #6)

4 Statistical Regression Modeling

Linear and logistic regression models are the most commonly used models in classical statistics. This part of the assignment uses the traditional statistical approaches to modeling, building, and implementation.

4.1 Linear Regression Models

Choose a continuous variable as a response to perform a linear regression analysis. Please use several subsections to organize your analysis that contain the following components.

- Statement of the question(s), the purpose of this analysis: association analysis or predictive analysis?
- Justify whether the data set has sufficient information to address the question(s)
- Model building process: initial model, diagnostics, further transformations (in addition to the one in the EDA), key performance metrics of model assessment, and final model selection (based on appropriate performance metrics). You are expected to
  - create a model that contains a few practically important variables
  - create a model that includes additional variables that potentially influence the response
  - use certain variable selection methods to identify the optimal model (i.e., the final model)
- Interpretation of regression coefficients. If you transformed your response variable, you need to do some algebra to convert the transformed response variable back to the original scale before you interpret the regression coefficient.
- Summary/discussion/recommendation

You could open a subsection for each bullet point.

4.2 Logistic Regression Analysis

Choose a binary variable as a response to perform a logistic regression analysis. If your data set does not have a binary categorical variable that can be used for the logistic regression model, you can dichotomize a continuous response **in a meaningful way** and then build a logistic regression model with the dichotomized variable.

Please use several subsections to organize your analysis that contain the following components.

- Statement of the question(s), the purpose of this analysis: association analysis or predictive analysis?
- Justify whether the data set has sufficient information to address the question(s)
- Model building process: initial model, diagnostics, transformation and scaling (in addition to the one in the EDA), key performance metrics of model assessment, and final model selection (based on certain performance metrics). For practice, you are expected to
  - create a model that contains a few practically important variables
  - create a model that includes additional variables that potentially influence the response
  - use certain variable selection methods to identify the optimal model (i.e., the final model)
- The interpretation of the final model: interpret the regression coefficient and applications of the model.
- Summary/discussion/recommendation

You could open a subsection for each bullet point.

Part III: Report Revision- Including Cross Validation (for week #7)

5 Predictive Modeling

The idea is to use **data-driven approaches** to data splitting and then apply cross-validation methods to select the final model from a pool of candidate models based on **predictive performance metric** such as **MSE** for linear regression models and **accuracy**, **sensitivity**, or **specificity** for logistic regression models.

Suggested Components in the Predictive Analysis

- *random splitting* - using random splitting for all data partitions.
- *Two-way data splitting* - data split into 75% for training and validation and 25% for testing.
- *5-fold cross-validation* - using a 5-fold cross-validation algorithm on the training data

5.1 Prediction Linear Regression

The primary predictive performance metric for linear regression modeling is the mean square error (the average squared error between predicted and the observed values of the response variable in its original scale).

Other predictive performance metrics that can also be used are R^2 or R^2_{adj}.

Likelihood-based metrics such as AIC and SBC can be used if the likelihood functions of all candidate models are at the same scale. These measures are not as intuitive as the MSE since MSE is a squared '**distance**' in the Euclidean space.

If the response variables in all candidate models are at the same scale, the MSE is expected to be used in the cross-validation for model selection.

5.2 Logistic Predictive Modeling

The primary tool for assessing the global predictive performance of logistic models is ROC curve analysis (this includes the area under the ROC curve - AUC). ROC curve suggested for this assignment.

Other predictive performance measures that can be considered are **accuracy**, **sensitivity**, and **specificity**.

Reporting ROC and AUC is required when comparing candidate models.

5.3 Testing Performance

After the final model is identified, you need to use the 25% testing data set to report the performance of the corresponding models on the **new data**.