

Project One: Part II - Regression Analysis

(You are expected to give a descriptive title)

Contents

1	Classical Regression Modeling	1
2	Linear Regression Models	1
3	Logistic Regression Analysis	2

1 Classical Regression Modeling

To address the practical questions formulated last week, we employ **both** linear regression and logistic regression models, utilizing the analytical dataset created previously. This assignment builds upon **last week's** exploratory data analysis (EDA) and feature engineering work and will **later** be combined with **next week's** task on predictive modeling and cross-validation to form a comprehensive project report.

This assignment focuses on the classical regression analysis.

2 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.

3 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