# **Hands-on Assignment 2**

Your report should document each step that you have taken and reflect your understanding of the concepts, theories, rules, and procedures involved in the specific data mining tasks.

Part 1:

## **Discuss how to develop Multiple Linear Regression models, including details on data partitioning, model setting, feature selection methods, etc.**

Multiple linear regression is a key concept in statistical modeling used to predict the impact of multiple predictors on a single outcome. It is widely applicable in various disciplines and as a statistical method for data analysis and interpretation. This technique can account for the influence of numerous independent variables on an outcome. However, multicollinearity among predictors needs to be considered. Unlike simple linear regression, multiple linear regression models include multiple predictors, allowing researchers to include relevant variables. Developing a clear model for interpretation and prediction is essential for successful regression modeling. Pitfalls in multiple linear regression include neglecting collinearity and using variables that do not meet regression assumptions, which can bias results. Resolving collinearity and verifying assumptions are necessary before proceeding with model development.

Before creating regression models, data must be partitioned into training, validation, and test sets. The training set is used to create the model, the validation set checks performance, and the test set generates final statistics. Sets must be representative of the data to avoid poor model performance. Techniques like stratification and random sampling can be used. Overfitting occurs when the model fits training data well but not test data. More data can improve performance, but sample size, predictor variables, and effect size must be considered. Balancing data across sets ensures a robust model.

When developing a multiple linear regression model, one sets the regression equation, identifying predictors and determining if the relationship is linear or nonlinear. In this guide, we only consider linear associations. Parameters of interest are the true constants used to predict values, including the intercept. The relationship between independent variables and Y is linear with its own unique slope parameter. The purpose of this investigation is to analyze the relationship between financial distress and selected firms.

The right features are crucial for model impact. Choosing features can lead to underfitting or overfitting problems. Underfitting causes poor predictive accuracy due to oversimplicity, while overfitting generates overly complicated models. Features can be discarded or generalized to solve these issues. Feature selection can be manual, hybrid, or algorithmic. Techniques like backward elimination and stepwise regression are used in manual approaches. Informed domain knowledge is important for manual selection. Algorithmic approaches utilize automated algorithms like lasso and ridge regression. The cost of interpretability is a critical factor. Fewer variables enhance efficiency and interpretability, leading to better decision-making. Selecting the right features is important for robust multiple linear regression models.

Metrics such as R-squared, adjusted R-squared, and RMSE will be used to assess model performance. Additionally, residual analysis will be employed to assess model fit and identify potential issues. Significance tests and validation techniques will be used to evaluate overall model performance and generalizability. Lastly, interpreting regression results and avoiding incorrect inferences will be discussed. The ultimate goal is to determine the effectiveness and nature of the relationship between explanatory and dependent variables. Evaluating regression studies should inform decision-making by extracting the signal from the noise.

## Build 3 multiple linear regression models using 3 different feature selection methods, and for each feature selection method used, you should analyze the best subset candidates, justify and select the corresponding best subset model.

## Compare and evaluate the 3 selected best subset models and discuss which one is a better model. Note that it is possible that different feature selection methods may lead to the same best subset model.

From the output generated in R, we can review metrics and other useful information like residuals, coefficients, residual standard error, multiple r-squared , F-statistic, P-value, RMSE or Root Mean Squared Error, and more. In all of the three best selected subset models we find the following variables are common across the three models CRIM, ZN, INDUS, CHAS, NOX, RM, DIS, RAD, TAX, PTRATIO, LSTAT, and CAT. MEDV. To compare models we will be working with the holdout set and reviewing relevant metrics to assess which subset model has the best performance on validation data.

For the model that applied stepwise regression (both directions) we can easily observe that the predictors are CRIM, ZN, INDUS, CHAS, NOX, RM, DIS, RAD, TAX, PTRATIO, LSTAT, and CAT. MEDV. This model has an adjusted r-square value of 0.8464. Regarding error between predicted and actual values, this model has a RMSE of 3.892 (rounded). This model has a p-value of <2.2e-16 which is very close to zero. The p-value for this model is statistically significant and worth analyzing because it is able to explain variance well.

For the model that applied stepwise regression (backward) we can easily observe that the predictors are CRIM, ZN, INDUS, CHAS, NOX, RM, DIS, RAD, TAX, PTRATIO, LSTAT, and CAT. MEDV. Similar to the previous model, this model has an adjusted r-square value of 0.8464. Regarding error between predicted and actual values, this model has the highest RMSE result of 3.897 (rounded) which is not ideal when striving to obtain predictive accuracy. This model has a p-value of <2.2e-16 which is very close to zero. The p-value for this model indicates that it is statistically significant and worth analyzing because it is able to explain variance well.

For the model that applied stepwise regression (forward) we can easily observe that the predictors are CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, LSTAT, and CAT. MEDV. When compared with the previous model, this model included the AGE variable. This model has an adjusted r-square value of 0.8459 which is lower than the first two models. This model's 0.0005 drop in adjusted r-squared suggests that it may be overfitting due to the addition of the AGE variable that failed to improve or add benefit to the model. Regarding error between predicted and actual values, this model has a RMSE of 3.892 (rounded) and surprisingly matches the RMSE from the model with stepwise regression in both directions. This model has a p-value of <2.2e-16 which is very close to zero and indicates this model is statistically significant to analyze.

Ideally we want a model that is able to balance these metrics. For example, we might want to have the highest r-squared outcome, low RMSE, meet statistical significance and or simplicity goals. Assessing relevant goals can be helpful to figure out which metric to prioritize. If we prioritize prediction accuracy then we want to go for models with lower RMSE values, such as the models that applied stepwise regression in “both directions” or “forward.” If we prioritize better model fit, then we want to go for models that have higher adjusted r-squared values, such as the models with stepwise regression in “both directions” or “backward.” All models exhibit similar residual standard errors (around +/- 3.628), but the model with forward selection has a slightly better fit in terms of residual spread. Also, all models can explain about ~85% variance in the response variables. After analyzing all three models across various metrics, it seems that the models with “backward” and “both directions” selection are the two most robust because they have better predictive performance overall. To help us choose a single model, we may decide to solely focus on the adjusted r-squared metric and ensure we reduce the risk of excluding important variables in a simple manner.

## Analyze what predictors are included in the final selected model. Recap what you have learned in the data exploration step (i.e., hands-on report 1) regarding predictors and discuss how you can use what you have learned in report 1 to reduce the number of predictors.

During the data exploration step, we reviewed the structure and distribution of the dataset. In addition, we used visualizations to explore the predictors and their relationships. In the hands-on report, we determined LSTAT was negatively correlated with MEDV and RM was positively correlated with MEDV. The correlation matrix showed that certain variables, like TAX and RAD were highly correlated. Variables that are highly correlated can introduce multicollinearity and can be removed without affecting model performance. The heatmap showed weaker correlations between MEDV and ZN or INDUS which could also be excluded.

The predictors in the final selected model using backward stepwise regression include CRIM, ZN, INDUS, CHAS, NOX, RM, DIS, RAD, TAX, PTRATIO, LSTAT, and CAT. MEDV. The significant values included LSTAT and RM, as indicated by the data preprocessing. Predictors like ZN and INDUS are retained, but on the borderline of significance.

## Discuss the recommended model and its performance. Interpret the results of the model, make recommendations/suggestions.

In order to select the best model, the best measure to base this decision on is adjusted R squared. Adjusted R squared is a modification of R squared adjusted for the number of predictors in a model. Both of the models we analyzed gave us an adjusted R squared of 0.8464, and the third one had an adjusted R squared of 0.8459. We chose the model given to us by the backward method with an adjusted R squared of 0.8464 as the recommended model, as it has the highest adjusted R squared. This result means that around 85% of the variation in the output variable is explained by the input variables. This selected model also provided us with a residual standard error of 3.628. Residual standard error measures the standard deviation of the residuals in the model. The p value provided by this model is also extremely low p-value: < 2.2e-16, indicating that the model is statistically significant.

Variable Age was not included in any of the models, indicating that this variable is not a good predictor for the model. We constructed a third model forcing variable age, and it gave us an even lower adjusted R squared and a higher standard error. In the future, we can exclude this variable from the analysis and save the costs associated with researching this variable.

The combination of both a high adjusted R squared, and low p-value indicate that this model is a good fit, and it is statically significant, indicating the predictors are doing a good job predicting the output variable.

## Provide details such as screenshots of ASDM or R procedures, graphs, outputs, and **justifications** of your actions, etc.

## **Part 2: Ch. 10 Logistic regression example**

Review the textbook section covering the example of using logistic regression model to classify flights that are likely to be delayed. In the same Word file, summarize the whole data mining process, what was done in each step, the rationale and justification for each action taken. Describe and evaluate the selected model and elaborate how this model can be used in practice. Summarize what you have learned from this example.

In the logistic regression example presented in the textbook, six predictors are used to determine if a flight in the dataset will be delayed. The six predictors are day of the week, departure time, origin, destination, carrier and weather. The binary outcome variable is delayed, and is coded 1 for delayed and 0 for otherwise.

**Data mining process**

Data Visualization

What was Done: In the data visualization step, the data relationships were explored through visuals including bar graphs and a heatmap.

Rationale and Justification: Data visualization helps show relationships between the variables which will assist with later steps. We want to ensure important variables are included in the model and irrelevant ones are not selected.

Data Preprocessing

What was Done: In the data preprocessing step, the data was partitioned into a training set (60%) and validation set (40%). In this case, data was already cleaned and categorical variable transformed into dummy variables. If there were any outliers, they should have been removed as well.

Rationale and Justification: Converting the categorical variables into dummy variables is essential for logistic regression models as it requires numerical inputs. In addition, the training set is used to fit a model and the validation set is used to evaluate the model’s performance.

Model Fitting and Estimation

What was Done: Using the six categorical predictors mentioned above, the estimated logistic regression model was fitted to the dataset.

Rationale and Justification: To determine the categorical outcome (delayed-1 or other-0) the logistic regression model interprets the outcome and how each predictor impacts the likelihood of delay.

Model Interpretation

What was Done: After fitting the model, the coefficients are analyzed to understand how each predictor impacted the delay and to what extent.

Rationale and Justification: Interpretation of the model is essential for understanding the model and which predictors are most important.

Model Performance

What was Done: Model performance was interpreted by the confusion matrix, error rates and a lift chart.

Rationale and Justification: Evaluating the performance of the model ensures the model can generalize well to new data beyond the training set. In addition, the confusion matrix can give insights into the number of false positives and false negatives.

Variable Selection

What was Done: In this step, variables are analyzed to see if they could be removed or merged with another variable. In addition, selection methods such as stepwise selection, forward selection and backward elimination can be used to eliminate insignificant variables.

Rationale and Justification: By simplifying the model, we can increase interpretability and reduce overfitting.

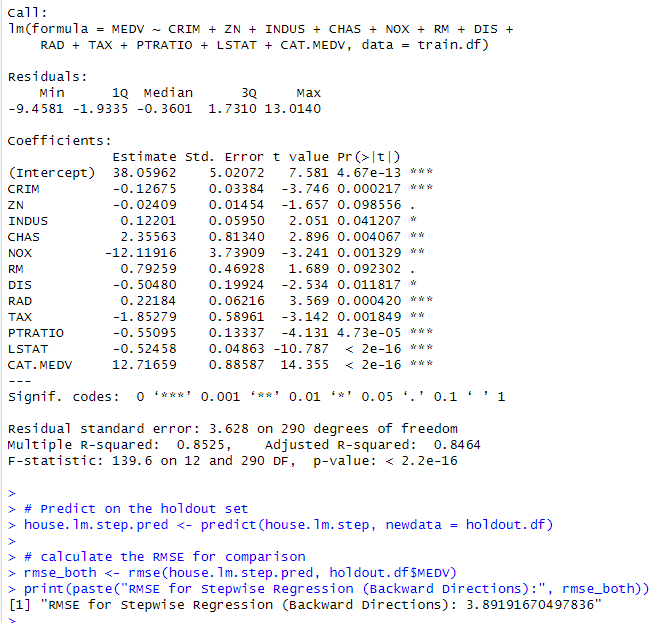
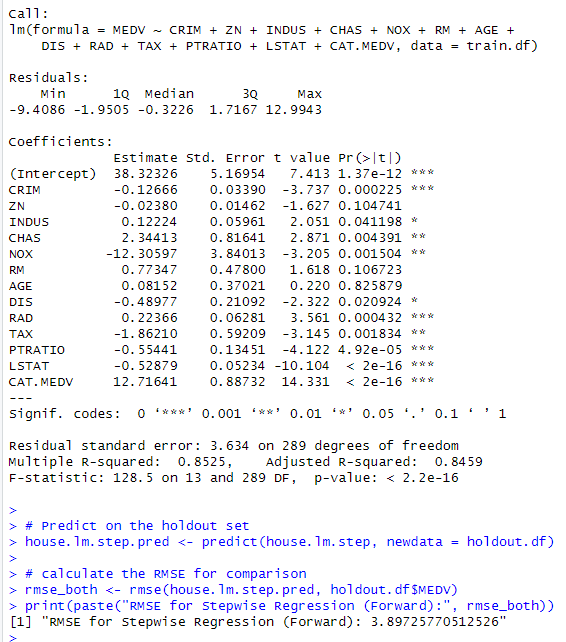
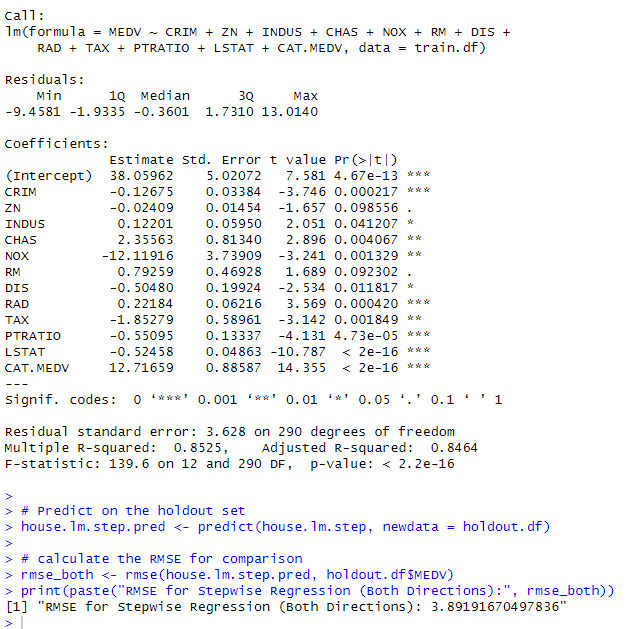
**Selected model**

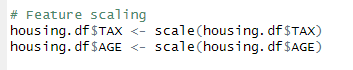
The selected model uses six predictors that need knowledge of the carrier, day of the week, hour of the day and whether the day had inclement weather. This model shows that those predictors are most likely to influence a delay.

This model can be used integrated into an airline’s system to predict delays. It could be used to proactively manage delays, improve customer satisfaction by communicating delays in advance and optimize resource allocation by focusing on flights that are more likely to be delayed.

**Conclusion**

This example highlighted the importance of data preprocessing, model performance and variable selection for building an effective logistic regression model. Proper handling of categorical values is essential for a logistic model. The model’s interpretability helps understand the driving factors of the outcome. The model’s performance metrics ensures the model can be used with any dataset. The variable selection reduces the model’s complexity and improves performance without decreasing interpretability.





Correlation heat Map