W5451 – Statistical Learning with R and Python

winter semester 2022/2023

Topic. Application Of Different Multiple Linear Regression Models Using R Programming.

Examiner: Prof. Dr. Yuanhua Feng

Assistant: Mr. Sebastian Letmathe

Project beginning: November 28, 2022

Project due: January 16, 2023, 11:59 p.m.

Submitted by Group No. 06:

Name	Matriculation Number	Contribution
Trusha Patel	6872844	• Code and text for all problems

Contents

Introduction

Statistical learning is a branch of statistics that deals with the use of statistical models to analyze and understand patterns in data. It is a broad field that encompasses a wide range of techniques, including supervised and unsupervised learning methods, for gaining knowledge, constructing models, making predictions, and making decisions. The goal of statistical learning is to find useful patterns in data that can be used to make informed decisions, predictions, or generalizations about the underlying data. It is a powerful and flexible approach that can be applied in a wide variety of fields such as finance, healthcare, marketing and many more.

Under supervised learning focus is on building a statistical model for estimating an output based on one or more inputs. Whereas unsupervised learning involves inputs but no supervising output. Researchers are interested in observing the relationships and structure from such unsupervised data.

One of the most used supervised statistical learning approaches for making predictions is multiple linear regression model. It's a generalization of simple linear regression in which there are more than one predictor variable. There are several cases of this model such as polynomials, interactions, and categorial predictor variables(Ambrosius, 2007). The Multiple linear regression is based on same assumptions as simple linear regression.

This study firstly provides a brief explanation of the contexts from chapter 2 to 5 of book : An Introduction to Statistical Learning with Application in R (James et al., 2013). Then multiple linear regression models were fitted with different combinations of variables in two examples and conduct an empirical analysis. Finally, some concluding comments were provided on the overall study.

Summary of Chapters 2-5

2.1 Chapter 2: Statistical Learning

Statistical learning is a set of approaches for estimating f. Suppose we observe a quantitative response Y and p different predictors, $X_1, X_2, \dots, X_p$. According to our assumption the relationship between Y and X, can be written as $Y = f(X) + \epsilon$. Here, f is some fixed but unknown function of X, and ϵ is a random error term, which is independent of X and has mean zero. The f is estimated for two reasons: prediction and inference. Since in the setting error term is average to zero, Y can be predicted by using $\hat{Y} = \hat{f}(X)$. The accuracy of $\hat{Y}$ depends on two quantities: reducible error and irreducible error. As the irreducible error bound is almost always unknow while in practice, to improve the accuracy of prediction estimate f must focus on minimizing reducible error. In inference setting, the focus diverts towards knowing the exact form of X and understanding the association between Y and X rather than making predictions for Y.

For estimating f the statistical methods used are characterised as either parametric or non-parametric. Under parametric method, assumption regarding the functional form is made and then a method is chosen to estimate model parameters. Those parameters are obtained by fitting selected model to training data. Most used method for estimation is ordinary least square (OLS). In non-parametric methods assumptions are not made for the functional form of f rather the

idea is to seek estimate of f that gets close to the data points without being too rough. This method is more useful for complex functional forms.

Variables are characteries in two categories quantitative or qualitative. The problems with quantitative responses are regression problems, while those with qualitative response are classification problems. The least squares linear regression is used with quantitative response whereas methods like logistic regression, K-nearest neighbors and boosting can be used for either quantitative or qualitative responses. Selecting the best approach is challenging and most important part of performing statistical learning in practice. Some of the most important concepts in this regard are measuring the quality of fit, the bias-variance trade off and classification setting.

2.2 Chapter 3: Linear Regression

Linear regression is parametric approach widely used statistical learning tool for predicting quantitative response. Simple linear regression assumes that there is approximately a linear relationship between X and Y, modeled as $Y \approx \beta_0 + \beta_1 X$. As β is unknown, data must be used to estimate the coefficients. Least square approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize RSS. $Y = \beta_0 + \beta_1 X + \varepsilon$ is the population regression line which is best linear approximation to the true relationship between X and Y. The accuracy of estimates of β_0 and β_1 is assessed by computing the standard errors, these errors can be further used to compute the confidence intervals and hypothesis testing. The general hypothesis usually tested is whether there exists a relationship between X and Y or not. Mathematically, $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 \neq 0$, the decision depends on how far standard deviations is from 0 (using t-statistic).

A multiple linear regression model with p predictors is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$, where X_j represents j^{th} predictor. β_j is interpreted as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed. Variables can be selected by using forward, backward, and mixed selection methods. Nonlinear relationships are accommodated by polynomial regression. Potential problems face while fitting a linear regression model are non-linearity of response predictor relationships, correlation of error terms, non-constant variance of error terms, outliers, high leverage points and collinearity.

One of the best and simplest non-parametric method is K-nearest neighbors' regression (KNN regression). Given a value K and prediction point x_0, KNN first identifies $\mathcal{N}_0$, the K training observations that are close to x_0. Then $f(x_0)$ is estimated using the average of all the training responses in $\mathcal{N}_0$. Mathematically, $\hat{f}(x_0) = \frac{1}{K}\sum_{x_i \in \mathcal{N}_0} y_i$. The optimal value for K depends on the bias-variance trade-off. KNN performs only a little worse than least square regression when value of K is large, while performs far worse when K is small.

2.3 Chapter 4: Classification

The process of predicting qualitative response is known as Classification. For a binary qualitative response, its still appropriate to use linear regression. Logistic regression models the probability that Y belongs to a particular category. To fit the non-linear model's maximum likelihood is used. The estimates are chosen to maximize this likelihood function. For multiple

predictors, multiple logistic regression model is effective. The multinomial logistic regression classifies a response variable that has more than two classes.

There are generative models like linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and Naive Bayes for classification. The LDA classifier results from assuming that the observation within each class comes from a normal distribution with specific mean and common variance and plugging the estimates for this parameters in Bayes classifier. In the case of multiple predictors, LDA assumes a multivariate Gaussian (or multivariate normal) distribution with some correlation between each pair of predictors. QDA classifier results from assuming a gaussian distribution and plugging estimates for the parameters into Bayer's theorem to perform prediction. Naive Bayes classifier makes a single assumption that within the kth class, the p predictors are independent. Its powerful as it considers not only marginal distribution of each predictor but also joint distribution of predictors. When the true decision boundaries are linear, LDA and logistic regression performs well, whereas moderately non-linear, QDA or naive Bayes provides better results.

2.4 Chapter 5: Resampling Methods

Resampling methods are one of the indispensable tools of modern statistics involved in repeatedly drawing samples form a training set and refitting a model on each sample to obtain additional information about the fitted model. It could be computationally expensive as it involves fitting the same statistical method multiple times using different subsets of the training data. Most used resampling methods are cross-validation and the bootstrap.

Cross validation (CV) is used to estimate the test error associated with a given statistical learning method to evaluate its performance, or to select the appropriate level of flexibility. This process of model performance evaluation is called model assessment. The process of selecting the proper level of flexibility for a model is called model selection. CV simple strategy called validation set approach involves in randomly dividing available observations into two sets, training and validation set. The model is fitted on training set, the resulting validation set error rate is assessed using MSE and provides estimate of test error. Another CV strategy is Leave-one-out cross-validation (LOOCV) where a single observation is used for validation set and remaining observations makes up the training set. LOOCV is very general method and can be used with all kinds of predictive models. Lastly, K-fold cross-validation approach which involves randomly dividing the set of observations into k groups or folds, of approximately equal size. First fold is treated as validation set and method is fit on the remaining k-1 folds. The bootstrap is a powerful tool that can be used to quantify the uncertainty associated with a given estimator or statistical method.

Detailed Description of Multiple Linear Regression

In practice more often considering more than a one predictor variable is important. The multiple linear regression model is used to directly accommodate multiple predictors. Mathematically model is represented as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{\mathrm{p}} X_{\mathrm{p}} + \epsilon,$$

Where X_j represents the j^{th} predictor and β_j quantifies the association between variable and the response. Thus β_j is interpreted as the average effect on Y of a one unit increase in X_j, while all other predictors are fixed. The regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are unknown and must be estimated. Those given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ is then used for prediction.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

The least square approach is used to estimate the parameters. The regression coefficients are chosen to minimize the sum of squared residuals. The RSS can be mathematically calculated as follows $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2$$

The value of estimated coefficients that minimize RSS are the multiple least squares regression coefficient estimates.

Few questions while performing multiple linear regression are important to answer.

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
   In multiple regression setting with p predictors, its important to know whether all the regression coefficients are zero i.e., $\beta_1 = \beta_2 = \ldots = \beta_p = 0$. The hypothesis test is set were, $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ verses $H_a : at\ least\ one\ \beta_j\ is\ non - zero$. This test is performed by computing F-statistic. When there is no relationship between the response and predictors, F-statistic take on a value close to 1. This approach works when p is relatively small and certainly small compared to n.

2. Do all the predictors help to explain Y, or is only a subset of predictors useful?
   Generally, all the predictors are associated with the response but its more often that response is only associated with a subset of predictors. The task called variable selection is used to determine which predictors are associated with response, to fit a single model involving only those predictors. Ideally, to perform variable selection different models are tested each containing different subset of the predictors.

   Variable selection can be realized by using methods such as forward selection, backward selection, or mixed selection. Forward selection starts with a null model and adds variables one at a time based on the lowest residual sum of squares (RSS). Backward selection starts with all variables in the model and removes the variable with the largest p-value. Mixed selection is a combination of forward and backward selection and can remedy the limitations of both methods. It is important to note that variable selection can become computationally infeasible for large numbers of predictors. The model quality then further be judged using criteria´s like Mallow´s C_p , Akaike information criterion (AIC), Bayesian information criterion (BIC) and adjusted R^2.

3. How well does the model fit the data?
   The most used numerical measures of model fit are the RSE and R^2. An R^2 is the square of the correlation of the response and variable. The R^2 value close to 1 indicates that model

has a large portion of the variance in the response variable. The RSE can be defined as $RSE = \sqrt{\frac{1}{n-p-1}RSS}$. The models with more variables tend to have higher RSE if the decrease in RSS is small relative to increase in p. If the RSE value is very close to the actual outcome value, then model fit the data well.

4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?
   The multiple regression model has three types of uncertainties: reducible error, which is related to the accuracy of the coefficient estimates, can be measured by confidence intervals; model bias, which exists as a linear model almost always an approximation of reality, and can be avoided by estimating the best linear approximation to the true surface; and irreducible error, which is caused by random error in the model and cannot be completely eliminated, even if the true values of coefficients are known.

Empirical Analysis

Example 1: Wind Power Generation

The data was collected from a wind turbine's SCADA system in Turkey over the course of year 2018. The data was recorded at 10-minute intervals throughout the year (Wind Power Curve Modeling, n.d.). The number of observations is 50530. The variables are as follows:

- AP (Active Power) in kW: The power generated by the turbine at a given moment.
- WS (Wind Speed) in m/s: The wind speed that the turbine uses for electricity generation.
- TP (Theoretical Power Curve) in KWh: The power values that the turbine generates at a given wind speed as specified by the turbine manufacturer.
- WD (Wind Direction) in degrees: The direction of the wind at the hub height of the turbine.

For testing the relationship between the generated power (which is AP variable) with other variables, multiple linear regression models are fitted. A general multiple linear regression model with p predictors: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$. In this example, AP is the dependent variable being predicted and WS, TP and WD are predictor variables. Firstly, the data is separated into training and test set using createDataPartition () from library (caret). Then, the models fitted with different combinations of variables on the training set which are as follows:

- Model 1: AP = β_0 + β_1WS + β_2TP + β_3WD +ϵ
- Model 2: AP = β_0 + β_1WS + β_2WD +ϵ
- Model 3: AP = β_0 + β_1WS + β_2WD + β_3 (WS*WD) +ϵ
- Model 4: AP = β_0+ β_1WS + β_2(WS^2) +ϵ
- Model 5: AP = β_0 + β_1WS + β_2TP + β_3WD + β_4(WS^2) + β_5(TP^2) + β_6(WD^2)+ϵ

An ANOVA(Analysis of Variance) table is used to compare the fit of different models by comparing the residual variances between models. The F-test, calculated using ANOVA, determines if there's a significant difference in fit of the models. If the p-value is less than a pre-determined level of significance, one model is considered a better fit than the others.

Table 1 : ANOVA results and adjusted $R^2$for all the fitted multiple linear regression models.

Model	Res. Df	RSS	Df	Sum of sq	F-statistics	Pr(>F) ***	Adjusted R^2
Model 1	40422	6.4696e+09	-	-	-	-	0.9071
Model 2	40423	1.1484e+10	-1	-5.01e+09	32693.2	2.2e-16	0.8351
Model 3	40422	1.1101e+10	1	3.83e+08	2494.4	2.2e-16	0.8406
Model 4	40423	1.1350e+10	-1	-2.4868e+08	1621.4	2.2e-16	0.8370
Model 5	40419	6.1990e+09	4	5.151e+09	8396.2	2.2e-16	0.9110

Note: Res. Df - residual degrees of freedom, RSS- residual sum of squares, Df- degrees of freedom, Sum of sq- difference in residual sum of squares, Pr(>F)- p-value associated with the F-statistic. Code: '***' 0.001.

Multiple linear regression models' analysis is as follows:

Model 1: Is a linear regression model with AP as dependent variable and WS, TP and WD as predictor variables. According to the adjusted R^2 value of 0.9071 and very low p-value for the F-statistics, the model is a good fit. As having a very low p-value indicates that all three predictors are significant of the dependent variable AP. This model has 40422 degrees of freedom for the residuals and a residual sum of squares of 6.4696e+09.

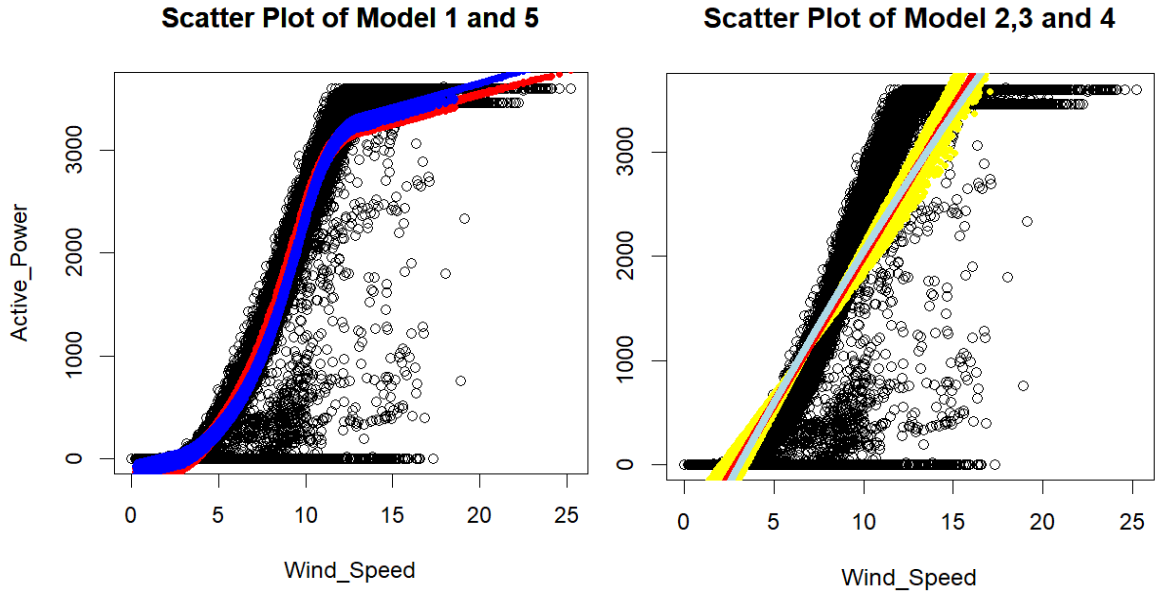
Model 2: The model is takes into consideration WS and WD predictor variables. The adjusted R^2value is 0.8351 which means 83.51% of the variance in the dependent variable AP can be explained by this model. The adjusted R^2 and R^2 has the same value which indicates that the model is not too complex. The F-statistic is also significant (p-value is less than 2.2e-16) which indicates that the overall model is significant. According to the Table 1, this model has 40423 degrees of freedom for the residuals and a residual sum of squares of 1.1484e+10, which is higher than the first model indicating that this model is not as good a fit as the first one.

Model 3: The independent variables in this model are WS, WD, and the interaction term WS*WD. The intercept term represents the predicted value of the dependent variable when all independent variables are set to zero. The coefficients value for WS and WD are 3.245 and 1.851, respectively. The coefficient for the interaction term between WS and WD is -0.268. All coefficients have p-values < 0.001 indicating they are highly statistically significant. The adjusted R^2 is 0.8406, indicating 84.06% of variation in the dependent variable is explained by independent variables. F-statistic is 7.105e+04, p-value is < 2.2e-16 indicating overall model is highly statistically significant. The model has 40422 residual degrees of freedom and residual sum of squares of 1.1101e+10, which is lower than model 2 but higher than model 1.

Model 4: The independent variables in this model are WS, I(WS^2) which is the square of WS. The coefficient value for WS is 329.3871, for (WS^2) is -2.5517. The p-values for all coefficients are less than 0.001, indicating that they are highly statistically significant. The adjusted R^2 value is 0.837, which means that 83.7% of the variation in the dependent variable is explained by the independent variables in the model. The F-statistic is 1.038e+05 and the p-value is less than 2.2e-16, indicating that the overall model is also highly statistically

significant. According to the Table 1, This model has 40423 degrees of freedom for the residuals and a residual sum of squares of 1.1350e+10, which is higher than model 1 and 3.

Model 5: A polynomial multiple linear regression model with a degree of 2 for independent variables is fitted. The coefficients for polynomial terms of WS, TP, and WD are highly statistically significant, except for the polynomial term of (WS^2). The adjusted R^2 value is 0.911, indicating that 91.1% of the variation in the dependent variable is explained by the independent variables in the model. The F-statistic and p-value are also highly statistically significant, indicating that the overall model is a good fit. The model has 40419 degrees of freedom for the residuals and the lowest residual sum of squares among all the models.



The scatter plots of Model 1 and 5 appear to be the most accurately fitted models. Model 5 (shown in blue) and Model 1 (shown in red) both closely follow the true values of the predictors. In contrast, the models represented in the scatter plots of Model 2, 3, and 4 have less accurate lines, indicating that they do not capture the true values of the predictors accurately. Overall, Model 5 is the best fitting model as it has the lowest residual sum of squares, the highest F-statistic, highest adjusted R^2 and the lowest p-value. Model 1 also has a good fit to the data, but it is not as good as Model 5. Model 2, 3 and 4 do not fit the data as well as Model 1 and 5.

Example 2: Energy Efficiency of Buildings

The dataset, created by Angeliki Xifara, a civil engineer, comprises information on energy efficiency. It was processed by the Oxford Centre for Industrial and Applied Mathematics in the UK (UCI Machine Learning Repository: Energy Efficiency Data Set, n.d.). The dataset includes 768 observations and 8 variables. The variables are as follows:

- X1 - Relative Compactness
- X2 - Surface Area
- X3 - Wall Area
- X5 - Overall Height
- X6 - Orientation

- X7 - Glazing Area
- X8 - Glazing Area Distribution
- Y1 - Heating Load

To design a building efficiently and determine the necessary equipment for comfortable indoor temperatures, the heating load must be calculated. This requires considering various building characteristics and the conditioned space. The dependent variable in this case is Y1, with the remaining variables being independent. To begin, the data is divided into a training and test set using the createDataPartition () function from the caret library. Next, various models are fitted using different combinations of variables on the training set.

- Model 1: Y1 = β_0 + β_1X1 + β_2X2 + β_3X3 + β_4X5 + β_5X6 + β_6X7+ β_7X8 + ϵ
- Model 2: Y1 = β_0+ β_1X2 + β_2X3 + β_3X5 + β_4X7 + ϵ
- Model 3: Y1 = β_0+ β_1X2 + β_2X5 + β_3(X2 * X5) + ϵ
- Model 4: Y1 = β_0 + β_1X1 + β_2X2 + β_3X3 + β_4X6 + β_5X7 + β_6X8 + β_7(X1^2) + β_8(X2^2) + β_9(X3^2)+ β_{10}(X6^2) + β_{11}(X7^2) + β_{12}(X8^2) + ϵ

Table 2 : ANOVA results and adjusted $R^2$for all the fitted multiple linear regression models.

Model	Res. Df	RSS	Df	Sum of sq	F-statistics	Pr(>F) ***	Adjusted R^2
Model 1	608	5102.9	-	-	-	-	0.9159
Model 2	611	5377.9	-3	-275.0	11.233	3.518e-07	0.9118
Model 3	612	9970.0	-1	-4592.0	562.746	2.2e-16	0.8367
Model 4	603	4920.5	9	5049.5	68.756	2.2e-16	0.9182

Note: Res. Df - residual degrees of freedom, RSS- residual sum of squares, Df- degrees of freedom, Sum of sq- difference in residual sum of squares, Pr(>F)- p-value associated with the F-statistic. Code: '***' 0.001.

Multiple linear regression models' analysis is as follows:

Model 1: The model is centered to predict a Y1 variable using a set of predictor variables (X1, X2, X3, X5, X6, X7, X8). The model has a good fit as adjusted $R^2$is 0.9159 and all predictor variables are significant with a p-value < 0.05 except X6. According to the Table 2, the model has 608 residual degrees of freedom and a residual sum of squares of 5102.9. This means that there are 608 observations in the dataset and the sum of the squared differences between the observed outcome values and the model's predicted values is 5102.9.

Model 2: The model predicts Y1 using X1, X2, X3, X5, and X7 predictor variables. It has a good fit with adjusted R^2 value of 0.9118 and all predictor variables are significant except X2 with p-value of 0.0577. The dataset has 611 observations and the sum of squared differences between observed and predicted values is 5377.9. Model 2 has 3 less degrees of freedom and a smaller residual sum of squares than Model 1. The F-statistic is 11.233 and the p-value is 3.518e-07, which is statistically significant.

Model 3 : The independent variables in this model are X2, X5, and the interaction term X2*X5. The model has a relatively strong relationship with the response variable, as indicated by the

adjusted R^2 of 0.8367. However, it's important to note that even though the interaction term (X2*X5) is significant with a p-value of 0.00146, which is less than 0.05, indicating that there is a statistically significant relationship between X2 and X5 and the response variable, the individual predictor variables X2 and X5 are not significant with p-values of 0.60255 and 0.08404 respectively. This means that model seems a good fit for the data, based on the R-squared value, but the individual effect of X2 and X5 on the response variable is not clear. Additionally, the F-statistic and p-value are lower than the previous models, indicating that this model is not fitting the data as well as the previous models.

Model 4: A polynomial model is fitted. According to the adjusted R^2 value 0.9182 and all predictor variables and quadratic terms are significant with a p-value < 0.05 except poly(X2^2) and poly(X6^2) with p-values of 0.146454 and 0.604802 respectively. The difference in residual sum of squares between model 4 and the previous model 3 is 9, 5049.5, meaning that the model 4 has 9 more degrees of freedom than the model 3 and the residual sum of squares of the model 4 is 5049.5 units less than the model 3. The F-statistic of the model is 68.756, and the p-value is < 2.2e-16. The p-value is much less than 0.05, so the model is considered statistically significant. The F-statistic is much higher than the previous models and the p-value is much lower, indicating that this model is fitting the data much better than the previous models.

Overall, Model 1 has the highest adjusted R^2 value of 0.9159, which indicates that it has a good fit to the data and it explains a large portion of the variance in the response variable. However, Model 4 has a slightly higher adjusted R^2 value of 0.9182 but it has 9 less degrees of freedom than Model 1. Model 2 has a lower adjusted R^2 value of 0.9118 but it has 3 less degrees of freedom than Model 1. Model 3 has the lowest adjusted R^2 value of 0.8367 and it has the lowest F-statistics and the highest p-value which means that it is not a good fit to the data and it does not explain the response variable well.

Conclusion

This study initial discussed the field of Statistical learning, which is considered as a powerful and flexible approach used to analyse and understand patterns in data. A brief explanation of the concepts covered in chapters 2-5 of the book "An Introduction to Statistical Learning with Application in R" such as estimating fixed but unknown function of predictors, parametric and non-parametric methods, and different methods used for quantitative and qualitative variables are provided. In general, it has been seen that selecting the best approach is a challenging and important part of performing statistical learning in practice, and mention key concepts such as measuring the quality of fit, bias-variance trade off and classification.

The study focuses on fitting Multiple Linear Regression models on two examples. In the Example 1, multiple linear regression models were fitted to test the relationship between the generated power (AP) and other variables such as wind speed (WS), theoretical power curve (TP), and wind direction (WD) collected from a wind turbine's SCADA system in Turkey over the course of 2018. The data was separated into training and test sets, and five models were fitted with different combinations of variables on the training set. The fit of the models was compared using ANOVA table. The analysis of the multiple linear regression models shows that Model 5, which is a polynomial multiple linear regression model with a degree of 2 for

independent variables, is the best fit for the data. This is determined by its high adjusted R^2 value of 0.911, indicating that 91.1% of the variation in the dependent variable is explained by the independent variables in the model. Additionally, Model 5 has the lowest residual sum of squares, the highest F-statistic, and the lowest p-value among all the models, which also confirms its good fit. Model 1 is also a good fit for the data, but it's not as good as Model 5. Model 2, 3, and 4 are not as good a fit as Model 1 and 5.

The Example 2 is based on energy efficiency dataset where dependent variable is Y1, which represents the heating load of a building. The data is divided into a training and test set using the createDataPartition () function from the caret library. Then, multiple linear regression models are fitted using different combinations of variables on the training set. The models are compared using ANOVA and adjusted R^2 same as example 1.

In conclusion, the analysis of the multiple linear regression models shows that Model 4, which is a polynomial model, is the best fit for the data. This is determined by its high adjusted R^2 value of 0.9182, indicating that 91.98% of the variation in the dependent variable is explained by the independent variables in the model. Additionally, Model 4 has a lower residual sum of squares, a higher F-statistic, and a lower p-value than the previous models, which confirms its good fit. Model 1 also has a good fit with an adjusted R^2 value of 0.9159, but it's not as good as Model 4. Model 2 and Model 3 have lower R-squared value and lower F-statistic and p-value than Model 4, indicating that they are not as good a fit as Model 4. It's worth mentioning that Model 4 is a polynomial model, this means that the relationship between the predictors and the response variable is non-linear, and this model should be used with caution when interpreting the coefficients.

It's important to note that this analysis is based on the specific data provided in the examples and the results may be different for other datasets. In addition, this analysis assumes that the data is linear and normally distributed. The model selection is a complex task that requires consideration of multiple factors, including the sample size, the number of predictors, the complexity of the data, and the goal of the analysis. Therefore, it is important to consider the assumptions, limitations, and potential sources of error when interpreting the results of the analysis.

References

- Ambrosius, W. T. (Ed.). (2007). *Topics in biostatistics*. Humana Press.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer. https://doi.org/10.1007/978-1-4614-7138-7
- *UCI Machine Learning Repository: Energy efficiency Data Set*. (n.d.). Retrieved 16 January 2023, from https://archive.ics.uci.edu/ml/datasets/energy+efficiency#
- *Wind power curve modeling*. (n.d.). Retrieved 15 January 2023, from https://kaggle.com/code/winternguyen/wind-power-curve-modeling