|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Matriculation number:** | **2** | **6** | **5** | **2** | **1** |



**Examination Assignment**

Module: Data Analysis and Statistics

Exam part: Data Analysis and Statistics

Examiner: Prof. Dr. Schwind, Dipl.-Biol. Ralf Darius Deadline for the submission: 31.08.2019, 11:59 pm

|  |  |  |
| --- | --- | --- |
| **Study program** | **Begin of studies** | **Last name, First name** |
| Information Engineering and Computer Science (M.Sc.) | SoSe – 2019 | Prajapati, Bijalben |

Assessment criteria and number of points that can be achieved:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Maximum number of**  **points** | Skills and Expertise | Systematic and scientific Quality | Quality of the results | Presentation of the results |
| **100** | 45 | 15 | 30 | 10 |

Result:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Points** | **Mark** | Skills and Expertise | Systematic and scientific Quality | Quality of the results | Presentation of the results |
|  |  |  |  |  |  |

**DECLARATION OF AUTHENTICITY**

This report is the result of my own work. Material from the published or unpublished work of others, which is referred to in the report, is credited to the author in the text.

**Bijalben Prajapati**

**Abstract**

The procedure of extracting, compiling and modelling of raw data in order to acquire structured information such that it can be applied to formulate conclusions is referred as data analysis. It can also help in the estimation of outcomes or even supporting decisions in scientific and business field. There are several data analysis methods, some which involves descriptive analysis, diagnostic analysis, predictive analysis, prescriptive analysis.

This report shows the overall report on the case study "number of factors that relate to wages for a group of males from the Atlantic region of the United States" that has been developed as a part of examination assignment for Data Analysis and Statistics. The report explains the research question, context and the basic goal. It also states methods for solving the given tasks based on R programming.

R provides a very rich computing environment for wide range of numerical calculations. Here in this case study, by using R a stochastic model is trained for the finding factors that relate to wage.

**Table of Contents**

|  |  |
| --- | --- |
| 1. Introduction    1. Motivation    2. Research questions    3. Goal    4. Context    5. Approach 2. Tools and methods    1. RStudio    2. Methods       1. Hypothesis testing       2. Normality Test       3. Correlation Test       4. Feature Selection (Wrapper Method)       5. Multiple Linear Regression       6. Dummy variable       7. Interaction Method       8. randomForest 3. Result 4. Implementation    1. Relationship using Hypothesis Testing    2. Type and strength of relationship    3. Level of accuracy    4. Which factors contribute to wage?    5. Is the relationship linear?    6. Are there interaction effects? 5. References   Appendix | 7  7  7  8  8  9  9  9  9  9  10  10  11  12  12  12  12  13  14  14  15  16  17  17  19  21  22 |

**List of Figures**

|  |  |
| --- | --- |
| Fig 2.2.4: Wrapper Method working flow  Fig 4.3: Output of prediction model  Fig 4.5.1: Scatter Plots Matrix  Fig 4.5.2: Prediction using Random Forest (Finalized Model)  Fig 4.6.1: Interaction between year and education  Fig 4.6.2: Interaction between age and education  Fig 4.6.3: Interaction between age and year | 11  16  17  18  19  19  20 |

1. **Introduction**

The Current Population Survey (CPS) is venerable and justifiable surveys in the United States. Its stipulate information such as work, wages and education which decide people differ and as a community. The Data specified in the Introduction and The Study System sections are taken from the Examination Assignment provided by Dipl.-Biol. Ralf Darius, dataset contains income and other information for 3000 male workers in the Atlantic region of the United States. This report will try to understand the dependency between an employee’s age, educational attainment as well as the calendar year and an employee’s wage (ref).

The procedure of extracting, compiling and modelling of raw data in order to acquire structured information such that it can be applied to formulate conclusions is referred as data analysis. It can also help in the estimation of outcomes or even supporting decisions in scientific and business field. There are several data analysis methods, some which involves descriptive analysis, diagnostic analysis, predictive analysis, prescriptive analysis.

* 1. **Motivation**

The Current Population Survey and the data provided help us to explore more about age, education and year of Atlantic males affects on the wages of those males. With the introduction of R programming language, it has become an easy task to achieve solutions related to understanding available data and then analyzing various related aspects. R programming is highly popular as it is open source and easily accessible. This is a very obvious choice when a thorough understanding of data using various graphs and statistical methods is needed. The available packages in R can be used to perform specific functions like partitioning the data, classifying and combining data, finding hidden layers in the data, etc. Also, R has various functions which can handle many probability functions required to accomplish our task. This report helps us to gain much deeper knowledge about Linear Regression and Scientific Programming using R. In addition to that it provides us better statistical data summarization and visualization.

* 1. **Research Questions**

A research question refers to the vital part of a research project, case study or literature review. It focuses on the study, determines the methods and manages all the different stages of inspection, analysis and reporting. For the given case study, the research questions can be categorized based on:

1. How to find relationship among variables using correlation?
2. How to handle categorical variable in regression problem?
3. What is hypothesis test?
4. How to conduct hypothesis test for correlation?
5. How to define relationships (strong week positive and negative) among variables?
6. What are various methods used for correlation?
7. How to implement normality test for given data?
8. Working of multiple linear regression.
9. How to check the performance of model?
10. What is training and testing data?
11. How to calculate root mean square error?
12. What is confidence interval?
13. Feature selection methods for regression?
14. What is dummy variable and its problem?
    1. **Goal**

The major goal of this report is to investigate how year, age and education attainment of male workers in the central Atlantic Region affect the wages of those employee. Regression Predictive Modelling and its methodology is used to find relationship of other factor to wage and which factors most contribute to wages.

* 1. **Context**

Some of the major facts and figures that have to be considered at the time of modelling and simulation process of survey data for male employee in the central Atlantic region of the USA are as follows:

* Data contains 3000 workers in the Mid-Atlantic Region from 2003 to 2007.
* Have age of males between 18 to 80.
* They are either married, unmarried or divorced.
* Everyone is from different race.
* Everyone had a minimum education qualification of High School Grade and maximum qualification of advanced degree.
* Every person is from middle Atlantic region and all are working in different job field.
* Some of them do not have health insurance and also the health condition is different for everyone according to their health.
* Details such as the wages they have are in the range of 20.08554 to 318.3424. The lowest wage they had in their life is in the range of 3 to 5.763128
  1. **Approach**

The four-stage prediction method is used for predicting accurate wages. First step is to understand the given data and then find out the noisy data if there are any. In next step, relationships and strength between depended variable and independent variables is found. The following step is feature selection. In fourth phase, models are created using training data and wages are predicted using test data. Lastly, best fit model is selected by considering error ratio.

1. **Tools and methods**
   1. **RStudio**

**RStudio** is an IDE for R which is free and an open source developed by **RStudio, Inc.** It is written in C++, Java, JavaScript and uses Qt framework for its GUI. It supports direct code execution and is used for statistical computing and graphics. RStudio runs on all major desktop platforms like Windows, Mac OS, Ubuntu, etc. It has own server **RStudio Server** can also be used as a server which provides users to access IDE using a web browser. Nowadays RStudio is a more comfortable workbench for working in [SQL](https://blog.rstudio.com/2018/10/02/rstudio-1-2-preview-sql/), [Stan](https://blog.rstudio.com/2018/10/16/rstudio-1-2-preview-stan/), [Python](https://blog.rstudio.com/2018/10/09/rstudio-1-2-preview-reticulated-python/), and [D3](https://blog.rstudio.com/2018/10/05/r2d3-r-interface-to-d3-visualizations/).  (RStudio Team, 2019)

* 1. **Methods**

There are ample packages and functions in R for explore and analysis data. These are the methods and function used for data analysis which explained in detail as follows:

* + 1. **Hypothesis Testing**

A statistical hypothesis is the distribution of value. Similarly, a statistical hypothesis specifies a possible set of distribution of value which is for true statement. A hypothesis that specifies a single distribution for value is called **simple;** a hypothesis that specifies more than one distribution for value is called **composite** (DataFlair Team, 2019).

There are two types of statistical hypotheses.

* **Null hypothesis**: The observations result purely from chance. It is denoted by Ho.
* **Alternative hypothesis**: The observations are influenced by some non-random cause. It is denoted by Ha.

The strength of evidence in support of hypothesis is measured by the **p-value** (Burdess. N, 2010).

H0🡪 p-value = 0.5

Ha🡪 p-value ≠ 0.5

**Syntax: t.test(data.1, data.2)** – The basic method of applying a t.test is to compare two vectors of numeric data.

There are two types of tests; one-tailed test and two-tailed test. In one-tailed the region of rejection is one side of sampling distribution and in two-tailed region of rejection is both sides of sampling distribution.

* + 1. **Normality Test**

There are several methods for **normality test** such as **Kolmogorov-Smirnov (K-S) normality test** and **Shapiro-Wilk’s test**. **Shapiro-Wilk’s** method which is based on the correlation between the data and the corresponding normal scores, is most widely used for normality test as it provides better power than K-S normality. There are two packages required for shapiro test which are **dplyr** and **ggpubr**.

Syntax: **shapiro.test**() # find normality for one variable.

In outcome the p-value > 0.05 means that the distribution of the data is not significantly different from normal distribution. we can assume normality (STHDA, n.d.).

* + 1. **Correlation Test**

The correlation test is used to find out linear dependency between two variables. It depends on the distribution of the data.

There are three methods to perform **correlation test**:

* **Pearson correlation, it** measures a linear dependence between two variables. It is most commonly used method in data analysis. It depends to the distribution of the data. It can be used when data is normal distribution.
* **Kendall** and **Spearman** which are rank-based and non-parametric correlation coefficients.

**Correlation Formula**

**Where, x** and **y** are two vectors of length

**mx** and **my** corresponds to the means of x and y, respectively.

**Spearman Correlation Formula**

The **Spearman Correlation** method computes the correlation between the rank of x and the rank of y variables. To find correlation selected method is Spearman because this is the only one method in which there is no assumption (Statistics Solutions, 2019).

Where, x′=rank(x) and y′=rank(y)

* + 1. **Feature Selection (Wrapper Method)**

In wrapper methods, train model using subset of features. These methods are usually computationally very expensive while using Realtime data ([Kaushik](https://www.analyticsvidhya.com/blog/author/sauravkaushik8/) , 2016).

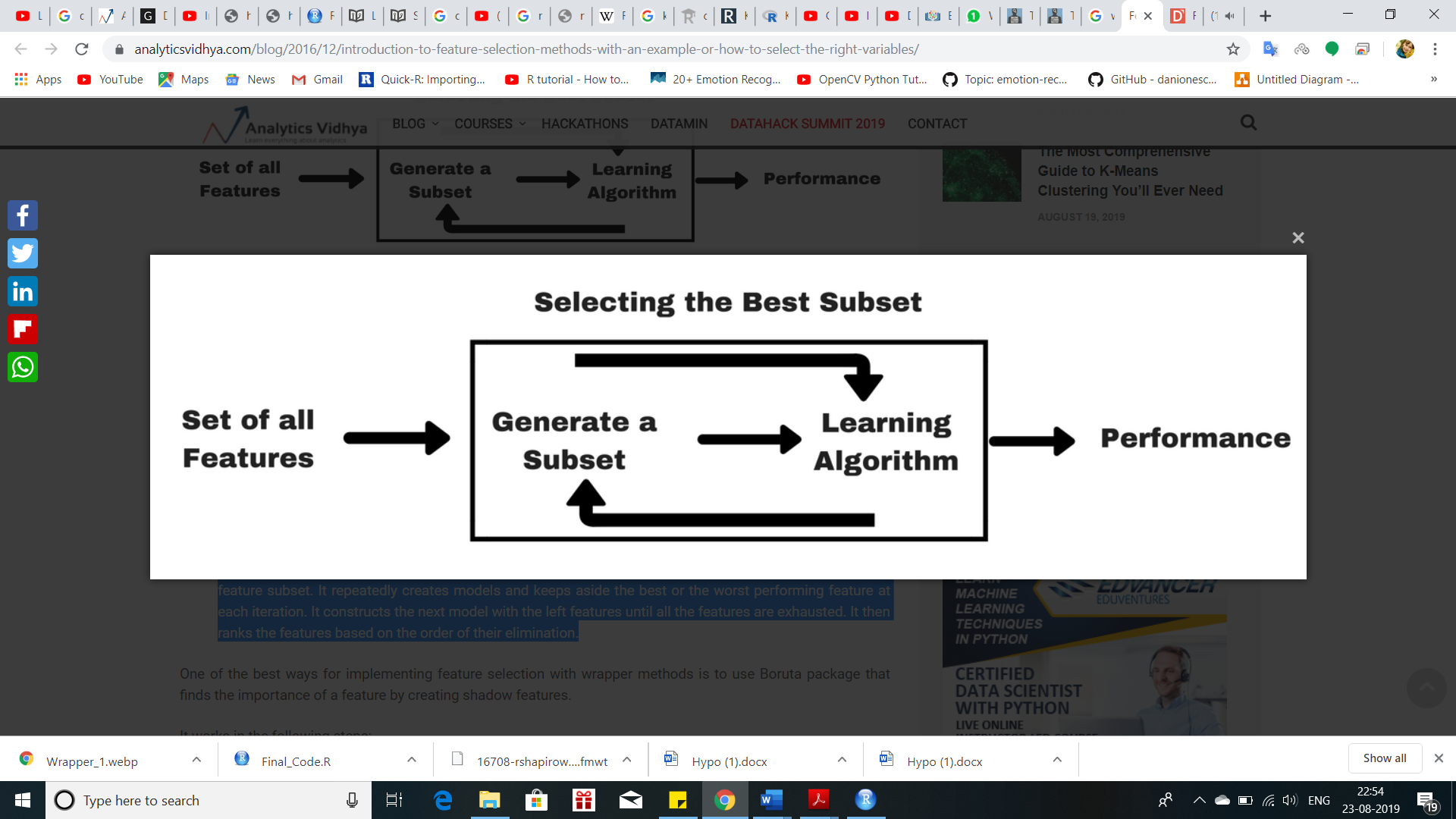


Fig 2.2.4: Wrapper Method working flow

Source: [*https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables*](https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/)

* + 1. **Multiple Linear Regression**

A linear regression model that contains more than one predictor variable is called a *multiple linear regression model*. The following model is a multiple linear regression model with two predictor variables, x1 and x2.

Y = β0 + β1 x1 + β2 x2 + ϵ

* + 1. **Dummy Variable**

Dummy variable method used when data has column which contains of categorical variable value. Dummy variable method is used because some algorithm does not handle categorical data (James, 2013).

**Syntax:** dummy\_cols() / dummy\_columns()

For example, there are two possible diagnoses: female and male. We could consider encoding these values as a quantitative, response variable, Y is

Y=

* + 1. **Interaction Method (Plot\_model)**

Interaction methods is use when the relationship between more than three variables. Plot\_model use to create plots from regression model, and it is a generic plot-function. In order to use plot\_model pakages such as sjplot, ggplot2, sjmic are loaded and installed.

**Syntax**: plot\_model(model, type)

Where, model is a regression model object, type is type of plot

Generally, there are three types of group for plot: Coefficients, Marginal Effects,Model diagnostics (Lüdecke, 2019).

* + 1. **RandomForest**

The randomForest performs prediction using the average of the prediction of multiple individual base models. For random forest method required libraries are ggplot2 and randomForest.

The randomForest package optionally produces two additional pieces of information: a measure of the importance of the predictor variables, and a measure of the internal structure of the data.

Syntax: randomForest (type ~ ., data, importance = TRUE, do.trace = 100)

The randomForest function returns an object of class "randomForest" (Canty, 2002) ([Sagar](https://dataaspirant.com/author/chaitanya-sagar/), 2018).

1. **Result**

From hypothesis testing it has been concluded that age, year and education has some relationship between wages and age/year/education. Using correlation matrix, the strength and type of relationship is carried out where all the variable has week relationship with wages. All possible regression feature selection algorithm is applied for selecting variables. If all features are considered, the outcome of model data is fitting good on regression line. Considering all the variables, multiple linear regression model is trained, and prediction of wages is done using test dataset which lead to the 39.1 Root mean square error. To reduce error and improve prediction, interaction between all the variable was found, but we observed interaction between age and education only. Using this interaction in linear regression model, wages are predicted which leads to 38.23 Root mean square error which means model is improved minor. Furthermore, normalize data of age is taken to train model and predict wages but Root mean square error is hiked to 39.37. Moreover, to improve the accuracy the model is trained using random forest algorithm in which log of age and interaction between age and education are taken as input parameters. By using this model 35.37 root mean square error is observed. At the end a model finalized and developed, which take two input values that are age and education to predict wages. It can conclude that around 65% of time this model predicts correct wage by considering root mean square error.

1. **Implementation**

The implementation of this task is realized using R. The practical task provided helps in investigating the availability and variability of resources which factors (age, year, education) affect the wage of the Atlantic male. This analysis helps us to understand the relationship, important variables and interaction between predictor and other variables, which helps to prediction wages accurately. The implementation explained below in detail.

* 1. **Relationship Using Hypothesis Testing**

The hypothesis testing of correlation is used to find relationship between depended and independent variables. There are two types of hypothesis, correlation value, H0 is called as null Hypothesis while correlation value Ha is not called as Alternate Hypothesis.

Before moving forward to relationship data understanding is important so it is found that there is one categorical variable which can’t be handle in hypothesis testing.

Solution for this problem is dummy variable. The strong reason for taking dummy variable is inequality problem; suppose we consider categorical variable as a factor then sometime in training dataset few category values were missing because training and testing data set are selected randomly. In many cases this issue of inequality leads to an error while doing prediction using test set.

Output:

Shapiro-wilk normality test

W = 0.87957, p-value < 2.2e-16

Spearman’s rank correlation rho

S = 4161669391, p-value = 3.754e-05

Sample estimates: rho - 0.07518448

Selected level of significance is .05 and Confidence Interval is 95%. This is two tail tests so, it is divided in level of significance by 2 which will give us .025 on both the tail.

There are three methods for correlation which are Pearson, Kendall and Spearman. Most widely used method is Pearson correlation, but there is an assumption that both variables should be normally distributed. Shapiro test is implemented to check weather wages are normally distributed or not. Outcome of this test is, p-value is < 2.2e-16 that means wages data is not normally distributed because p-value is < 0.05. From Shapiro test it is decided that Pearson method is not suitable method to find correlation. Spearman method is selected to find out correlation because this is the only method which have not assumption. After implementation of Spearman Correlation method, the outcome is p-value of all the test is not zero so Alternate Hypothesis is accepted for all tests, and Null Hypothesis is rejected. Thus, there is some relationship between wages and all other variables.

* 1. **Type and Strength of Relationship**

Correlation value is used to find the types and strength of relationship. If correlation value is positive then there is positive relationship whereas, the value is negative then there is a negative relationship. Correlation value is (+ and -) 0.8 or higher it indicates strong correlation, (+ and −) 0.5 to 0.7 is indicates medium correlation and (+ and −) 0.4 or low it represents weak correlation.

Output:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | year | age | wage | College\_Grad | Advanced\_Degree | HS\_Grad | Some\_  college | Less\_HS\_grad |
| Year | 1.00 | 0.04 | 0.08 | 0.00 | 0.02 | 0.01 | -0.02 | -0.01 |
| Age | 0.04 | 1.00 | 0.23 | 0.02 | 0.09 | -0.01 | -0.07 | -0.02 |
| Wage | 0.08 | 0.23 | 1.00 | 0.22 | 0.35 | -0.29 | -0.01 | -0.25 |
| College\_Grad | 0.00 | 0.02 | 0.22 | 1.00 | -0.22 | -0.38 | -0.29 | -0.17 |
| Advanced\_Degree | 0.02 | 0.09 | 0.35 | -0.22 | 1.00 | -0.28 | -0.21 | -0.13 |
| HS\_Grad | 0.01 | -0.01 | -0.29 | -0.38 | -0.28 | 1.00 | -0.36 | -0.22 |
| Some\_College | -0.02 | -0.07 | -0.01 | -0.29 | -0.21 | -0.36 | 1.00 | -0.16 |
| Less\_HS\_Grad | -0.01 | -0.02 | -0.25 | -0.17 | -0.13 | -0.22 | -0.16 | 1.00 |

Correlation value of Age and Wage is 0.23 which means **Positive Week Correlation**. There are few cases observed the age decreases and wage increases and it can happen in 20% of time. Correlation value of year and wage is 0.08 it implies **Positive Week Correlation**. There are few cases shown the year decreases and wage increases it can happen in 10% of time.

Correlation value of Hs\_Grad and Wages is -0.29 which indicates **Negative Weak Correlation**. There are few cases displayed when the number of Hs\_Grad student increases and wages decreases; it can happen in 29% of time. While, Some\_College and Wages is -0.01 (Negative Weak Correlation). There are few cases observed when the number of Some\_College student increases wages decreases; it can happen less than 1% of time. Correlation value of College\_Grad and wage is 0.22 which is Positive Weak Correlation. There are few cases displayed when the number of College\_Grad student decreases wages increases; it can happen in 22% of time. Less\_HS\_Grad and Wages is -0.25 means Negative Weak Correlation. There are few cases that when the number of less\_HS\_Grad student increases wages decreases; it can say in 25% of time it happens. Advanced\_Degree and Wages correlation value is 0.35 which is Positive Weak Correlation. There are few cases shown that when the number of Advanced\_Degree student decreases wages increases, it can happen 35% of time.

* 1. **Level of accuracy**

Prediction of wages is done by using multiple linear regression where input parameter is age, education and year.

There are few steps which performed to find accuracy. Make Dummy variable of education. Remove one of the category values of education to solve the problem of dummy variable trap and divided the data into to part 80% in training dataset and other 20% in testing. Using multiple linear regression algorithm, the model is trained, and training dataset is used in it. On the bases of train model, this result have predicted wages for testing dataset. In the final set root mean square error is calculated which is 39.1.

Output:

RMSE - 39.1156813

RSquared – 0.2138564

MAE – 27.4876104

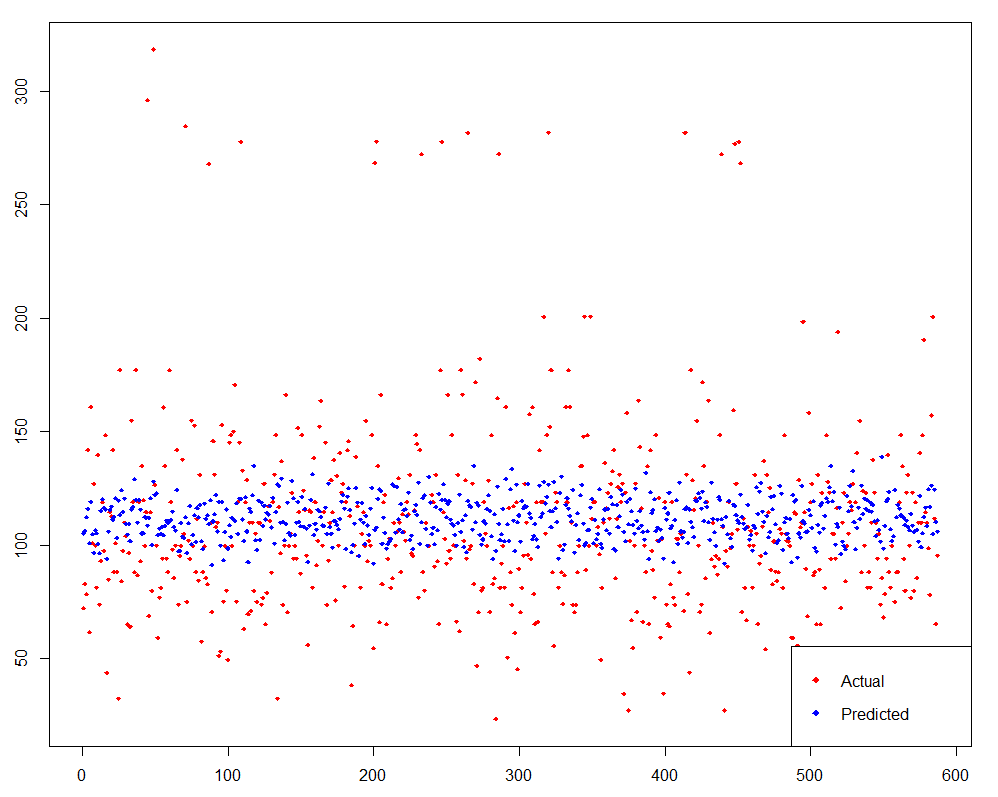


Fig 4.3: Output of Prediction Model

Fig 4.3 describes that red points are the actual testing values and blue points are the predicted values by the model. Root Mean Square error is 39.1 which is quite high so we predicted that this model doesn’t have good accuracy which could derive that age, education and year has week relationship with wages.

* 1. **Which Factors Contribute to Wage?**

In this scenario feature selection method is used to find which factor contribute to wage. There are three methods of feature selection: filter method, wrapper method and embedded method. Among them, wrapper method is selected and implemented all possible approach. Model trained by considering all variables. Education is taken as a factory for training model because this model is used for feature selection and cannot select one or two value of that category. Finally, it was needed to select whole category for prediction.

Output:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Index | N | Predictors | R-Square | Adj. R-Square | Mallo’s Cp |
| 1 | 1 | education | 0.235 | 0.234 | 107. |
| 2 | 1 | age | 0.0383 | 0.0380 | 904. |
| 3 | 1 | year | 0.00430 | 0.00396 | 1042. |
| 4 | 2 | age education | 0.259 | 0.258 | 9.64 |
| 5 | 2 | year education | 0.238 | 0.237 | 95.8 |
| 6 | 2 | year age | 0.0416 | 0.0410 | 892. |
| 7 | 3 | year age education | 0.262 | 0.260 | 1 |

In outcome, R-Square value is calculated for every possible combination of predictor variable. we can say that the model trained with all variables is the best on the bases of R-Square value which is 0.262 which is highest. That means, by using that model or on that scenario this model fits with data well as compare to other models.

* 1. **Is the Relationship Linear?**

By using pairs function, it can clearly observe that when any variable plotted with wages it doesn’t fit with all or high amount of data in approximately straight line not even around 70% of data.

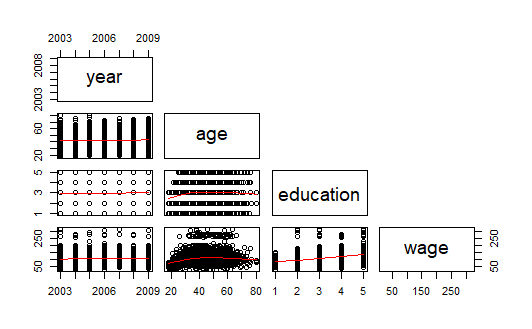


Fig. 4.5.1: Scatter Plots Matrix

Moreover, when I find strong or weak relationship between wage and other variables as already discussed in Section 4.3, multiple linear regression was performed whose result is 39.1 root mean square error and R-Square value is 0.21. By considering these two points we can conclude that relationship is not linear. Linear model will not fit the data well. In this case we had observed 0.21 R-Square value which means only 21% of data is fitting on a straight-line. When the relationship is not linear between predictor and other variables, at that time, first we try to us interaction variable in model development or we can also transform data from nonlinear to linear form by using functions link log.

Output:

Use interaction

RMSE - 38.2256813

RSquared – 0.2194372

MAE – 26.7576320

Log value

RMSE - 39.3826816

RSquared – 0.2142381

MAE – 27.4893628

RandomForest with log age and interaction of age and education

RMSE - 35.3826583

RSquared – 0.2688404

MAE – 24.6324849

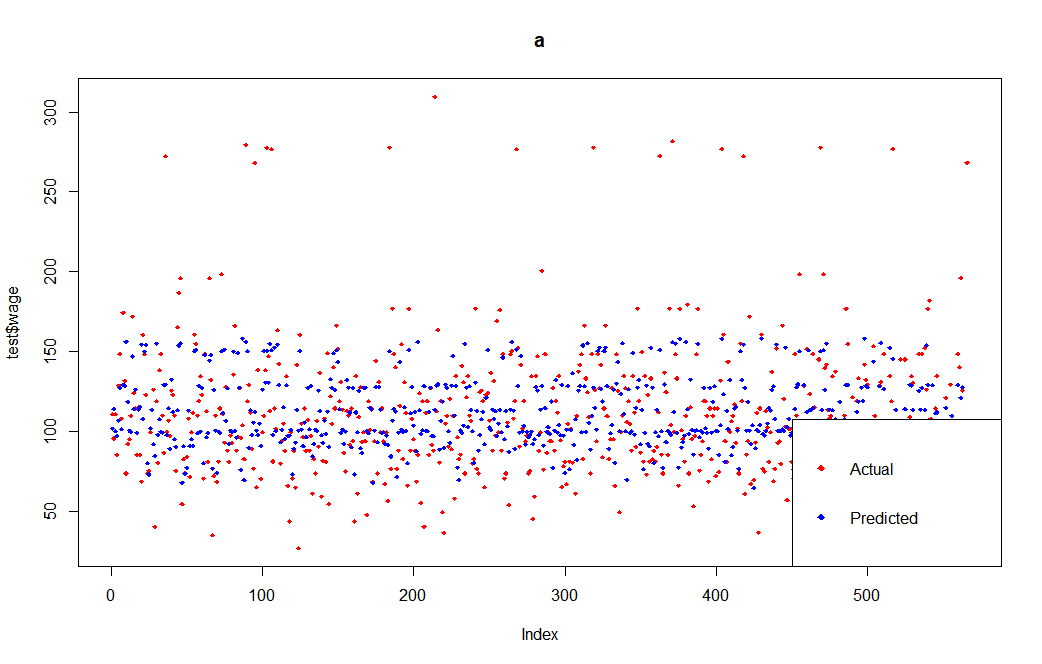


Fig. 4.5.2: Prediction using Random Forest (Finalized Model)

If transformation technique doesn’t work then we have to apply other algorithms like Random Forest, Support Vector Machine and Neural Network that can handle nonlinear data.

* 1. **Are there Interaction Effects?**

To find interaction effect we had train model with various interaction combination and check whether there is a slop difference or not. If there is variation of slop then it can conclude that there is an interaction between that variable and if slop of both variables is equal then there is no interaction between variable.

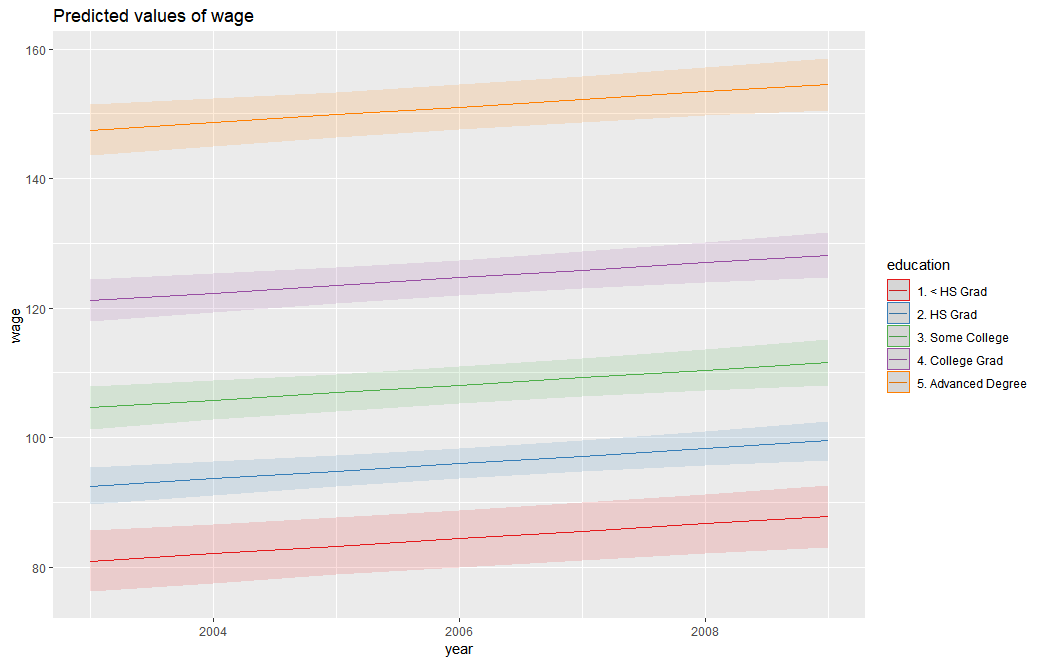


Fig 4.6.1: Interaction between year and education

There is no variation in slop observed so, it can say that there is no interaction between variable.

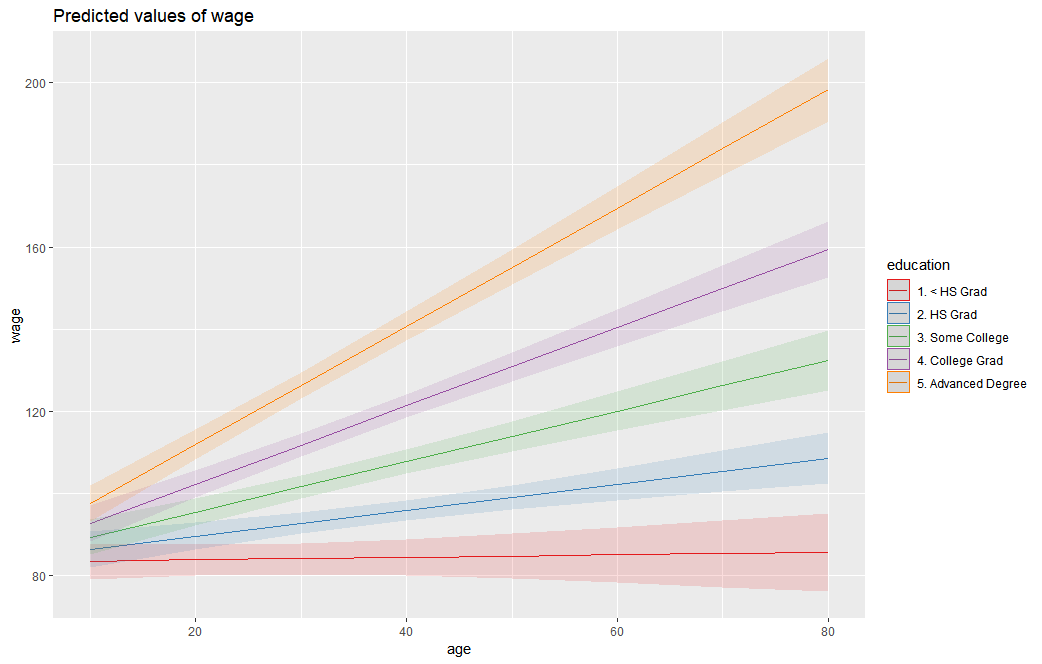


Fig 4.6.2: Interaction between age and education

There is variation in slop observed so, it can say that there is interaction between variable.

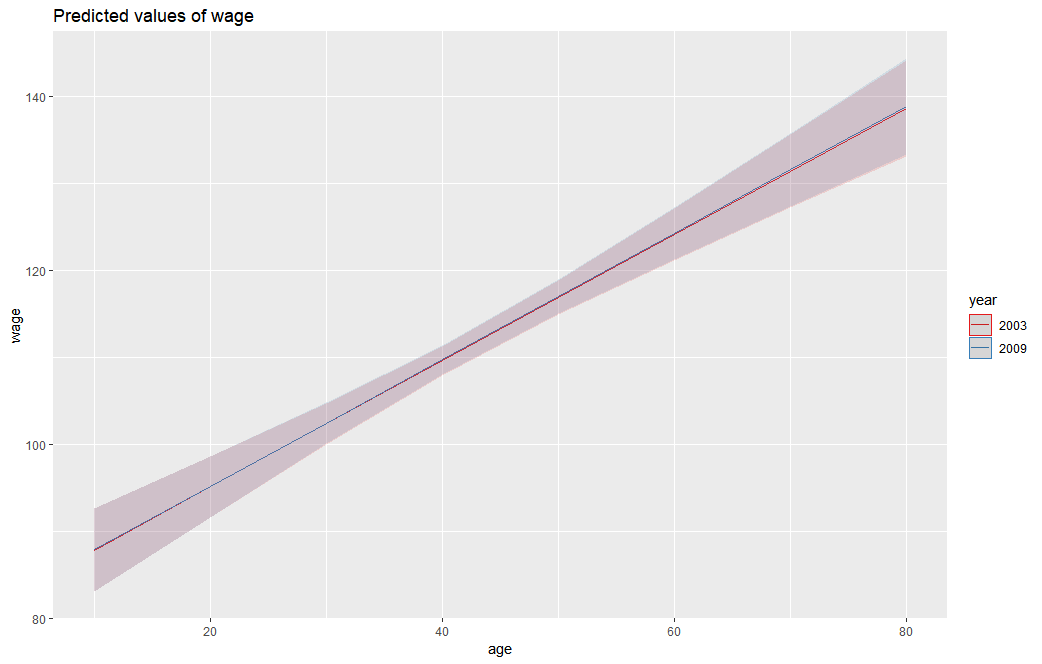


Fig 4.6.3: Interaction between age and year

There is no variation in slop observed so, it can say that there is no interaction between variable.

1. **References**

<https://www.census.gov/programs-surveys/cps/technical-documentation/methodology.html>

Bock, T. (August 29, 2018). What are Dummy Variables? [Blog post]. Retrieved from https://www.displayr.com/what-are-dummy-variables/#targetText=Dummy variables are the main is the number of employees.

Burdess, N. (2010). Introducing hypothesis testing. Starting statistics: A short, clear guide (pp. 127-137). London: SAGE Publications Ltd. doi: 10.4135/9781446287873

Canty, A. (December, 2002). Resampling Methods in R: The boot Package. The Newsletter of the R Project, 2(3), 18-19. Retrieved from <https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf>

DataFlair Team. (July 13, 2019). Introduction to Hypothesis Testing in R – Learn every concept from Scratch!. [Blog post]. Retrieved from <https://data-flair.training/blogs/hypothesis-testing-in-r/>

G. James et al., An Introduction to Statistical Learning: with Applications in R,Springer Texts in Statistics, DOI 10.1007/978-1-4614-7138-7 1, © Springer Science+Business Media New York 2013.

[Kaushik](https://www.analyticsvidhya.com/blog/author/sauravkaushik8/), S. (December 1, 2016). Introduction to Feature Selection methods with an example (or how to select the right variables?). Retrieved from <https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/>

Lüdecke, D. (August 02, 2019). Plotting Estimates (Fixed Effects) of Regression Models. Retrieved from <https://cran.r-project.org/web/packages/sjPlot/vignettes/plot_model_estimates.html>.

RStudio Team (2011). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. URL <https://blog.rstudio.com/2011/02/28/rstudio-new-open-source-ide-for-r/>

RStudio Team (2019). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. URL <https://blog.rstudio.com/2019/04/30/rstudio-1-2-release/>

[Sagar](https://dataaspirant.com/author/chaitanya-sagar/), C. (January 15, 2018). Feature Selection Techniques with R. [Blog post]. Retrieved from https://dataaspirant.com/2018/01/15/feature-selection-techniques-r/

Statistics Solutions .(2019). Correlation (Pearson, Kendall, Spearman). Retrieved from <https://www.statisticssolutions.com/correlation-pearson-kendall-spearman/>

STHDA. (n.d.). Normality Test in R. [Blog post]. Retrieved from http://www.sthda.com /english/ wiki/ normality-test-in-r.

**Appendix**

#------------------------------------------------------------

# Name: Bijalben Prajapati

# Matriculation number: 26521

# Course name: Data Analysis and Statistics

# Date: 31/08/2019

#------------------------------------------------------------

rm(list=ls())

install.packages('ISLR')

while(!is.null(dev.list()))

{

dev.off()

}

require(ISLR)

library(ISLR)

attach(Wage)

library(caret)

library(ISLR)

library(ggplot2)

library(dplyr)

library(ggpubr)

library(olsrr)

library(randomForest)

library(sjPlot)

library(Boruta)

assessment\_dataframe <- Wage[sample(nrow(Wage), 3000), ]

colnames(assessment\_dataframe)

Sel\_data <- assessment\_dataframe[,c(1,2,5,11)]

colnames(Sel\_data)

# understand data description

summary(Sel\_data)

# check if data contaion any null value or not

is.null(Sel\_data$age)

is.null(Sel\_data$wage)

is.null(Sel\_data$education)

is.null(Sel\_data$year)

# handly categorical varaible

Dummy\_results <- fastDummies::dummy\_cols(Sel\_data)

colnames(Dummy\_results)

Sel\_Data\_Dummy<-Dummy\_results[,c(1,2,4:9)]

colnames(Sel\_Data\_Dummy)

# give proper name

names(Sel\_Data\_Dummy)[names(Sel\_Data\_Dummy) == "education\_5. Advanced Degree"] <- "Advanced\_Degree"

names(Sel\_Data\_Dummy)[names(Sel\_Data\_Dummy) == "education\_4. College Grad"] <- "College\_Grad"

names(Sel\_Data\_Dummy)[names(Sel\_Data\_Dummy) == "education\_1. < HS Grad"] <- "less\_HS\_Grad"

names(Sel\_Data\_Dummy)[names(Sel\_Data\_Dummy) == "education\_3. Some College"] <- "Some\_College"

names(Sel\_Data\_Dummy)[names(Sel\_Data\_Dummy) == "education\_2. HS Grad"] <- "HS\_Grad"

colnames(Sel\_Data\_Dummy)

summary(Sel\_Data\_Dummy)

head(Sel\_Data\_Dummy)

# check normality of wage

shapiro.test(Sel\_Data\_Dummy$wage)

# check how much data is fit to linear line

#pairs (Sel\_data, upper.panel = NULL)

pairs (Sel\_data, upper.panel = NULL, panel = panel.smooth)

# Hypothesis Test for finding relationship

cor.test(Sel\_Data\_Dummy$year, Sel\_Data\_Dummy$wage, method = "spearman", exact = F, conf.level = 0.95)

cor.test(Sel\_Data\_Dummy$age, Sel\_Data\_Dummy$wage, method = "spearman" , exact = F, conf.level = 0.95)

cor.test(Sel\_Data\_Dummy$College\_Grad, Sel\_Data\_Dummy$wage, method = "spearman" , exact = F, conf.level = 0.95)

cor.test(Sel\_Data\_Dummy$Some\_College, Sel\_Data\_Dummy$wage, method = "spearman" , exact = F, conf.level = 0.95)

cor.test(Sel\_Data\_Dummy$less\_HS\_Grad, Sel\_Data\_Dummy$wage, method = "spearman" , exact = F, conf.level = 0.95)

cor.test(Sel\_Data\_Dummy$HS\_Grad, Sel\_Data\_Dummy$wage, method = "spearman" , exact = F, conf.level = 0.95)

cor.test(Sel\_Data\_Dummy$Advanced\_Degree, Sel\_Data\_Dummy$wage, method = "spearman" , exact = F, conf.level = 0.95)

# Find Stringth and Type of relationship

Correlation <- cor(Sel\_Data\_Dummy,method = "spearman")

round(Correlation, 2)

# Feature Selection for traning a model

Feature\_Selection\_model <- lm(wage ~ ., data = Sel\_data)

ols\_step\_all\_possible(Feature\_Selection\_model)

# handle dummy variable trap problem

Sel\_Data\_Model\_Dummy <- Sel\_Data\_Dummy[,!(names(Sel\_Data\_Dummy) %in% c("Some\_College"))]

colnames(Sel\_Data\_Model\_Dummy)

# train model and predict wage with multiple regression

set.seed(222)

ind <- sample(2, nrow(Sel\_Data\_Model\_Dummy), replace = T, prob = c(0.8, 0.2))

train <- Sel\_Data\_Model\_Dummy[ind==1,]

test <- Sel\_Data\_Model\_Dummy[ind==2,]

Wage\_Relationship\_by\_error\_model <- lm(wage ~ ., data = train)

WRE\_Prediction\_value <- predict(Wage\_Relationship\_by\_error\_model, test)

postResample(pred = WRE\_Prediction\_value, obs = test$wage)

# predicted result Vs. actual result

plot(test$wage,col='red',main='a',pch=18,cex=0.7)

points(WRE\_Prediction\_value,col='blue',pch=18,cex=0.7)

legend('bottomright',legend=c('Actual','Predicted'),pch=18,col=c('red','blue'))

# find interaction between variables

interaction\_YE <- lm(wage ~ year:education, data = Sel\_data)

plot\_model(interaction\_YE, type = "int")

interaction\_AE <- lm(wage ~ age:education, data = Sel\_data)

plot\_model(interaction\_AE, type = "int")

interaction\_AY <- lm(wage ~ age:year, data = Sel\_data)

plot\_model(interaction\_AY, type = "int")

#train model and predict wage with multiple regression by use interection variable in this

Improve\_Interaction\_model <- lm(wage ~ year+age+College\_Grad+Advanced\_Degree+HS\_Grad+less\_HS\_Grad+

age:(College\_Grad+Advanced\_Degree+HS\_Grad+less\_HS\_Grad), data = train)

II\_Prediction\_value <- predict(Improve\_Interaction\_model, test)

postResample(pred = II\_Prediction\_value, obs = test$wage)

#train model and predict wage with multiple regression by normalizing age

Improve\_Log\_model <- lm(wage ~ year+log(age)+College\_Grad+Advanced\_Degree+HS\_Grad+less\_HS\_Grad+log(age):College\_Grad+Advanced\_Degree+HS\_Grad+less\_HS\_Grad, data = train)

IL\_Prediction\_value <- predict(Improve\_Log\_model, test)

postResample(pred = IL\_Prediction\_value, obs = test$wage)

rf <- randomForest(wage ~ log(age)+(age):College\_Grad+Advanced\_Degree+HS\_Grad+less\_HS\_Grad, data = train)

rf\_Prediction\_value <- predict(rf, test)

postResample(pred = rf\_Prediction\_value, obs = test$wage)