|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Matriculation number:** | 2 | 6 | 5 | 0 | 9 |



**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.) | Summer Semester  March - 2019 | Turankar, Rohit |

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 |
|  |  |  |  |  |  |

**Statement of Authorship**

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 Material from the published or unpublished work of others, which is referred to in the report, is credited to the author in the text.

Rohit Turankar

Matriculation Number: 26509

**Introduction**

**Research Questions**

The research objectives are as follows:

* What are the dependent and independent factors which are affecting the income/wage for a group of males from the Atlantic region of the United States?
* How the sample data is being distributed with respect to age, calendar year and education level.
* How strong is the relationships amongst these factors?
* Programming language, tool and method to be used for accurate results.

A detailed material is provided in the assignment that briefly describes everything about the income survey. A brief explanation about the Research methods is also mentioned.

**Motivation**

The sample data has been taken from the assignment given by Prof. Dr. Schwind, Dipl. – Biol. Ralf Darius.

Interest has been invoked in simulation of “March 2011 Supplement to Current Population Survey data” by Steve Miller, of open BI. The inspiration is taken from the Human Capital Theory and the field of labor economics. The Current Population Survey (CPS) is a well-recognized U.S. survey that offers data about many of the things that define us as people and as a society — our job, income, and education. The labor market situation of low-educated people is particularly critical in most advanced economies like United States, especially among youngsters. The analysis can help in deriving a relation between education attainment and given wage for a specific age group.

**Goals**

The main purpose of this report is to simulate the income survey data for males in the central Atlantic region of the USA and to find the effects of different parameters on the income/wage of the males.

**Context**

Below facts have been considered about the income survey data.

* The sample data is having 3000 records.
* The calendar year range lies between 2003-2009.
* The age of males lies between 18 to 80-year-old.
* There are 5 levels in which education field is categorized i.e. Below HS (high School degree), HS Degree, Some College, College Grad and Advanced Degree(highest).
* In total 11 fields are present but only 4 fields i.e. age, calendar year, wage, education has been taken so that relevant analysis can be done.

**Tools & Methods**

After analyzing the assignment and doing the research on such kind of simulations for income

survey, the tools and methods were selected to find the accurate results using the best programming language that can represent analysis and statistics in a convenient way.

The programming language that is used for this task is R.

The application which helps for visualization and writing code is R Studio.

The methods that are used in simulating the data are as follows:

* ANOVA – Analysis of Variance
* Pearson correlation coefficient
* Spearman's rank-order correlation
* Kendall's Tau-b
* Regression (Linear and Multiple)
* Pairwise t test
* Interaction effect

ANOVA – Analysis of Variance

An ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help to figure out whether one need to reject the null hypothesis or accept the alternate hypothesis. One-way or two-way refers to the number of independent variables (IVs) in our Analysis of Variance test. One-way has one independent variable (with 2 levels) and two-way has two independent variables (can have multiple levels).

aov(formula, data = NULL, projections = FALSE, qr = TRUE,

contrasts = NULL, ...)

Formula here means a formula specifying the model

Pearson correlation coefficient

The Pearson correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by r. It can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. The stronger the association of the two variables, the closer the Pearson correlation coefficient, r, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. The two variables have to be measured on either an interval or ratio scale.

Spearman's rank-order correlation

The Spearman rank-order correlation coefficient is a non-parametric measure of the association intensity and direction between two factors measured at least on an ordinal scale. The symbol rs (or the Greek letter ρ, pronounced rho) denotes it. The test is used either for ordinal variables or for ongoing information that failed the Pearson's product-moment correlation.

It is only appropriate to use a Spearman’s correlation if your data "passes" two assumptions that are required for Spearman’s correlation to give you a valid result.

#1 Two variables should be measured on an ordinal, interval or ratio scale.

#2 There is a monotonic relationship between the two variables.

Kendall's Tau-b

Kendall's tau-b (τb) correlation coefficient (Kendall's tau-b, for short) is a nonparametric measure of the strength and direction of association that exists between two variables measured on at least an ordinal scale. It is considered a nonparametric alternative to the Pearson’s product-moment correlation when your data has failed one or more of the assumptions of this test.

It is also considered an alternative to the nonparametric Spearman rank-order correlation coefficient (especially when you have a small sample size with many tied ranks).

Regression

The simple linear regression is used to predict a quantitative outcome y on the basis of one single predictor variable x. The goal is to build a mathematical model (or formula) that defines y as a function of the x variable.

Once, we built a statistically significant model, it’s possible to use it for predicting future outcome based on new x values.

Multiple regression generally explains the relationship between multiple independent or predictor variables and one dependent or criterion variable. A dependent variable is modeled as a function of several independent variables with corresponding coefficients, along with the constant term.

Multiple regression requires two or more predictor variables, and this is why it is called multiple regression.

Pairwise t test

Another approach for determining which pairwise groups are significantly different following ANOVA is to use multiple t tests followed by one of the following tests to deal with familywise error: Bonferroni, Dunn-Sidàk, Holm’s, Hochberg, Benjamini-Hochberg or Benjamini-Yekutieli.

In our case we are using Bonferroni method post ANOVA test to do the analysis.

Interaction effect

In regression, an interaction effect exists when the effect of an independent variable on a dependent variable change, depending on the value(s) of one or more other independent variables.

In a regression equation, an interaction effect is represented as the product of two or more independent variables. Typical regression equation without an interaction:

ŷ = b0 + b1X1 + b2X2

**Results**

Following are the tasks which has been performed to fulfill the requirements of demonstrating the assignment appropriately.

* Demonstration of distribution of data using histograms.
* Presentation of association between variables using Scatter plots.
* Creation of tables for doing descriptive statistics of ordinal categorical variable with ANOVA and T test.
* Presenting final analysis using ggplots.

**Evidence of association between Age/Education/Year and Wage**

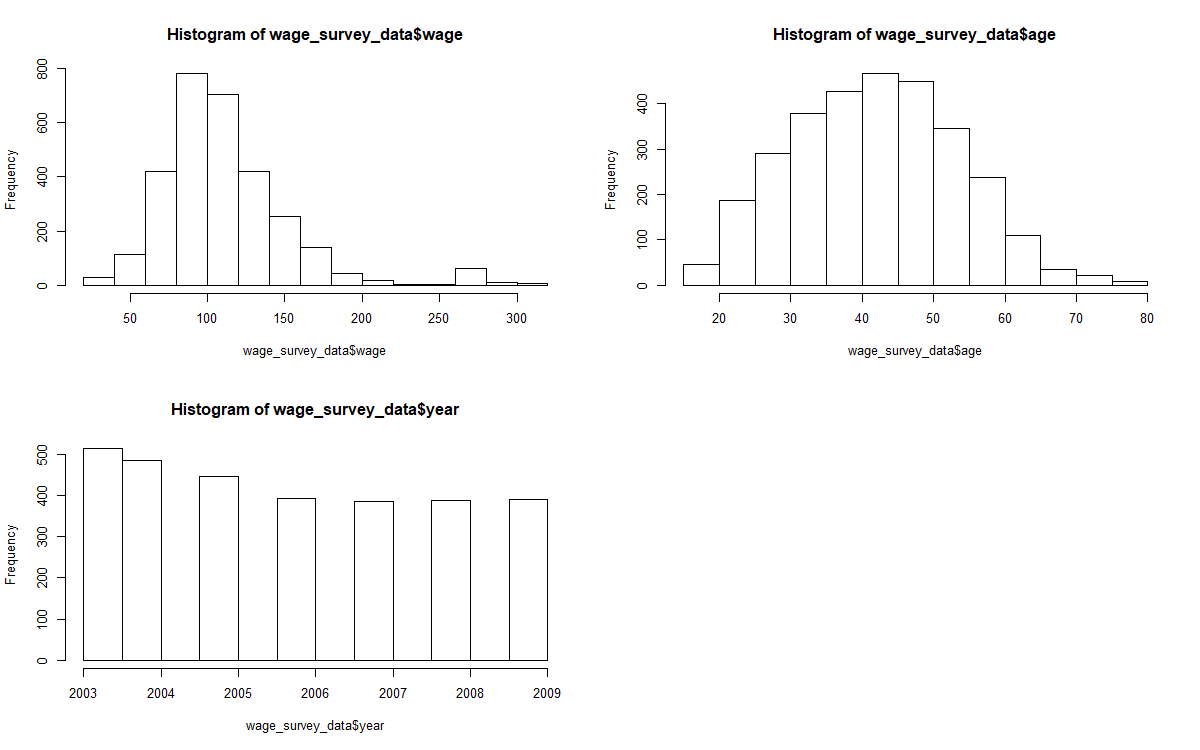


Figure 1: Visual representation of the Numeric Variables in the data set.

From Figure 1 it can be observed that age appears to be primarily concentrated about 30 – 50-year-old men with the counts tapering off beyond this range. It is peculiar that there is a worker who is almost 80 years of age in this data set. As for the wage, for the most part, it seems to be distributed stronger to 100 (thousand) with high wages about 250 (thousand) and above in a smaller set.

In the year graph, a higher number of workers was noted in the earlier years between 2003 and 2005, nonetheless, there is a steady decrease to this number until the year 2006, the number then plateaus around 390 workers detected for the other years.

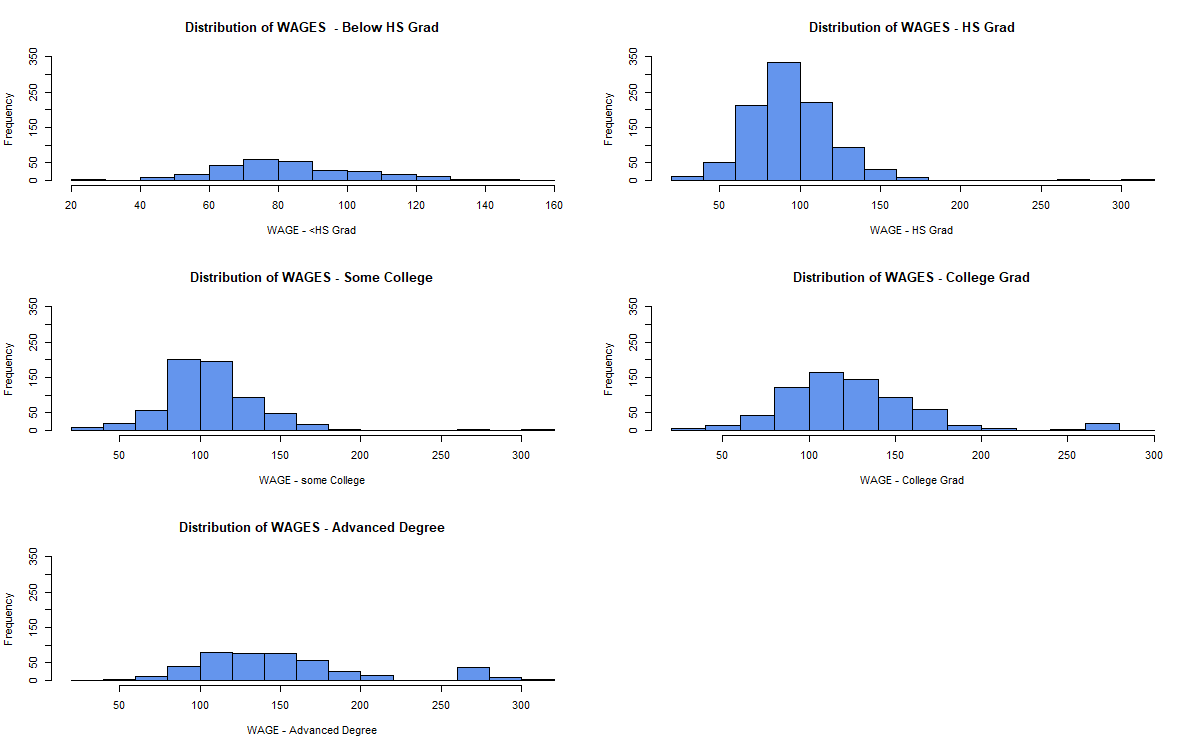


Figure 2: Visual representation of the distribution of wages among all education levels in the data set.

From Figure 3 it can be observed that how wages have been distributed among the different education level within the data set. Here y-axis shows the frequency i.e. number of males. There were high number of males approximately 350 who were getting wage between $90-110 and has a qualification of HS grad. A few were only getting wages more than $150.

Similar is the case for males whose education was “Some College” but the frequency is comparatively low. Although the count is less for the males which has Advanced Degree level, the number is significant who were getting high wage i.e. above $200. Going forward, from < HS Grad to Advanced Degree, the count of males increases who get more than $200 wage.

These graphs provide an evidence of association between age/education/calendar year and wage which can be explained in detail in the further report.

**Strength of relationships between dependent (wage) and independent variables(age/education/year)**

**Wage ~ Year**

Firstly, as both the variable i.e. dependent variable 🡪 Wage and independent variable 🡪 Year are numeric, there are multiple tests which can be performed to find the correlation between them.

After performing Kendal’s Tau test as the year data seems to be ordinal numeric where it starts from 2003-2008, Kendal’s tau coefficient **rk\_wage\_year** value is **0.05432**.

We have tried to plot graph using below linear regression model formula.

**yearmodel <- lm(wage ~ year, data = wage\_survey\_data)**

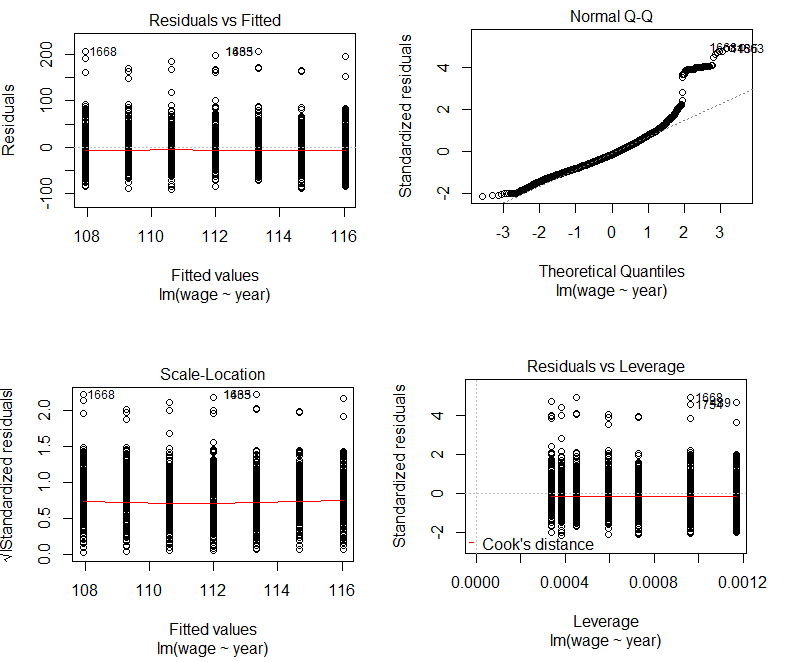


Figure 3: Graphic view of the Residual vs Fitted for finding Linear regression between wage and year.

After observing the value of correlation coefficient and Figure 3 graphs where we can see a straight line having almost **slope 🡪 0. The relation between wage and year is weak.**

**Wage ~ Age**

In this scenario, the independent and dependent variable is numeric and demonstrated in range i.e. wage is having a range from $0 to $350 and age of males from 18 to 80 years. The correlation test which gives best results for this type of data is Pearson’s Correlation Coefficient test.

After performing Pearson’s correlation coefficient, the value of **rp\_wage\_age** is **0.19563.**

We have tried to plot graph using below linear regression model formula.

**agemodel <- lm(wage ~ age, data = wage\_survey\_data)**

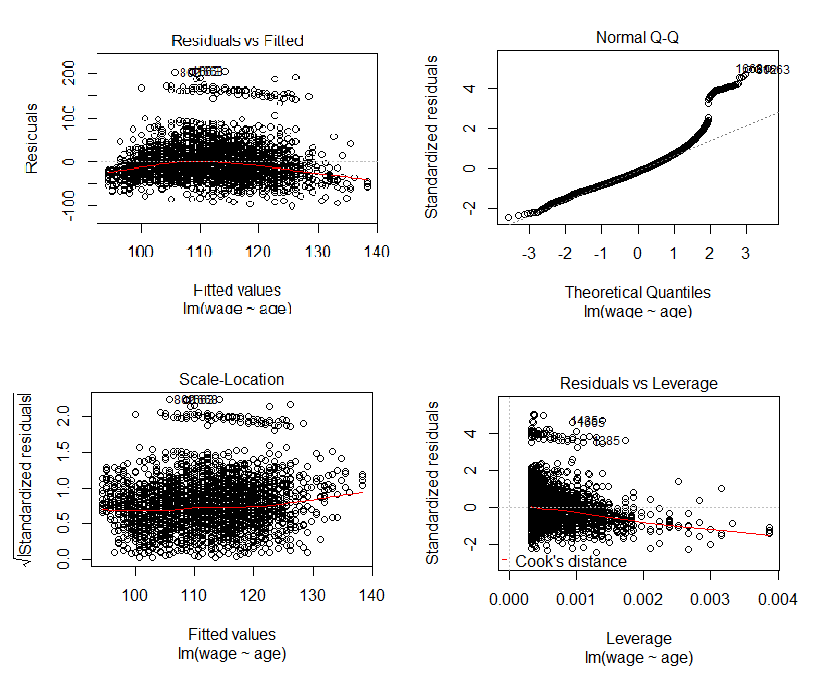


Figure 4: Graphic view of the Residual vs Fitted for finding Linear regression between wage and age.

After analyzing the result from correlation test (**rp\_wage\_age** = **0.19563)** and graphs the relation is moderately strong. We can observe the slope more clearly in Figure 5.

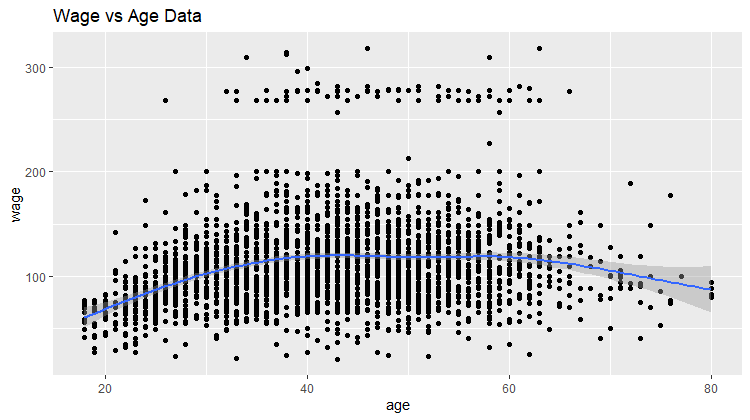


Figure 5: Scatter plots between Wage and Age with regression line.

The slope of the regression line is positive till a certain age i.e. 40. It means as the age increase from 18 to 40, wage of the males also increases. Once it reaches 40 it becomes stable till 60 which is the retirement age for most of the working employees. There is no change in the wage for this range. After 60 we can observe a negative slope which is illustrates that the increase in age after 60 years will impose a negative impact on wages of the males.

From the above graphs and result from the Pearson’s correlation test, we can strongly comment that the **relation between wage and age is moderate**.

**Wage ~ Education**

In this analysis we examined the relationship between education level and wage. An education beyond high school can be extremely expensive and can take graduates years to repay once they enter the job market. An assurance of a higher wages after graduation can be a motivating factor in seeking to attain a higher education level before entering the job market.

Education is an ordinal categorical variable with the levels < HS Grad, HS Grad, Some College, College Grad, Advanced Degree, and represent the highest level of education attained by the males. Wage is a continuous numeric variable and represents the income of the males.

A subset of the wage survey data set was taken consisting of the variables Education and Wage.

A picture containing scoreboard

Description automatically generated

Table 1: Statistics of wages with respect to education level

In the table 1 above we can see descriptive statistics for wage by education. We can see that average wage increases as the education increases with the lowest average wage of $84.10 for

< HS Grad and the highest average income of $150.92 for Advanced Degree, standard deviation also increases with education level.

ANOVA is the best choice for analyzing our data because we have a categorical variable with five level, Education, and a numerical variable, wage.

With ANOVA our null hypothesis is that all levels of education have the same mean income and any observed difference between education levels is due to chance. While the alternate hypothesis is that at least one education level has a mean income that is different than the other mean wages.

Next we check the conditions for ANOVA i.e distribution of the observations in each education level are normally distributed.

A screenshot of a video game

Description automatically generated

Figure 6: Boxplot to check distribution of data across all education levels.

From the boxplots we can see the variability is approximately the same across all education groups.

In the summary in table 2 we can see the results of the ANOVA test. Our F value from the ANOVA is 229.8059 and our p-value is 2.915932e-172. Because of this we can reject the null hypothesis at the alpha = 0.05 level and conclude that at least one income level is different than the other income levels.

Formula for Anova Analysis

**fit <- aov(wage ~ education, data = subset)**

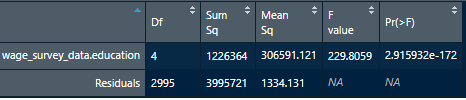


Table 2: Summary table

Next we can use pairwise t-tests and the Bonferroni correction to determine which education levels are significantly different from each other. Below is the code used for the tests and the code that generates of results for the test in table 3.

Formula for Pairwise T test

**ttests <- pairwise.t.test**

**(subset$wage\_survey\_data.wage,subset$wage\_survey\_data.education,p.adj="bonferroni")**

A close up of a scoreboard

Description automatically generated

Table 3: Pairwise ttests results

From the above graphs and result from the ANOVA test and pairwise t test, we can strongly comment that the **relation between wage and education is strong**.

**Prediction of wage based on Age and Education (2 Factors)**

For prediction of wage of males which is the dependent variable, we are taking on 2 independent variables i.e Age and Education. It is so because from the above analysis of all the 3 independent variables, we have observed that the relationship between wage and year is least significant. It eventually does not contribute or impact to wages taken by males in Central Region of United states.

We are using **Multiple Regression Model** for predicting wage along with further analysis on interaction effect between independent variables.

Formula for Regression - predictor

**model1 <- lm(wage ~ age + education, data = wage\_survey\_data)**

summary(model1)

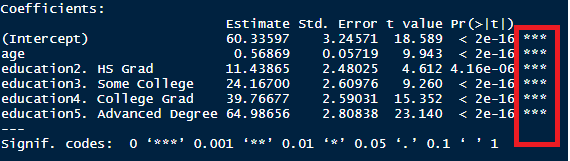


Table 4: Summary of multiple regression predictor model

From Table4: It shows how significant is the values for age and education levels. These values and coefficients are used to derive the prediction formula for wage taking age and education levels as variables.

**Rt = 60.33597 + 0.56869 \* Age + 11.43 E2 + 24.16 E3 + 39.77 E4 + 64.99 E5**

Rt 🡪 Variable for prediction of wage

E1🡪 Variable for Education level - <HS Grad

E2🡪 Variable for Education level - HS Grad

E3🡪 Variable for Education level - Some College

E4🡪 Variable for Education level – College Grad

E5🡪 Variable for Education level - Advanced Degree

To calculate wage for a male having specific education level and age. With the help of above prediction formula, wage can be calculated.

For example, a male having Age – 35 and Education level E3 i.e. Some College. We will calculate the wage.

Rt = 60.33597 + 0.56869 \* (35) + 11.43 (0) + 24.16 (1) + 39.77 (0) + 64.99 (0)

**Rt = 104.400**

A picture containing screenshot

Description automatically generatedFigure 7: Plotting multiple regression model with regression lines

From Figure 7 and Table 4 it was observed that the reference level for this multiple regression model is E1 i.e. < HS Grad and the respective coefficient are the distance between the reference line(red) to every regression line above it.

We have observed that there are many outliers which can not be ignored. Although the Figure 7 depicts a linear graph between wage, age and education levels, but outliers contribute to standard error which is significant.

Hence there are other factors also which affect this relationship.

**Interaction Effect on Multiple Regression Model**

Below is the model which is used to test the interaction effect.

**model2 <- lm(wage ~ age + education + age:education, data = wage\_survey\_data)**

**summary(model2)**

A black sign with white text

Description automatically generated

Table5: Summary of interaction model

This interaction effect can change the prediction formula which is used earlier for predicting wage. We observed that interaction effect between 4th and 5th education level is more significant with age. These 2 education levels are more impacted by age factor.

RTI = 68.09017 + 0.38316 \* age + 7.24752 E2 + 0.86720 E3 + 36.83562 E4 + 61.13781 E5 + 0.10113 age:E2 + 0.56573 age:E3 + 0.07277 age:E4 + 0.09876 age:E5

RTI is the prediction formula considering the interaction effect. The lines which are parallel in Figure 7 will not be non-parallel now.

The interaction effect is significant and can change the prediction values.

Using GGPLOT to present the final analysis

A picture containing screenshot

Description automatically generated

Figure 8: Boxplot + Scatterplot for demonstrating the final analysis

From figure 8: it was observed that the wages are least deviated for 1. <HS Grad and highly deviated for 5. Advanced Degree education level. There are high number of males which get wages more than $200 for 5. Advanced Degree education level. These points are outliers but are significant in number. There is high density of data for all age group males for education level 2. HS grad. Going forward from 1st to 5th level mean of the wages increases. Also, there are low count of males who are younger and have Advanced Degree and are college grad. The relation between age/education with wage is duly presented with the above figure

**Interpretation and Discussion**

It can be observed from the above conducted tests that for a given person age, it is feasible to predict its wage using Linear Regression models. Moreover, the regression lines are not straight and have negative slope after a certain interval. This shows that there are other factors also which impact wage i.e education level. The calendar year is the independent factor but has insignificant effect on wage. Hence the slope of such variable with wage tends to 0 which is correct with given data set.

Age and education are the independent variable which are significantly related to the wage. In order to make this model Multiple regression model used considering ordinal categorical data.

It was also observed from the ANOVA analysis test that how strongly level of education impacts the wage.

This is an indicator that to get an accurate prediction of a man’s wage, it is necessary to combine the age and his level of education.

**References**

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) An Introduction to Statistical Learning with applications in R, www.StatLearning.com, Springer-Verlag, New York

Andrew Beckerman, Owen Petchey (2012). Getting Started With R. Oxford Biology

Yosef Cohen and Jeremiah Y. Cohen (2008). Statistics and Data with R. Wiley

**Code**

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

## Reset R's brain

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

#rm(list=ls())

#

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

## Reset graphic device

## As long as there is any dev open (exept "null device")

## close the active one!

## Caution: closes all open plots!!!!

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

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

#{

# dev.off()

#}

#

## install.packages("ISLR") -- run this code to install ISLR package - first time only

##install.packages("ISLR")

#require(ISLR) #loading the package

#library(ISLR) # enabling package

#attach(Wage)

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

#

##checking the summary of dataset

#summary(assessment\_dataframe)

#

##checking the structure of dataset

#str(assessment\_dataframe)

#

##checking in case null values is present in the data frame

#assessment\_dataframe[!complete.cases(assessment\_dataframe),]

#

##checking in case NA is present in any columns

#assessment\_dataframe[is.na.data.frame(assessment\_dataframe),]

#

##Extract the data of the columns year, age, education and wage and work on those only.

#wage\_survey\_data <- assessment\_dataframe[, c("year", "age", "education","wage")]

#

#str(wage\_survey\_data)

#summary(wage\_survey\_data)

#

##checking the levels, whether levels are correct or not

#levels(wage\_survey\_data$education)

#

##changing the row names to NULL to sync the index

#row.names(wage\_survey\_data) <- NULL

#

## Histograms for showing the distribution of data(number of males) - Wage , Age and Year

###run all the 6 lines together for getting the desired histograms

##Below is the par function used to get all the graphs at once in the screen

#

#parBackup <- par()

#par(mfrow=c(1,3))

#wagehist <- hist(wage\_survey\_data$wage, breaks=15)

#agehist <- hist(wage\_survey\_data$age, breaks=15)

#yearhist <- hist(wage\_survey\_data$year)

#par(parBackup)

#

### Seperate histograms in one plot to check the number of males corresponding to each level of education

#

#parBackup <- par()

#par(mfrow=c(3,2))

#WageHISTHS <- hist(wage\_survey\_data[,4][wage\_survey\_data[,3]=="1. < HS Grad"],

# breaks=15,

# col="cornflowerblue",

# xlab="WAGE - <HS Grad",

# ylim = c(0, 350),

# main="Distribution of WAGES - Below HS Grad")

#WageHISTHsGrad <- hist(wage\_survey\_data[,4][wage\_survey\_data[,3]=="2. HS Grad"],

# breaks=15,

# col="cornflowerblue",

# xlab="WAGE - HS Grad",

# ylim = c(0, 350),

# main="Distribution of WAGES - HS Grad")

#WageHISTSomeCollege <- hist(wage\_survey\_data[,4][wage\_survey\_data[,3]=="3. Some College"],

# breaks=15,

# col="cornflowerblue",

# xlab="WAGE - some College",

# ylim = c(0, 350),

# main="Distribution of WAGES - Some College")

#WageHISTCollegeGrad <- hist(wage\_survey\_data[,4][wage\_survey\_data[,3]=="4. College Grad"],

# breaks=15,

# col="cornflowerblue",

# xlab="WAGE - College Grad",

# ylim = c(0, 350),

# main="Distribution of WAGES - College Grad")

#WageHISTAdvancedDegree <- hist(wage\_survey\_data[,4][wage\_survey\_data[,3]=="5. Advanced Degree"],

# breaks=15,

# col="cornflowerblue",

# xlab="WAGE - Advanced Degree",

# ylim = c(0, 350),

# main="Distribution of WAGES - Advanced Degree")

#par(parBackup)

#

#

#### installing ggplot package

#install.packages("ggplot2")

#

##enabling ggplot library

#library(ggplot2)

#

#

####Calculate the correlation coefficient for quantitatve (numeric) fields i.e wage and age,

####wage and year with Pearson, Kendall and Spearman test respectively

#

###wage and year (quantitative ---> quantitative)

####Kendals correlation test

###here rk\_wage\_year is the kendal's tau correlation coefficient

#

#rk\_wage\_year <- cor(wage\_survey\_data[,"wage"],

# wage\_survey\_data[,"year"],

# method = c("kendall") )

#

###rk\_wage\_year = 0.05432

#

#yearmodel <- lm(wage ~ year, data = wage\_survey\_data)

#plot(yearmodel)

#

### wage and age (quantitative --> quantitative)

####Pearson correlation test

#

#rp\_wage\_age <- cor(wage\_survey\_data[,"wage"],

# wage\_survey\_data[,"age"],

# method = c("pearson") )

#

#agemodel <- lm(wage ~ age, data = wage\_survey\_data)

#plot(agemodel)

#

### qplot for checking the slope between Wage and Age

#scatter\_01 <- qplot(age, wage, data=wage\_survey\_data, main="Wage vs Age Data")

#wageagescatter <- scatter\_01 + geom\_smooth()

#wageagescatter

#

#######Anova test for finding relation between ordinal categorical variable education with continuous numeric variable wage

#

#install.packages("xtable")

##used to create formated html tables

#library(xtable)

#

##used to subset data set

#library(plyr)

#

##used for dollar format

#library(scales)

#

###subsetting the data frame for performing ANOVA with 2 relevant variables

#

#subset <- data.frame(wage\_survey\_data$education, wage\_survey\_data$wage)

#subset <- na.omit(subset)

#

##subset of discriptive statistics

#summary <- ddply(subset, "wage\_survey\_data.education", summarise, N = length(wage\_survey\_data.wage),

# MEAN = mean(wage\_survey\_data.wage), STD = sd(wage\_survey\_data.wage),

# MIN = min(wage\_survey\_data.wage), MAX = max(wage\_survey\_data.wage),

# MEDIAN = median(wage\_survey\_data.wage), IQR = IQR(wage\_survey\_data.wage))

#

##renaming column of summay matrix

#names(summary)[1] <- "EDUCATION"

#

##converting values to dollar format

#summary <- cbind(summary[,1:2],

# apply(summary[c("MEAN","STD","MIN","MAX","MEDIAN","IQR")],2,dollar))

#

#

##creating and printing table of summary statistics

#stat.table <- xtable(summary, caption = "Table 1")

#print(stat.table, type = "html")

#

#

#g <- ggplot(subset, aes(x = wage\_survey\_data.education, y = wage\_survey\_data.wage, fill = wage\_survey\_data.education))

#g <- g + geom\_boxplot()

#g <- g + xlab("Education") + ylab("Wage") + labs(title = "Wage by Education")

#g <- g + guides(fill=guide\_legend(title=NULL))

#print(g)

#

##code for anova analysis

#fit <- aov(wage ~ education, data = subset)

#

##code for anove table

#anova.table <- xtable(fit, caption = "Table 2")

#print(anova.table, type = "html")

#

#

##code for pairwise t-tests

#ttests <- pairwise.t.test(subset$wage\_survey\_data.wage,subset$wage\_survey\_data.education,p.adj="bonferroni")

#

##table for pairwise t-test results

#ttest.table <- xtable(ttests$p.value, caption="Table 3")

#print(ttest.table, type = "html")

#

#####Multiple Regression for determining the effect of 2 independent variables on dependent variable

#

##### Age + education ----> Wage

#

## Actually R does the dummy coding for us when using a character or

## factor variable as a predictor in our lm model

#

#model1 <- lm(wage ~ age + education, data = wage\_survey\_data)

#

#str(model1)

#summary(model1)

#confint(model1, conf.level=0.90)

#plot(model1)

#

#####Fitted Regression Equation

#### E1, E2, E3, E4 and E5 are the levels of the education respectively

###Rt is the formula which is obtained from the model1

#

#### Rt = 60.33597 + 0.56869 Age + 11.43 E2 + 24.16 E3 + 39.77 E4 + 64.99 E5

#

###fitted Regression equation for education ---> Below HS Grad

### in this case E2, E3, E4 and E5 will be zero

#

### Rt1 = 60.33597 + 0.56869 Age

#

###fitted Regression equation for education ---> HS Grad

### in this case E3, E4 and E5 will be zero

#

### Rt2 = 60.33597 + 0.56869 Age + 11.43 (1)

### Rt2 = 71.76 + 0.56869 Age

#

###fitted Regression equation for education ---> Some College

### in this case E2, E4 and E5 will be zero

#

#### Rt3 = 60.33597 + 0.56869 Age + 24.16 (1)

#### Rt3 = 84.49 + 0.56869 Age

#

###fitted Regression equation for education ---> college Grad

### in this case E2, E3 and E5 will be zero

#

#### Rt4 = 60.33597 + 0.56869 Age + 39.77 (1)

#### Rt4 = 100.105 + 0.56869 Age

#

###fitted Regression equation for education ---> Advanced Degree

### in this case E2, E3 and E4 will be zero

#

#### Rt5 = 60.33597 + 0.56869 Age + 64.99 E5

##### Rt5 = 125.325 + 0.56869 Age

#

####Plotting the Regression Model (wage ~ age + education)

#

#plot(wage\_survey\_data$age[wage\_survey\_data$education=="1. < HS Grad"], wage\_survey\_data$wage[wage\_survey\_data$education=="1. < HS Grad"], col=2, ylim =c(0,350),xlim=c(0,100), xlab = "Age",ylab = "Wage",

# main="Wage Vs Age, Education")

#

#points(wage\_survey\_data$age[wage\_survey\_data$education=="2. HS Grad"], wage\_survey\_data$wage[wage\_survey\_data$education=="2. HS Grad"], col=3)

#

#points(wage\_survey\_data$age[wage\_survey\_data$education=="3. Some College"], wage\_survey\_data$wage[wage\_survey\_data$education=="3. Some College"], col=4)

#

#points(wage\_survey\_data$age[wage\_survey\_data$education=="4. College Grad"], wage\_survey\_data$wage[wage\_survey\_data$education=="4. College Grad"], col=5)

#

#points(wage\_survey\_data$age[wage\_survey\_data$education=="5. Advanced Degree"], wage\_survey\_data$wage[wage\_survey\_data$education=="5. Advanced Degree"], col=6)

#

#legend(5,350,legend=c("Below HS Grad","HS Grad","Some College","College Grad","Advanced Degree"), col=c(2,3,4,5,6), pch = c(1,100), bty ="n")

#

##adding regression line

#

#abline(a=60.33597 , b=0.56869, col=2,lwd=3)

#abline(a=71.76 , b=0.56869, col=3,lwd=3)

#abline(a=84.49 , b=0.56869, col=4,lwd=3)

#abline(a=100.105 , b=0.56869, col=5,lwd=3)

#abline(a=125.325 , b=0.56869, col=6,lwd=3)

#

#####GG plot to demonstrate the relation between ordinal categorical and quantitative data fields

#

#plot1 <- ggplot(data=wage\_survey\_data, aes(x=education, y=wage))

#

##adding layers

#plot3 <- plot1 + geom\_jitter(aes(colour=age)) + geom\_boxplot(alpha=0.7 , outlier.colour = NA)

#

### adding labels legend details and Title to the graph

#plot\_4 <- plot3 +

# xlab("Education") +

# ylab("Wage") +

# ggtitle("Wage by Education, Age")

#

###adding layers

#plot\_final <- plot\_4 + theme(axis.title.x = element\_text(colour = "DarkGreen", size=25),

# axis.title.y = element\_text(colour = "Red", size = 25),

# axis.text.x = element\_text(size=15),

# axis.text.y = element\_text(size = 20),

# legend.title = element\_text(size=25),

# legend.text = element\_text(size = 15),

# plot.title = element\_text(colour = "DarkBlue", size = 35,

# family = "Comic Sans MS"))

#####plotting the final graph

#plot\_final

#

#

########Interation between Age and Education

#

####firstly plot the graph between Age and Education

###non parallel lines

#

#model2 <- lm(wage ~ age + education + age:education, data = wage\_survey\_data)

#summary(model2)

#

##Fitted Regression Equation

#

#### Wage = 68.09017 + 0.38316 \* age + 7.24752 E2 + 0.86720 E3 + 36.83562 E4 + 61.13781 E5

#### + 0.10113 age:E2 + 0.56573 age:E3 + 0.07277 age:E4 + 0.09876 age:E5

#

##############################################################################