MID-ATLANTIC WAGE DATA ANALYSIS

DATA EXPLORATION

The dataset analyzed in this report is wage data found in the ISLR package. The data was assembled by Steve Miller, of Open BI, from March 2011 Supplement to Current Population Survey data (Gareth James, 2017). The summary of dataset is obtained using the following commands:

wage <- read.csv("Wage(1).csv")
dim(wage)
summary(wage)

```
> dim(wage)
[1] 3000   12
> summary(wage)
       X               year           age                      maritl           race                  education
 Min.   :   7373   Min.   :2003   Min.   :18.00   1. Never Married: 648   1. White:2480   1. < HS Grad        :268
 1st Qu.: 85622   1st Qu.:2004   1st Qu.:33.75   2. Married      :2074   2. Black: 293   2. HS Grad          :971
 Median :228800   Median :2006   Median :42.00   3. Widowed      :  19   3. Asian: 190   3. Some College     :650
 Mean   :218883   Mean   :2006   Mean   :42.41   4. Divorced     : 204   4. Other:  37   4. College Grad     :685
 3rd Qu.:374760   3rd Qu.:2008   3rd Qu.:51.00   5. Separated    :  55                   5. Advanced Degree:426
 Max.   :453870   Max.   :2009   Max.   :80.00
                region                  jobclass               health        health_ins      logwage          wage
 2. Middle Atlantic:3000   1. Industrial :1544   1. <=Good     : 858   1. Yes:2083   Min.   :3.000   Min.   : 20.09
                           2. Information:1456   2. >=Very Good:2142   2. No : 917   1st Qu.:4.447   1st Qu.: 85.38
                                                                                     Median :4.653   Median :104.92
                                                                                     Mean   :4.654   Mean   :111.70
                                                                                     3rd Qu.:4.857   3rd Qu.:128.68
                                                                                     Max.   :5.763   Max.   :318.34
```

Above summary highlights the following key characteristics of data:

- There are 3000 observations of 12 variables in the dataset.
- Age, wage and logwage are continuous variables.
- Maritl, health, health_ins, education, race and jobclass are all categorical variables.
- The range of variables year and age shows that data is collected over the period of 6 years from 2003-2009 for the people aged from 18 up to 80 years.
- The summary statistics of variable region shows that the data is collected only in the mid-Atlantic region.
- The dataset provides information on the wage of people categorized into two jobclass: industrial and information along with other data such as marital status, education levels etc.
- There is no information on the gender, so we cannot do gender-based analysis on wage.

CHECK FOR MISSING DATA

Before moving on to the detailed analysis of each variable it is important to check if it contains any missing values and clean and organize the data. To check for missing values following command was used for each variable and it was found that the dataset is free of any missing value:

table(is.na(wage$wage))

```
> table(is.na(wage$wage))

FALSE
 3000
> table(is.na(wage$age))

FALSE
 3000
```

EXPLORING CONTINUOUS VARIABLES

Continuous variables can take infinite number of values. The dataset consists of 3 continuous variables: age, wage and logwage. Statistical analysis of some features for these continuous variables is conducted below.

MEASURES OF LOCATION

	Wage
mean(wage$wage)	111.7036
median(wage$wage)	104.9215
	Age
mean(wage$age)	42.41467
median(wage$age)	42

	Logwage
mean(wage$logwage)	4.653905
median(wage$logwage)	4.653213

Table 1: Measures of Location

MEASURE OF SPREAD

	Wage
sd(wage$wage)	41.7286
quantile(wage$wage)	0% 25% 50% 75% 100% 20.08554 85.38394 104.92151 128.68049 318.34243
range(wage$wage)	20.08554 318.34243
	Age
sd(wage$age)	11.54241
quantile(wage$age)	0% 25% 50% 75% 100% 18.00 33.75 42.00 51.00 80.00
range(wage$age)	18 80
	Logwage
sd(wage$logwage)	0.3517526
quantile(wage$logwage)	0% 25% 50% 75% 100% 3.000000 4.447158 4.653213 4.857332 5.763128
range(wage$logwage)	3.000000 5.763128

Table 2: Measures of Spread

SYMMETRY

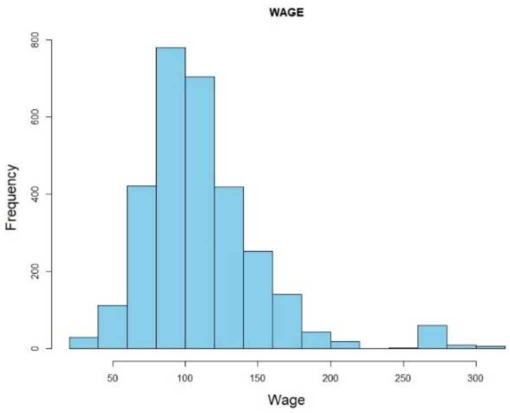


Figure 1: Histogram for Wage

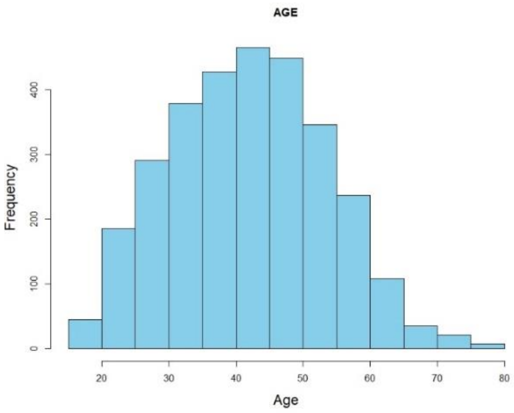


Figure 2: Histogram for Age

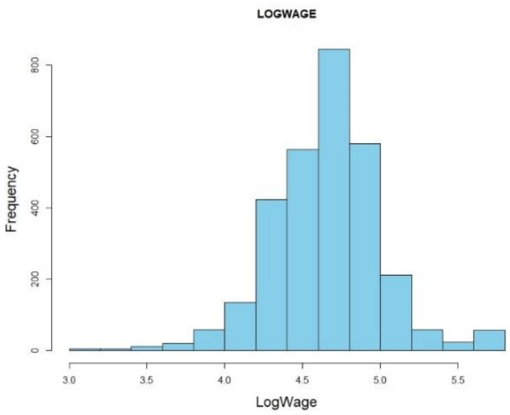


Figure 3: Histogram for Logwage

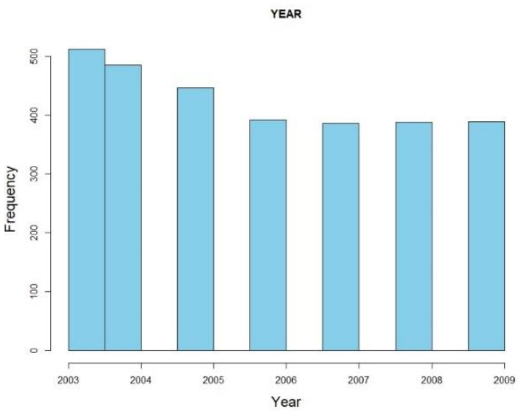


Figure 4: Histogram for Year

With the help of histograms, we found that the data for wage is skewed right and shows major number of people with wage around 100. The positive skewness of wage is also justified by mean and median values, where mean of wage is larger than median (see Table 1). The number of observations in dataset is spread equally over the years 2006-2009, whereas there is a slight tail to the left showing decrease in the number of observations collected in the year 2003-2005. The most common age of people in the dataset is 40-45 years. However, the data for age is also slightly right skewed and the histogram of logwage appears to be left skewed but their mean and median are almost same (see Table 1).

The histograms in figure:1-4 are created by using the following R code:

```
hist(wage$wage, col = "skyblue", main = "WAGE", xlab = "Wage", cex.lab = 1.5)
hist(wage$age, col = "skyblue", main = "AGE", xlab = "Age", cex.lab = 1.5)
hist(wage$logwage, col = "skyblue", main = "LOGWAGE", xlab = "LogWage", cex.lab = 1.5)
hist(wage$year, col = "skyblue", main = "YEAR", xlab = "Year", cex.lab = 1.5)
```

EXPLORING CATEGORICAL VARIABLES

Categorical variables can only take limited and fixed number of values. The dataset consists of 6 categorical variables: maritl, race, education, jobclass, health and health_ins. These variables can be analyzed by using pie charts, bar plots and tables.

COUNT OF EACH CATEGORY

To show the frequency of each category in variables: martial, race, education etc. we have created pie charts with the following line of codes for each variable:

```
label <- paste(names(table(wage$maritl)), "\n", table(wage$maritl))
titles <- "Pie Chart for Maritl"
pie(table(wage$maritl), labels = label, main = titles, cex = 1.5 , cex.main = 2 )
```

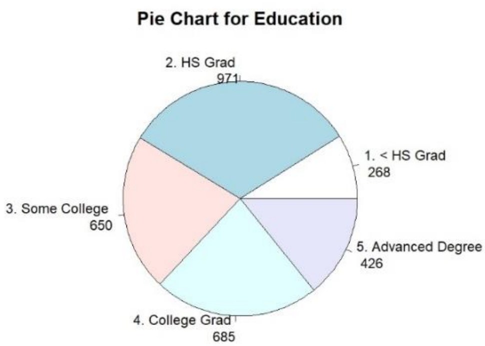


Figure 7: Pie Chart for Education

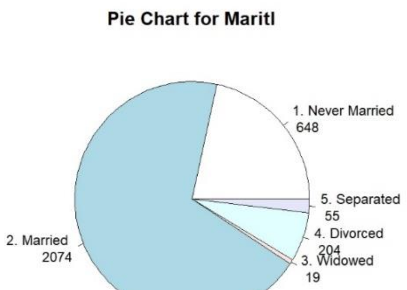


Figure 6: Pie Chart for Marital Status

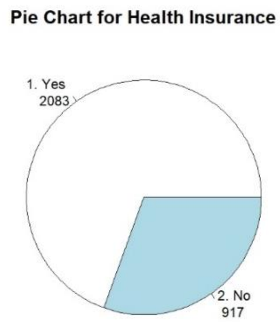


Figure 5: Pie Chart for Health Insurance

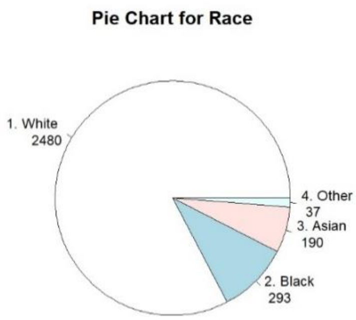


Figure 8: Pie Chart for Race

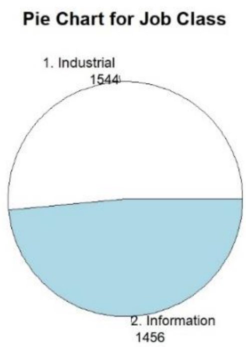


Figure 10: Pie Chart for Job Class

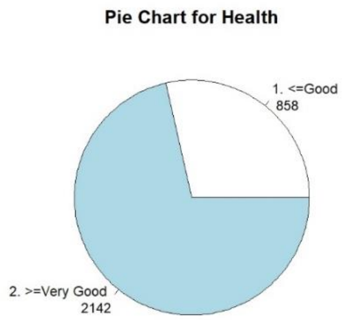


Figure 9: Pie Chart for Health

PROPORTION OF EACH CATEGORY

Pie chart for variable race (Figure 7) shows that the data is biased towards white people as there is a large proportion of white people in the chart. The marital status also appeared to be married for large proportion of people in Figure 6. The dataset for wage shows that out of 3000 people 2083 have their health insurance although most people are very healthy. We will discuss the relationships and impact of these proportions on the wage later. The pie charts show imbalanced categories of race and job classes however it is clear that there is no mislabeled category in any of the variables (Figure 5-10).

DISTRIBUTION OF VARIABLES

The normality of data can be checked with the help of Quantile-Quantile plots. In QQ-plot, R plots the quantiles of sample against the quantiles of normal distribution. To create the QQ-plot for wage following line of code was used in R.

```
qqnorm(wage$wage, col = "darkred", main = "QQ-Plot for Wage", cex.main = 2, cex.lab = 2)
qqline(wage$wage)
```

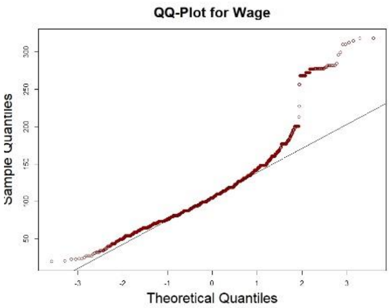


Figure 11: Normal QQ-Plot for Wage

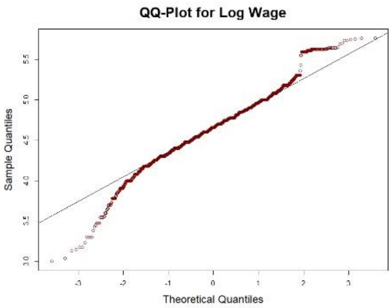


Figure 12: Normal QQ-Plot for Log Wage

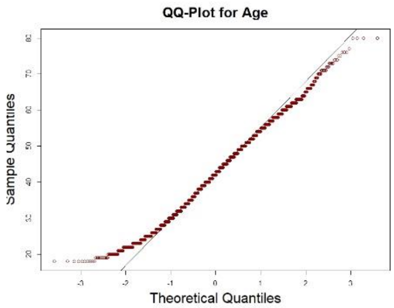


Figure 13: Normal QQ-Plot for Age

The QQ-plot for wage (Figure 11) shows that majority of the points of the sample are close to the line of normal distribution. However, there is some positive skewness in the plot indicating some people with wage values around 230 and some values even greater than 280.

The QQ-Plot for logwage shows that logwage is just the log transformation of the variable wage and it aligns the values more closely to normal distribution (Figure 12). Figure 13 shows that the age of people under observation has normal distribution.

OUTLIERS

Boxplot is an efficient way to illustrate possible outliers in the sample data and gives a concise illustration for quartiles. All the points beyond the whiskers of boxplot indicate outliers. There are many distant values indicating outliers in wage data, according to figure 14. These may be due to the variability in the data or may also be the result of experimental errors. Boxplot for wage is obtained by the following code:

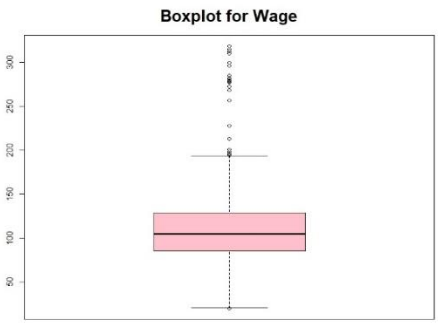


Figure 14: Boxplot for Wage

```
boxplot(wage$wage, main = "Boxplot for Wage", cex.main = 2, cex.lab = 1.5,
col = "pink")
```

Using the above line of codes, boxplots for other variables were obtained. There is only one distant value in figure 15, indicating only one observation of an 80 years old person. There are no outliers in years data.

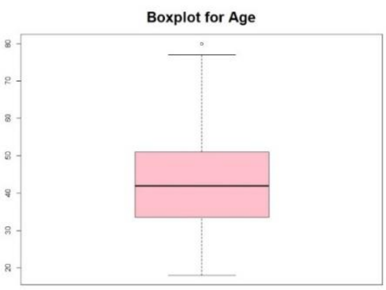


Figure 15: Boxplot for Age

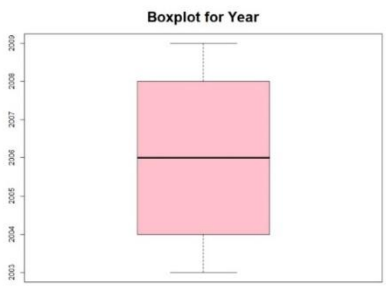


Figure 16: Boxplot for Year

RELATIONSHIP/ASSOCIATION BETWEEN VARIABLES

For plotting continuous variable:wage across the categorical variables: jobclass, education, race and maritl, we have used boxplot.

Relation between Wage and Jobclass

By categorizing the wage of people with the type of jobs they are doing we found that people working in information fields earn slightly more than the people working in industrial areas.

```
boxplot(wage$wage~wage$jobclass, main = "Wage Categorized by Job Class", cex.main = 2, xlab = "Job Class", ylab = "Wage", cex.lab = 1.5, cex.axis = 1.5, col = c("mistyrose", "powderblue"), medcol = c("red","darkblue"), whiskcol = c("red","darkblue"))
```

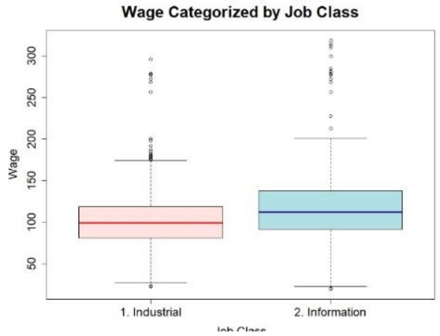


Figure 17: Wage Categorized by Job Class

Relation between Wage and Education

A direct influence of education is seen on the income of people in figure 18, where we have categorized the wage by education levels using boxplots for each factor of education. We can see with more advanced level of education the income also increases. Hence, there is a possible linear relation between education and wage (discussed later).

```
boxplot(wage$wage~wage$education, main = "Wage Categorized by Education", cex.main = 2, xlab = "Education Level", ylab = "Wage", cex.lab = 1.5, col = rainbow(length(unique(wage$education)), alpha = 0.2))
```

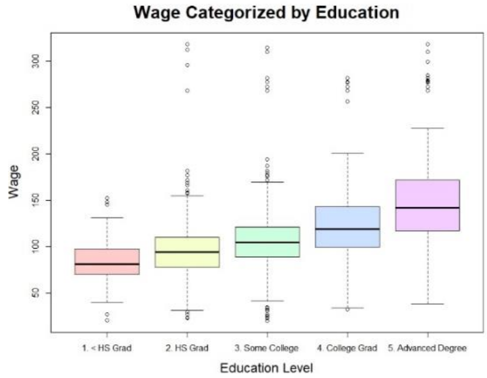


Figure 18: Wage Categorized by Education

Relation between Wage and Race

By categorizing the wage of people in Mid-Atlantic region, we found that Asians are likely to earn more than any other race although earlier we saw that the data collected was biased towards white people.

```
boxplot(wage$wage~wage$race, main = "Wage Categorized by Race", cex.main = 2, xlab = "Race", ylab = "Wage", cex.lab = 1.5, col = rainbow(length(unique(wage$education)), alpha = 0.2))
```

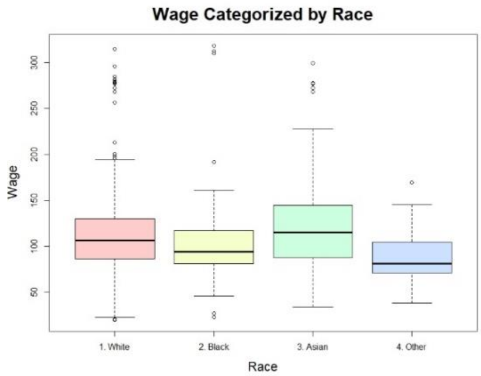


Figure 19: Wage Categorized by Race

Relation between Wage and Marital Status

The relation of marital status with the income of people is illustrated in this boxplot by plotting wage over different categories of marital status. Married people are found to have the highest wage based on the given data in Mid-Atlantic Region.

```
boxplot(wage$wage~wage$maritl, main = "Wage Categorized by Marital Status", cex.main = 2, xlab = "Marital Status", ylab = "Wage", cex.lab = 1.5, col = rainbow(length(unique(wage$education)), alpha = 0.2))
```

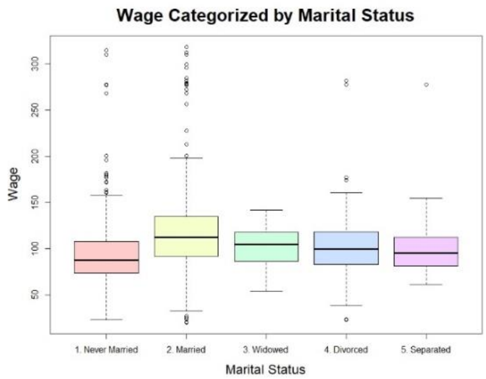


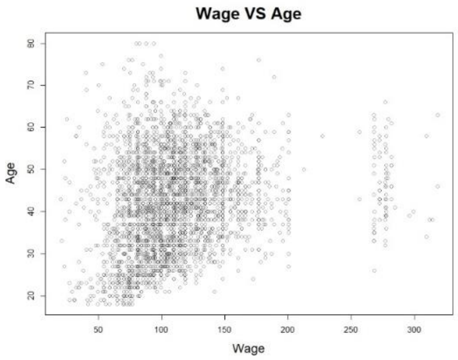
Figure 20: Wage Categorized by Martial Status

PAIR WISE ASSOCIATION BETWEEN VARIABLES/CORRELATION ANALYSIS

TEST OF CORRELATION BETWEEN AGE AND WAGE OF PEOPLE IN MID-ATLANTIC REGION

Preliminary Test Assumptions

Is the covariation linear? To show the relation between two continuous variables: age and wage, we have plotted these using a scatterplot. To test the linearity, we have added a line of linear model in the scatterplot. But the relation between age and wage is still unclear.



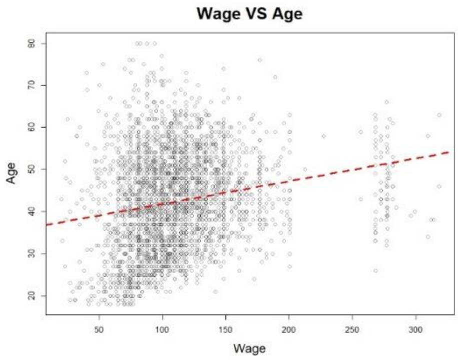


Figure 21: Scatterplot for Wage VS Age *Figure 22: Wage VS Age Scatterplot with Imposed Regression Line*

Is the data normally distributed? Yes, from normality plots in figure 11 and figure 13 we conclude that the population for wage and age follow a normal distribution.

Since both the variables follow normal distribution, we can use Pearson's Correlation test to check the dependence of wage on the variable age.

Pearson Correlation Test

We run the following commands in R for Pearson correlation test:

cor.test(wage$age, wage$wage, method = "pearson")

Test results in R:

```
        Pearson's product-moment correlation

data:  wage$age and wage$wage
t = 10.923, df = 2998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1609777 0.2298147
sample estimates:
      cor
0.1956372
```

Interpretation of the result:

The p-value for the test (2.2e-16) is less than the significance level (alpha=0.05), so we conclude that wage and age are correlated with correlation coefficient (0.195) at 95% confidence interval. Since the correlation coefficient is close to 0, we can state that the association between age and wage is not so strong.

TEST OF ASSOCIATION BETWEEN JOB CLASS AND EDUCATION LEVEL OF PEOPLE IN MID-ATLANTIC REGION

Preliminary Test Assumptions

To test if the two variables: jobclass and education are independent of each other that is there is no impact of education on the type of job people have we use Chi-Square Test of Independence. Therefore, we assume our null hypothesis as jobclass and education are independent variables at significance level alpha = 0.05. Alternative hypothesis will be: jobclass and education are not independent.

Chi-Square Test of Independence

First, we create a contingency table of two categorical variables: jobclass and education in R.

tbl <- table(wage$jobclass,wage$education)

Contingency table

	1. < HS Grad	2. HS Grad	3. Some College	4. College Grad	5. Advanced Degree
1. Industrial	190	636	342	274	102
2. Information	78	335	308	411	324

Now we apply Pearson's Chi-Square Test on the contingency table to find p-value.

chisq.test(tbl)

Test result in R:

```
        Pearson's Chi-squared test

data:  tbl
X-squared = 282.64, df = 4, p-value < 2.2e-16
```

Interpretation of the results:

The p-value of the test (2.2e-16) is less than the significance level (alpha = 0.05), so we reject the null hypothesis that jobclass is independent of education level. We can also see from the contingency table that people with jobs in information have a higher level of education than people working in industry. Hence, this association is also proved in chi-square test.

TEST OF ASSOCIATION BETWEEN EDUCATION AND WAGE

Preliminary Test Assumptions

Visualization of relationship between wage and education using boxplot in figure 18, illustrated that mean value of wage increases as the level of education increases with the wage for advanced degree being the highest. This means we can assume there is a linear relation between the two variables. To test linearity between an independent variable which is categorical and a dependent continuous variable we can use Linear Regression Model.

Linearity Test

Linear Regression:

reg.model <- lm(wage$wage~wage$education)
summary(reg.model)

Test result in R:

```
Call:
lm(formula = wage$wage ~ wage$education)

Residuals:
    Min      1Q  Median      3Q     Max 
-112.31  -19.94   -3.09   15.33  222.56 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       84.104      2.231  37.695  < 2e-16 ***
wage$education2. HS Grad          11.679      2.520   4.634 3.74e-06 ***
wage$education3. Some College     23.651      2.652   8.920  < 2e-16 ***
wage$education4. College Grad     40.323      2.632  15.322  < 2e-16 ***
wage$education5. Advanced Degree  66.813      2.848  23.462  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36.53 on 2995 degrees of freedom
Multiple R-squared:  0.2348,	Adjusted R-squared:  0.2338 
F-statistic: 229.8 on 4 and 2995 DF,  p-value: < 2.2e-16
```

Interpretation of results:

The estimated coefficient of intercept can be interpreted as the mean wage for people who are educated less than HS grads. Other estimated coefficients show that wage increases by 11K for HS Grad, by 23K for some college grads, by 40K for college grads and finally by 66K for people with advance degrees than people with less than HS education. This shows a linear increase in wage with education. The coefficient of variance (Multiple R-squared = 0.234) shows 23% variance in wage.

DESCRIPTIVE ANALYSIS

As dependent variable we have wages of people and as independent variables some are categorical:maritl, race, education, jobclass, health, health_ins and one variable is continuous:age. We have illustrated the association of wage with categorical variables: race, maritl, education and jobclass using boxplots in figure 17-20. The key findings of the analysis highlights that

- Wage of people in information job is slightly higher.
- Wage and education have simultaneous positive variation, but the Linear Regression Model depicts that the linear association is not terribly strong (R-squared = 0.234 is close to 0).
- whereas side-by-side boxplot of maritl, race and wage show fuzzy association.
- We must consider all the outliers in each association.
- Contrasts can only be applied to categorical variables with 2 or more levels we cannot use region variable in analysis of association.

We have computed Pearson's Correlation test for association between age and wage. The correlation coefficient = 0.195 is close to 0, which means the correlation is not strong. However, the positivity of coefficient shows that the variation between both variables is simultaneous.

Why one should focus on predicting log-wage first, and not wage directly?

We can see from figure 1: Histogram for Wage that its tail is extended towards the right. The mean of wage is also greater than it's median, which is clearly illustrated in figure 23 (below). This implies that data of wage is skewed to the right. However, mean and median for log-wage appears to be equal also illustrated by the plot (figure 24). This implies that data for log-wage is normally distributed.

Some statistical techniques might assume that the variable under test has normal distribution, but most techniques are not valid if data is skewed. Since, data for wage appears to have skewed distribution but the log of wage has normal distribution we should focus on predicting log-wage for accuracy of results.

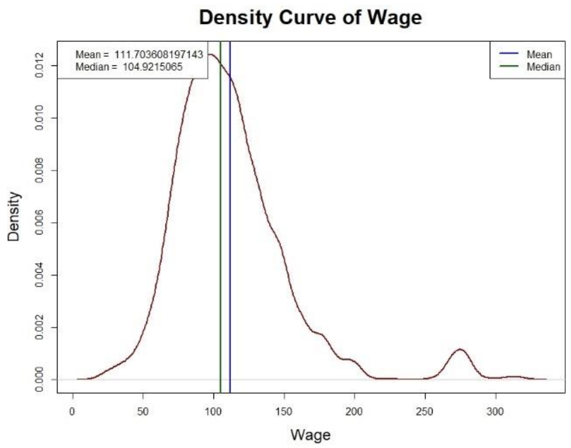


Figure 23:Density Curve of Wage

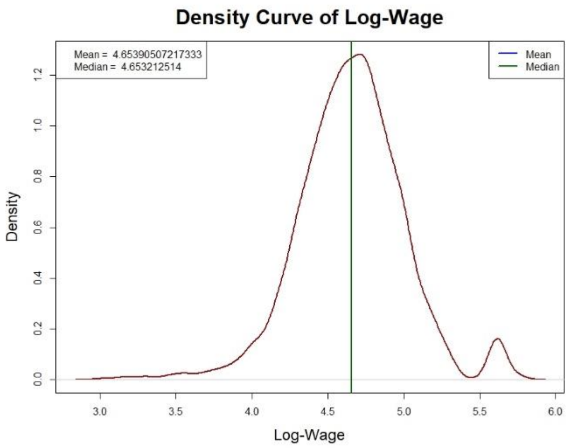


Figure 24: Density Curve of Log-Wage

MULTIPLE LINEAR REGRESSIONS

Which variables can be used to predict wage?

To construct a Multiple Linear Regression we use a call to lm() function in R:

```
AGE <- wage$age
YEAR <- wage$year
MARITAL_STATUS <- wage$maritl
RACE <- wage$race
EDUCATION <- wage$education
REGION <- wage$region
JOB_CLASS <- wage$jobclass
HEALTH <- wage$health
HEALTH_INSURANCE <-
wage$health_ins
LOG_WAGE <- wage$logwage
Wage <- wage$wage
mlr1 <- lm(Wage~AGE+YEAR+
MARITAL_STATUS+RACE+
EDUCATION+JOB_CLASS+
HEALTH+HEALTH_INSURANCE)
summlr1 <- summary(mlr1)
summlr1
```

Call result in R:

```
Call:
lm(formula = Wage ~ AGE + YEAR + MARITAL_STATUS + RACE + EDUCATION +
    JOB_CLASS + HEALTH + HEALTH_INSURANCE)

Residuals:
    Min      1Q  Median      3Q     Max
-100.33  -18.70   -3.26   13.29  212.79

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -2.423e+03  6.165e+02  -3.931 8.67e-05 ***
AGE                         2.707e-01  6.223e-02   4.350 1.41e-05 ***
YEAR                        1.241e+00  3.074e-01   4.037 5.54e-05 ***
MARITAL_STATUS2. Married    1.718e+01  1.720e+00   9.985  < 2e-16 ***
MARITAL_STATUS3. Widowed    2.052e+00  8.005e+00   0.256  0.79774
MARITAL_STATUS4. Divorced   3.967e+00  2.887e+00   1.374  0.16951
MARITAL_STATUS5. Separated  1.153e+01  4.844e+00   2.380  0.01736 *
RACE2. Black               -5.096e+00  2.146e+00  -2.375  0.01760 *
RACE3. Asian               -2.814e+00  2.603e+00  -1.081  0.27978
RACE4. Other               -6.059e+00  5.666e+00  -1.069  0.28505
EDUCATION2. HS Grad         7.759e+00  2.369e+00   3.275  0.00107 **
EDUCATION3. Some College    1.834e+01  2.520e+00   7.278 4.32e-13 ***
EDUCATION4. College Grad    3.124e+01  2.548e+00  12.259  < 2e-16 ***
EDUCATION5. Advanced Degree 5.395e+01  2.811e+00  19.190  < 2e-16 ***
JOB_CLASS2. Information     3.571e+00  1.324e+00   2.697  0.00704 **
HEALTH2. >=Very Good        6.515e+00  1.421e+00   4.585 4.72e-06 ***
HEALTH_INSURANCE2. No      -1.751e+01  1.403e+00 -12.479  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34 on 2983 degrees of freedom
Multiple R-squared:  0.3396,    Adjusted R-squared:  0.3361
F-statistic: 95.89 on 16 and 2983 DF,  p-value: < 2.2e-16
```

Interpretation of call results:

First, we interpret the p-value associated with F-statistics of the model is < 2.2e-16, which is significant, that means there is at least one predictor variables significantly associated with the dependent variable.

Next to interpret which variables are significantly associated we check their corresponding p-value in coefficients table. The variables with p-value < 0.001 are highly significant, which according to the result are: age, year, married, HS Grad, Some College, College Grad, Advance Degree, Health>=Very Good and Health_Insurance2. No. Other variables does not have significant p-values and can be ignored in the model.

The equation of Regression Model after ignoring the variables that are not significantly associated will be:

$$Est[Wage] = intercept + 2.7e^{-1}(Age) + 1.24(Year) + 1.72e^{1}(Married) + 7.76(HS\ Grad) + 1.83e^{1}(Some\ College) + 5.39e^{1}(Advance\ Degree) + 6.52(Health\ V.Good) - 1.75e^{1}(No\ Insurance)$$

Now we can predict the effect of each significantly associated variable on wage by using their corresponding regression coefficients given as estimates in the coefficient table, keeping all other variables constant.

For example, by keeping other variables constant, the wage of a 20 years old person can be estimated to increase by $2.7e^{-1} * 20$, approximately. Similarly, the wage is estimated to decrease by $1.75e^{1}$ for people with no health insurance than people with health insurance. The decrease is due to the negative regression coefficient. The effects all other variables can be interpreted in the same way using the regression equation.

MULTIPLE LINEAR REGRESSION OF LOGWAGE

Call to lm() of wage on all variables:

```
Call:
lm(formula = LOG_WAGE ~ AGE + YEAR + MARITAL_STATUS + RACE +
    EDUCATION + JOB_CLASS + HEALTH + HEALTH_INSURANCE)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.52379 -0.14652  0.00321  0.15584  1.24422 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 -2.102e+01  5.055e+00  -4.159 3.29e-05 ***
AGE                          2.541e-03  5.102e-04   4.981 6.69e-07 ***
YEAR                         1.260e-02  2.521e-03   4.997 6.15e-07 ***
MARITAL_STATUS2. Married     1.646e-01  1.410e-02  11.667  < 2e-16 ***
MARITAL_STATUS3. Widowed     4.898e-02  6.564e-02   0.746  0.45563    
MARITAL_STATUS4. Divorced    4.631e-02  2.367e-02   1.957  0.05047 .  
MARITAL_STATUS5. Separated   1.239e-01  3.972e-02   3.119  0.00183 ** 
RACE2. Black                -4.027e-02  1.759e-02  -2.289  0.02216 *  
RACE3. Asian                -2.120e-02  2.134e-02  -0.993  0.32078    
RACE4. Other                -5.953e-02  4.646e-02  -1.281  0.20021    
EDUCATION2. HS Grad          8.209e-02  1.943e-02   4.226 2.45e-05 ***
EDUCATION3. Some College     1.830e-01  2.066e-02   8.855  < 2e-16 ***
EDUCATION4. College Grad     2.827e-01  2.089e-02  13.530  < 2e-16 ***
EDUCATION5. Advanced Degree  4.333e-01  2.305e-02  18.798  < 2e-16 ***
JOB_CLASS2. Information      2.582e-02  1.086e-02   2.378  0.01745 *  
HEALTH2. >=Very Good         5.917e-02  1.165e-02   5.079 4.03e-07 ***
HEALTH_INSURANCE2. No       -1.936e-01  1.151e-02 -16.824  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2788 on 2983 degrees of freedom
Multiple R-squared:  0.3752,	Adjusted R-squared:  0.3718 
F-statistic:   112 on 16 and 2983 DF,  p-value: < 2.2e-16
```

Call to lm() of wage after step by step reducing the variables from the model:Call results:

```
Call:
lm(formula = LOG_WAGE ~ AGE + YEAR + MARITAL_STATUS + EDUCATION +
    HEALTH + HEALTH_INSURANCE)

Residuals:
     Min       1Q   Median       3Q      Max
-1.50102 -0.14454  0.00658  0.15664  1.24178

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -2.047e+01  5.058e+00  -4.047 5.32e-05 ***
AGE                          2.561e-03  5.079e-04   5.043 4.87e-07 ***
YEAR                         1.232e-02  2.522e-03   4.885 1.09e-06 ***
MARITAL_STATUS2. Married     1.674e-01  1.403e-02  11.936  < 2e-16 ***
MARITAL_STATUS3. Widowed     4.033e-02  6.566e-02   0.614  0.53912
MARITAL_STATUS4. Divorced    4.975e-02  2.364e-02   2.104  0.03546 *
MARITAL_STATUS5. Separated   1.265e-01  3.974e-02   3.184  0.00147 **
EDUCATION2. HS Grad          8.535e-02  1.940e-02   4.400 1.12e-05 ***
EDUCATION3. Some College     1.880e-01  2.057e-02   9.142  < 2e-16 ***
EDUCATION4. College Grad     2.930e-01  2.058e-02  14.240  < 2e-16 ***
EDUCATION5. Advanced Degree  4.464e-01  2.243e-02  19.901  < 2e-16 ***
HEALTH2. >=Very Good         6.043e-02  1.166e-02   5.185 2.31e-07 ***
HEALTH_INSURANCE2. No       -1.967e-01  1.147e-02 -17.150  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2792 on 2987 degrees of freedom
Multiple R-squared:  0.3727,    Adjusted R-squared:  0.3702
F-statistic: 147.9 on 12 and 2987 DF,  p-value: < 2.2e-16
```

Call to anova() to compare above models:

```
> anova(mlr2, mlr4)
Analysis of Variance Table

Model 1: LOG_WAGE ~ AGE + YEAR + MARITAL_STATUS + RACE + EDUCATION + JOB_CLASS +
    HEALTH + HEALTH_INSURANCE
Model 2: LOG_WAGE ~ AGE + YEAR + MARITAL_STATUS + EDUCATION + HEALTH +
    HEALTH_INSURANCE
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   2983 231.84
2   2987 232.76 -4  -0.91643 2.9478 0.01913 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation of results:

The difference between degrees of freedom is 4 for the two models because of reducing 4 variables in the second model. The p-value for the ANOVA test of two models is 0.019 which is not significant, implying that adding variable does not lead to significant improvement. There are two factors to consider for choosing the best fitted model. Minimum Residual Sum of Square RSS value and less variables without affecting the goodness of fit. Now we see residual plots for fitness of model.

Residual Plots

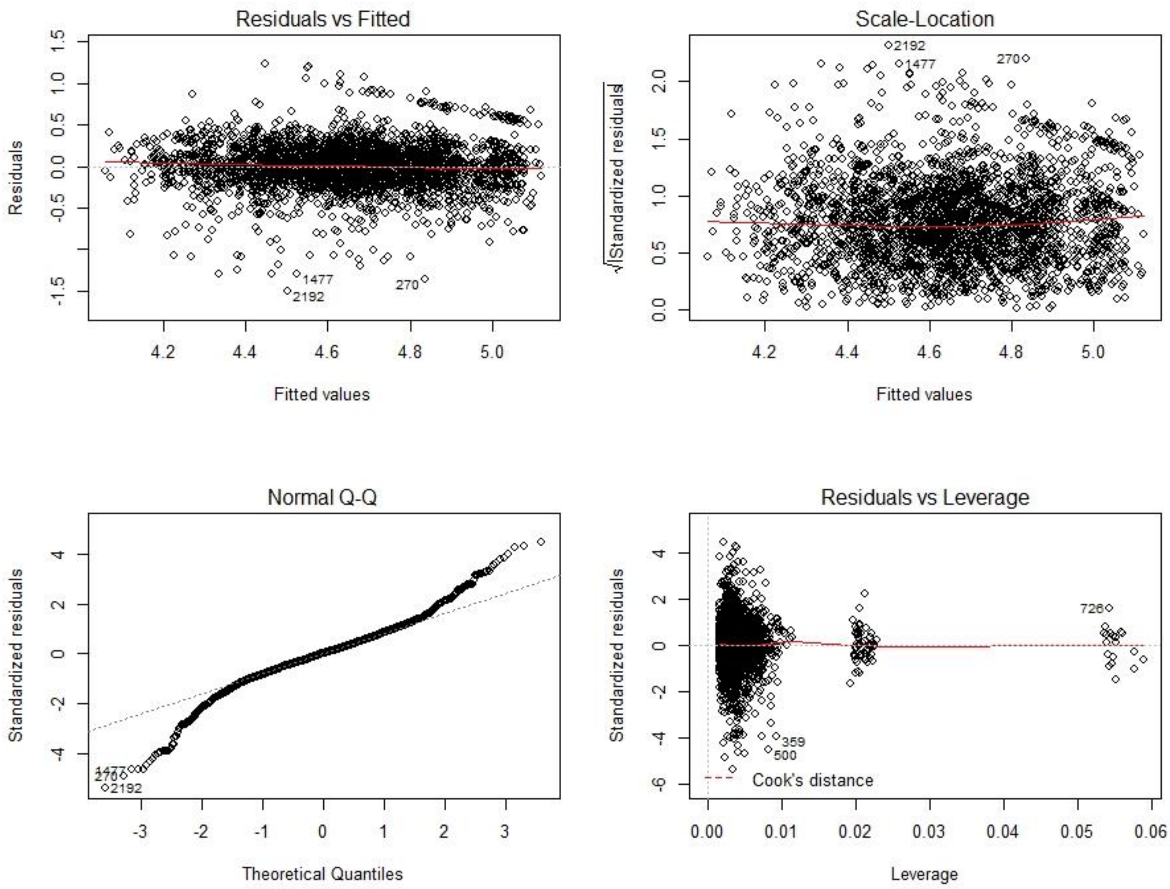
R summarizes the residual of a model in 4 plots:

Residual VS Fitted: It plots the residual over fitted values and the curve of the mode. The curves of both the models sits very straight.

Normal QQ-Plot: This plot shows the distribution of residual by comparing it with the normal curve. There is very slight difference between the curves of both models.

Scale-Location Plot: This plots the curve of squared standardized residual against the values of model. It is observed that the homoscedacity of both models appears to be almost same.

Residual VS Leverage: Smaller value of maximum leverage shows better fit of model. This plot illustrated a smaller leverage for model 2 (with reduces variables).



WEAKNESS AND EFFECTIVENESS OF ANALYSIS

Multiple linear regression provides a great way for analysis of association between the response and independent variables. It also provides a method of prediction of variance on the response variable. However, there are some limitations for this analysis:

- The model provides analysis of correlation between the variables which means the association should be linear. However, in our data there might be some other dependencies between the variables which cannot be analyzes using Multiple Linear Regression.
- This model is fit for analysis of data which is normally distributed. However, our primary responsible variable could not satisfy this assumption. Fortunately, the logarithm of wage transformed the data to satisfy the requirements of normality for accuracy of results.

References

Gareth James, D. W. (2017, 10 19). *Package 'ISLR'.* Retrieved from Cran.r-project.org: https://cran.r-project.org/web/packages/ISLR/ISLR.pdf