**Application of Linear Regression To Estimate Bill**

**Business Objective**: Martin leads the customer relationship team for a telecom giant. He analyzed the historical data for service request and found that there have been frequent requests for plan change by customers. As a proactive measure to resolve this issue, Martin contacted the Data Analytics and Business Insights team to build a tool that can estimate the right bill for the Customers.

Data analytics team proposed to build a regression model based on certain parameters that can be collected at the time of opening the account.

**Solution:**

Data Analytics team looked into the existing data of customers and extracted the following available parameters to consider in the model to estimate average monthly bill.

|  |
| --- |
| **Parameters** |
| Age |
| salary |
| Number of City lived |
| Type of Job |
| Bill Payer |
| Number of close relatives |
| Travel required |
| Number of Relatives Abroad |

**Linear Regression Details**

Regression gives an equation like Y=mX + C where C is intercept , m is slope of X

Application of Regression

Quantifying the relationship between two continuous variables

Predict (or forecast) the value of one variable from knowledge of the value of another variable

That is an estimating equation - a mathematical formula - will be developed

**Correlation**:

Correlation coefficient, **‘r’** lies between -1-0-+1 where 0 means relations is not linear

When the pattern of relationship is known, Correlation analysis can be applied to determine the degree to which the variables are related

Correlation analysis informs how well the estimating equation actually describes the relationship

For Stronger correlation r should be greater or less than 0.5

**Correlation Coefficients**

The strength of the relationship between two variables is measured by the coefficient of correlation coefficient ρ . For a sample we estimate rho using Pearson’s correlation coefficient r for sample and ‘R’ for population.

Negative correlation coefficients indicate negative relationships. i.e. As one variable increases, the other decreases

Stronger linear relationships have values closer to ± 1, weaker linear relationships have values closer to 0.

0 indicates no relationship at all and the relationship is not linear.

± 1 indicates a perfect relationship

**Correlation Causation**

Correlation analysis helps determine degree of relationship between two or more variables

It does not tell about cause and effect relationship

Even high degree of correlation does not necessarily mean a relationship of cause and effect exists between variables

Correlation does not imply causation though the existence of causation

Correlation does not imply causation though the existence of causation always imply correlation

The significance of a correlation is test using the same method as for the slope of the regression line.

**Coefficient of Determination**

The coefficient of determination **r2** measures how well the line fits the data.

It tells us how much of the variation in Y is explained by the relationship with X.

Ex. R2 = 0.75 means that the changes in Y relationship with X. Ex. R = 0.75 means that the changes in Y due to X are explained by 75% remaining 25% is due to chance or other influences.

**The Regression Model**

In general, the regression equation takes the form;

Y = ß0 + ß1x + e

Where

* y = the dependent variable
* x = the independent variable
* ß0 = The y-intercept
* ß1 = The slope of the line
* e = random error term

The line of best fit is the line that minimizes the spread of these errors

The term (y – yhat) is known as the error or residual.

The line of best fit occurs when the Sum of the Squared Errors is minimized

Ordinary Linear Square (OLS) method for estimating regression equation parameters are only valid if certain conditions below are met:

* The error variable is normally distributed
* The expected value of the error variable is zero
* The variance of the error is constant over the entire range of X values – Homoscadasticity
* The errors associated with any two Y values are independent

**Assessing Assumptions**

Graphical methods are particularly useful for studying potential violations of assumptions above

The simplest way to assess whether or not the residuals are normal is to draw a histogram and

visually inspect the distribution

Using least squares regression method ensures that the expected value of the error variable is zero

**Homoscedasticity** or constant variance is best evaluated by plotting the residuals against the predicted value of the Y variable.

For constant variance, residual and predicted value of the Y variable should not show any trend. Any increasing or decreasing trend is a sign of **Heteroescedasticity**

Residual plots against the X variable can also help us determine whether or not the simple linear model is the most appropriate model for the data. – Linearity assumption

A straight line plot for ‘Residual plots’ against the ‘X variable’ is appropriate for linear relationship while a curvilinear relationship looks appropriate for non linear data;

**Analyzing linear model output:**

* Inspection of a scatter plot of X and Y initially reveals whether the trend is linear
* Inspection of the residual plot also indicates whether the trend is linear
* Inspect the error variables. When the line fits the data well, the residuals are small and hence their variance is also small
* The variance of the residual can be estimated from the standard error of the estimate and is given by the computer
* The size of the residual standard error is however dependent on the sampling units and really only useful for comparing between models

**Significance of the relationship**

Hypothesis testing can be used to determine whether or not parameter estimate is significantly different from zero i.e if the slope is significant.

H0  🡪 slope is zero i.e no relationship

Ha  🡪 slope is not zero i.e relationship exists

Test: T statistics

**Case code and results**

##Set working directory

setwd("E:/self study/R-course-material/R\_WorkDir")

# Read data files for customer and their bill from the data csv.

cust\_billdata<-read.csv("billdata18052014.csv")

#Check if the data id populated/imported properly

head(cust\_billdata)

tail(cust\_billdata)

head(cust\_billdata)

age1 salary Num\_Citylived jobtype Payer ClsRelativesCnt Travel RelativesAbroad

1 21 10000 3 Other Parents 1 Low No

2 21 10000 3 Other Parents 1 Low No

3 16 10000 1 Government Parents 2 Low Yes

4 18 10000 3 Private Parents 3 Low No

5 15 10000 1 Government Parents 2 Low Yes

6 15 10000 3 Private Parents 3 Low No

AvgBill

1 286.25

2 239.25

3 310.25

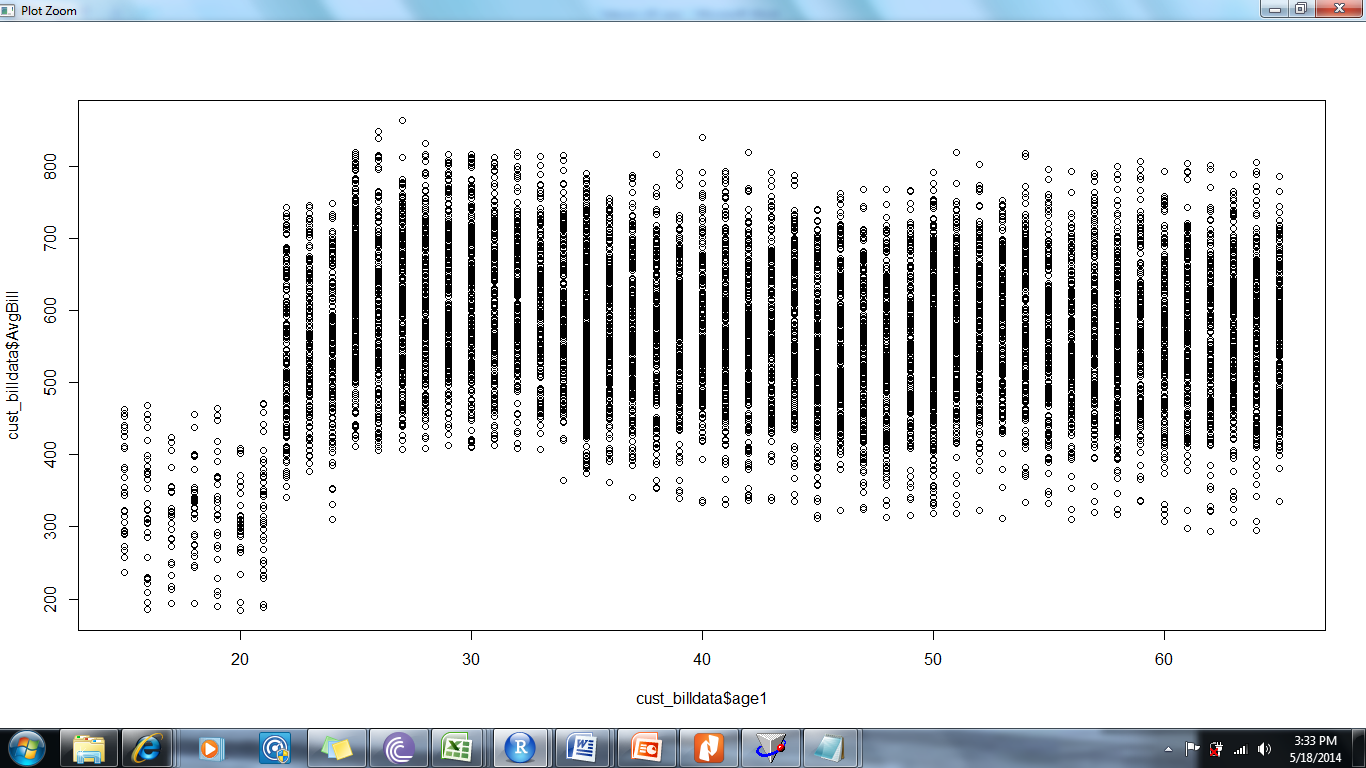
4 347.25

5 290.25

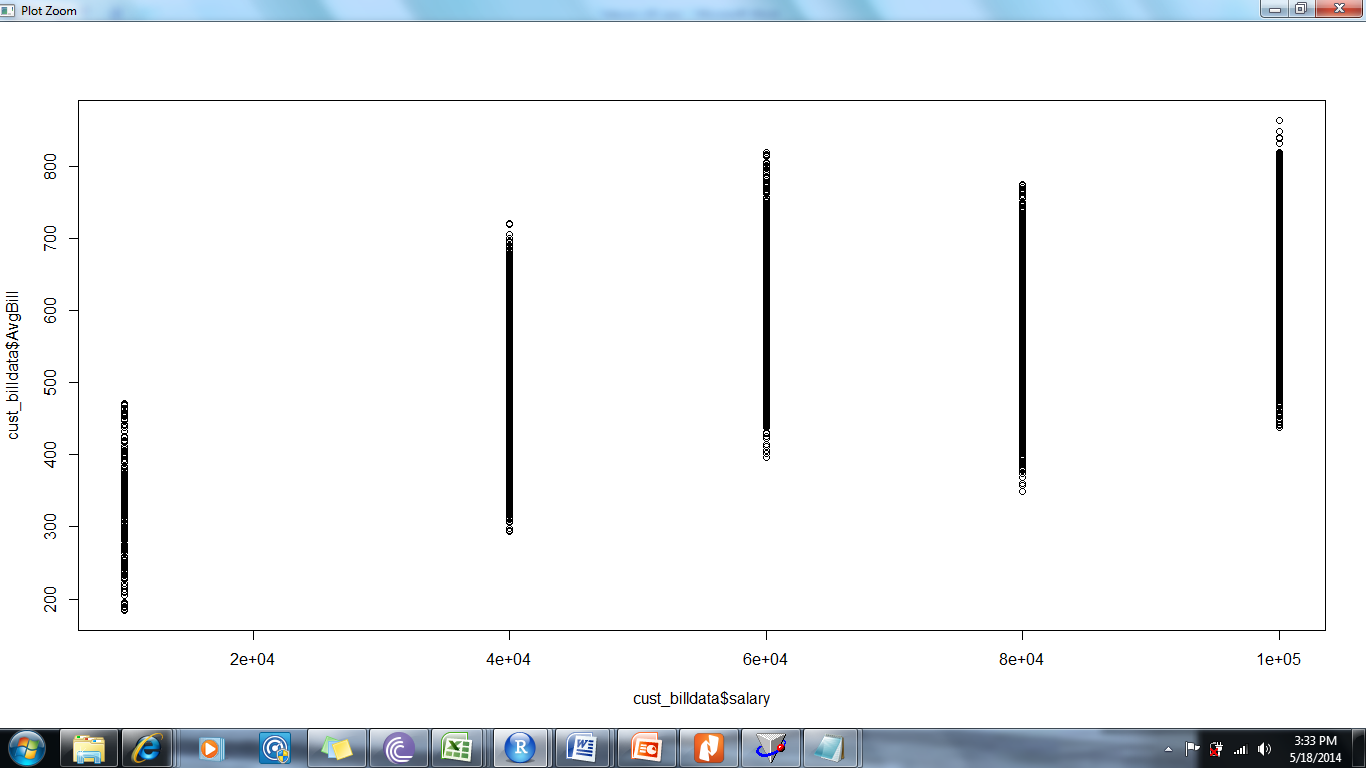
6 409.25

##Generate plots to see the relation between the independent variables and the AvgBill (dependent variable)

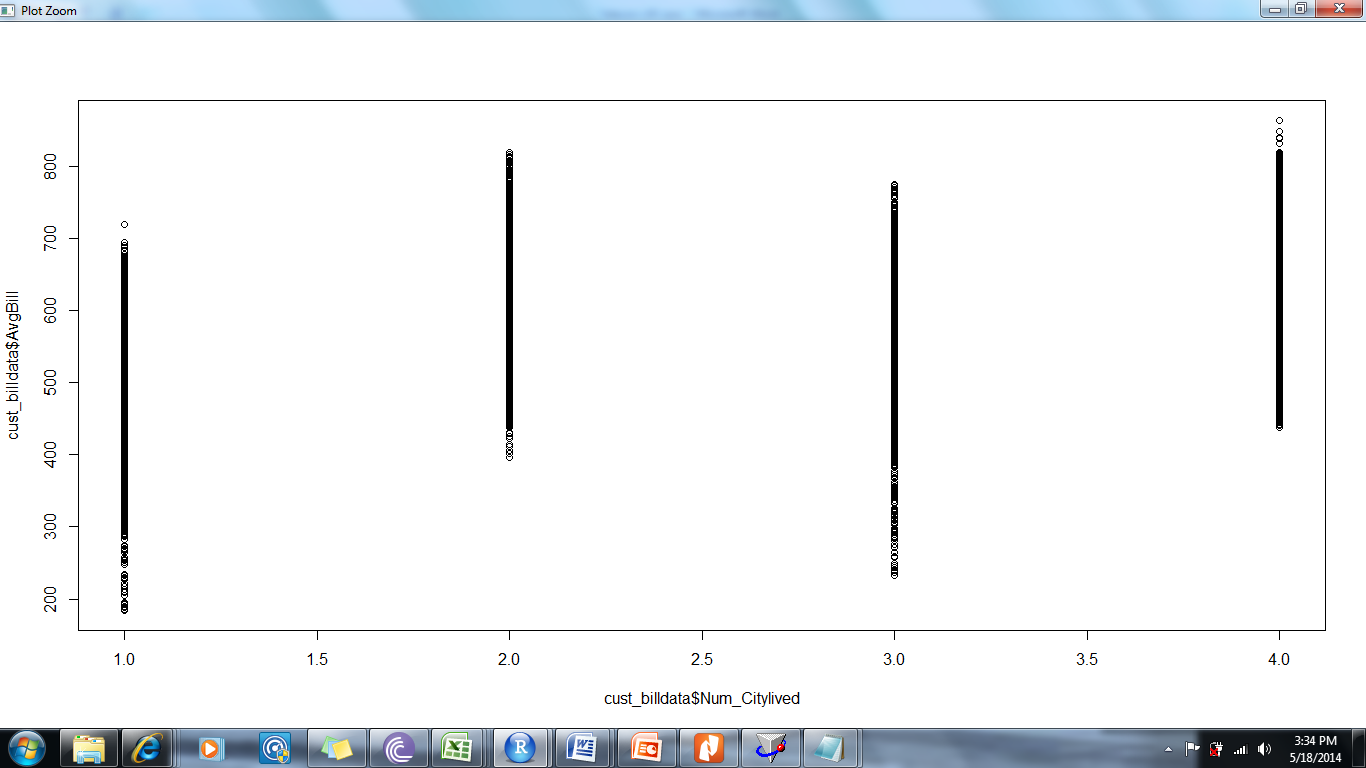
plot(cust\_billdata$age1, cust\_billdata$AvgBill)



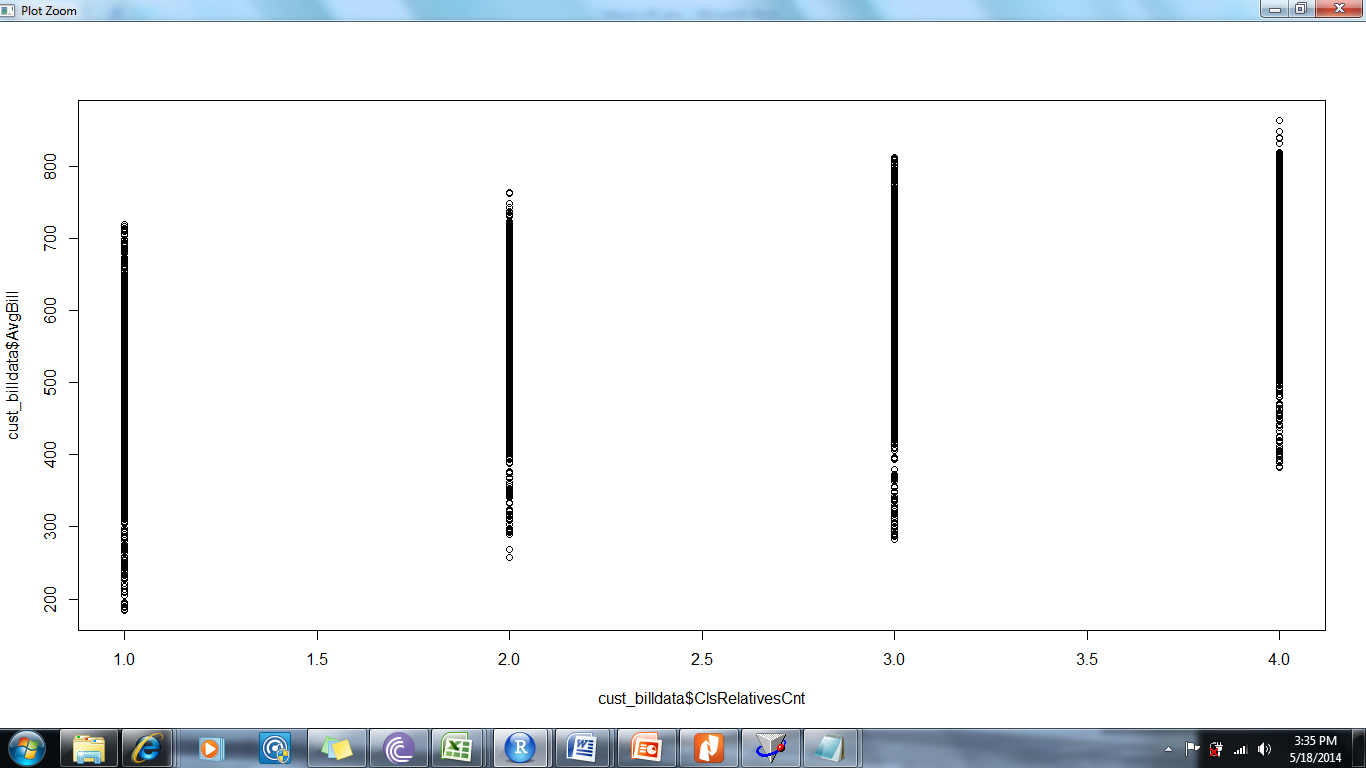
plot(cust\_billdata$salary, cust\_billdata$AvgBill)



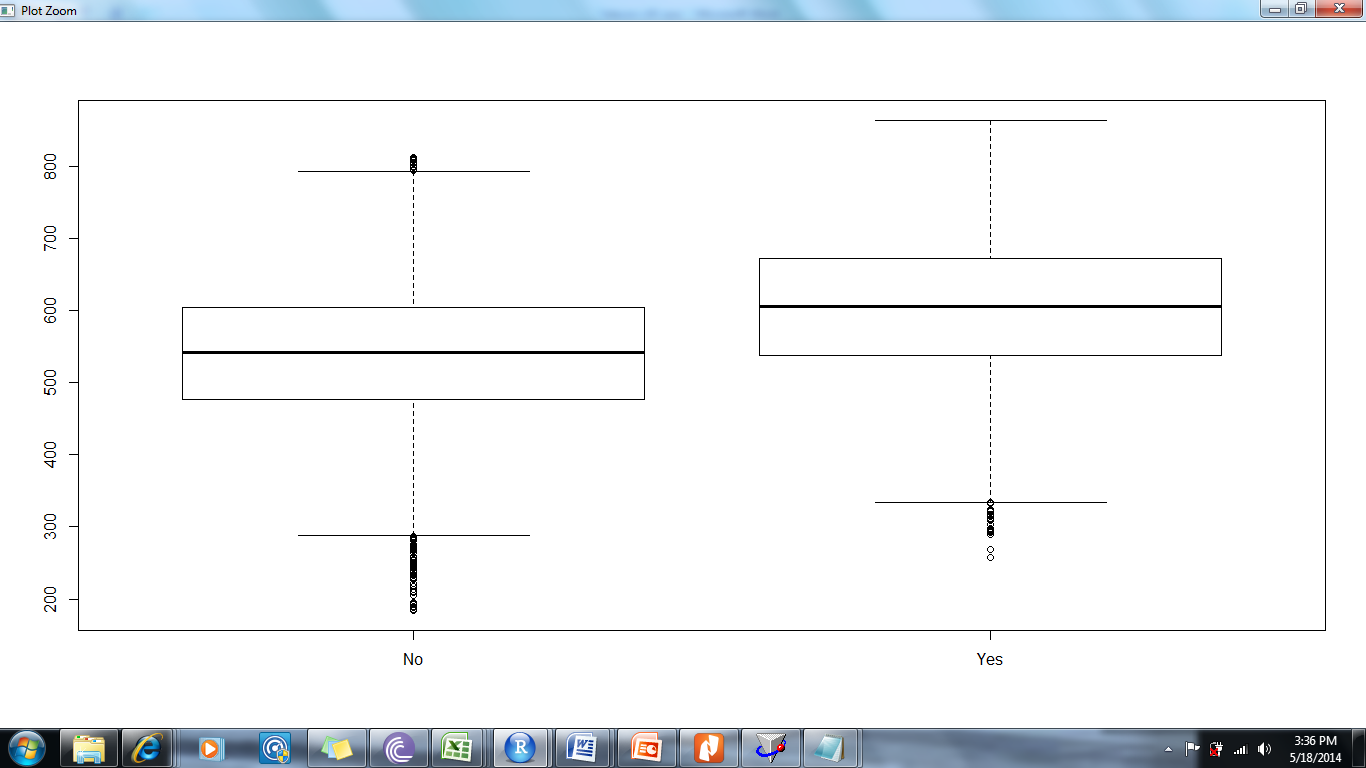
plot(cust\_billdata$Num\_Citylived, cust\_billdata$AvgBill)



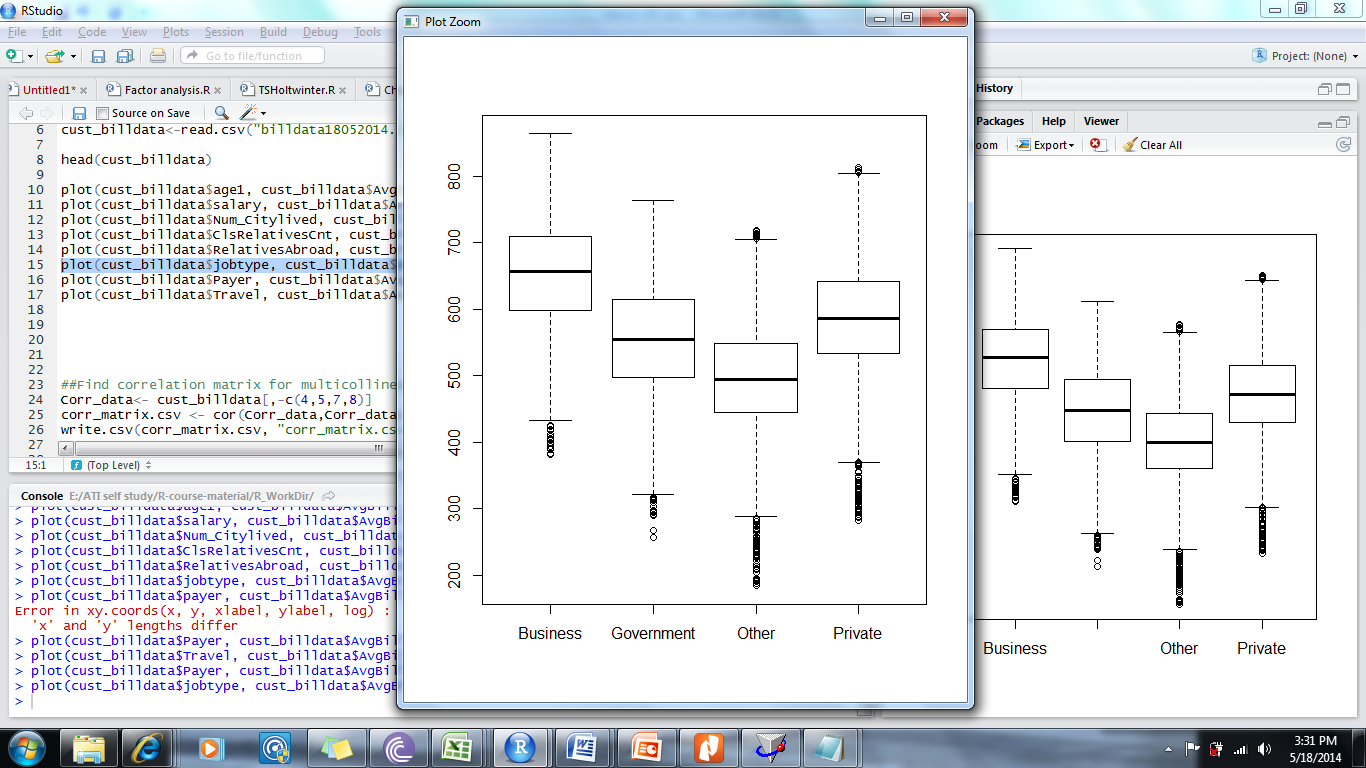
plot(cust\_billdata$ClsRelativesCnt, cust\_billdata$AvgBill)



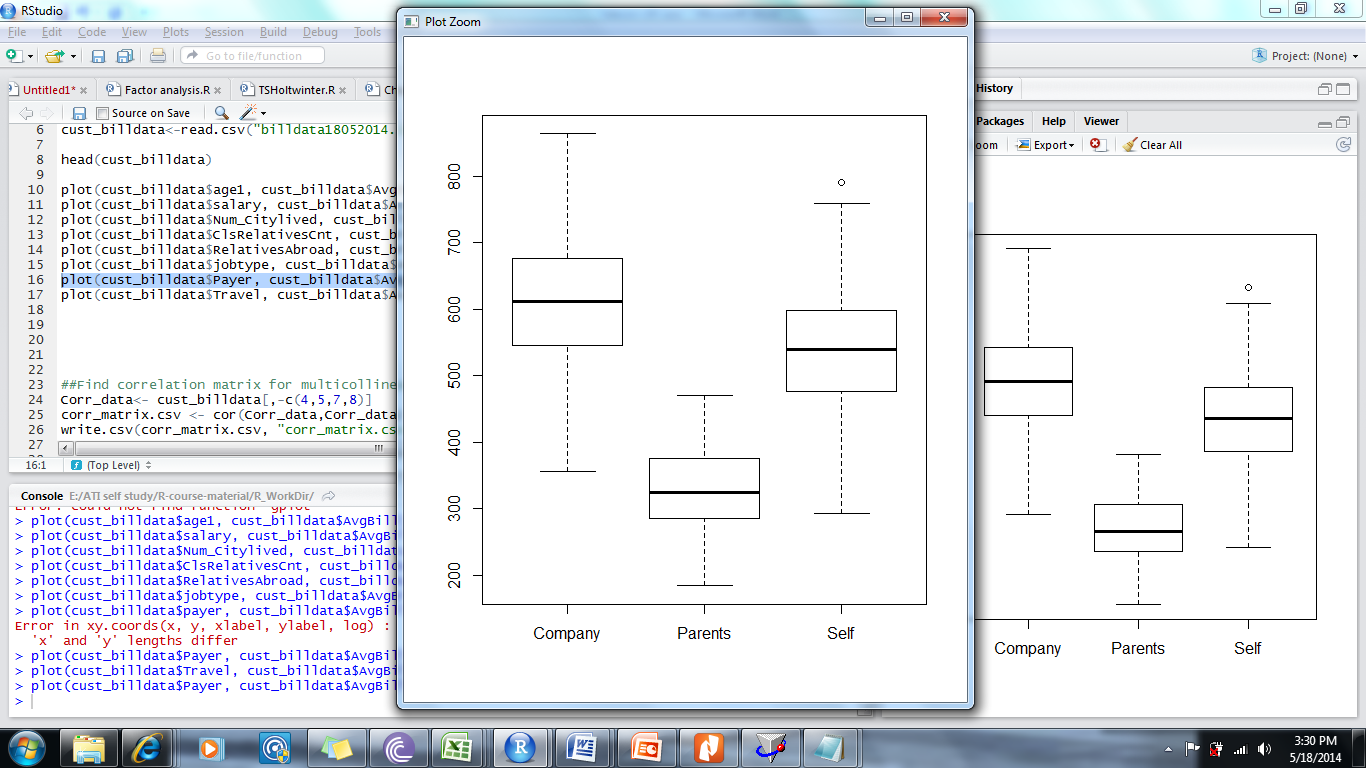
plot(cust\_billdata$RelativesAbroad, cust\_billdata$AvgBill)



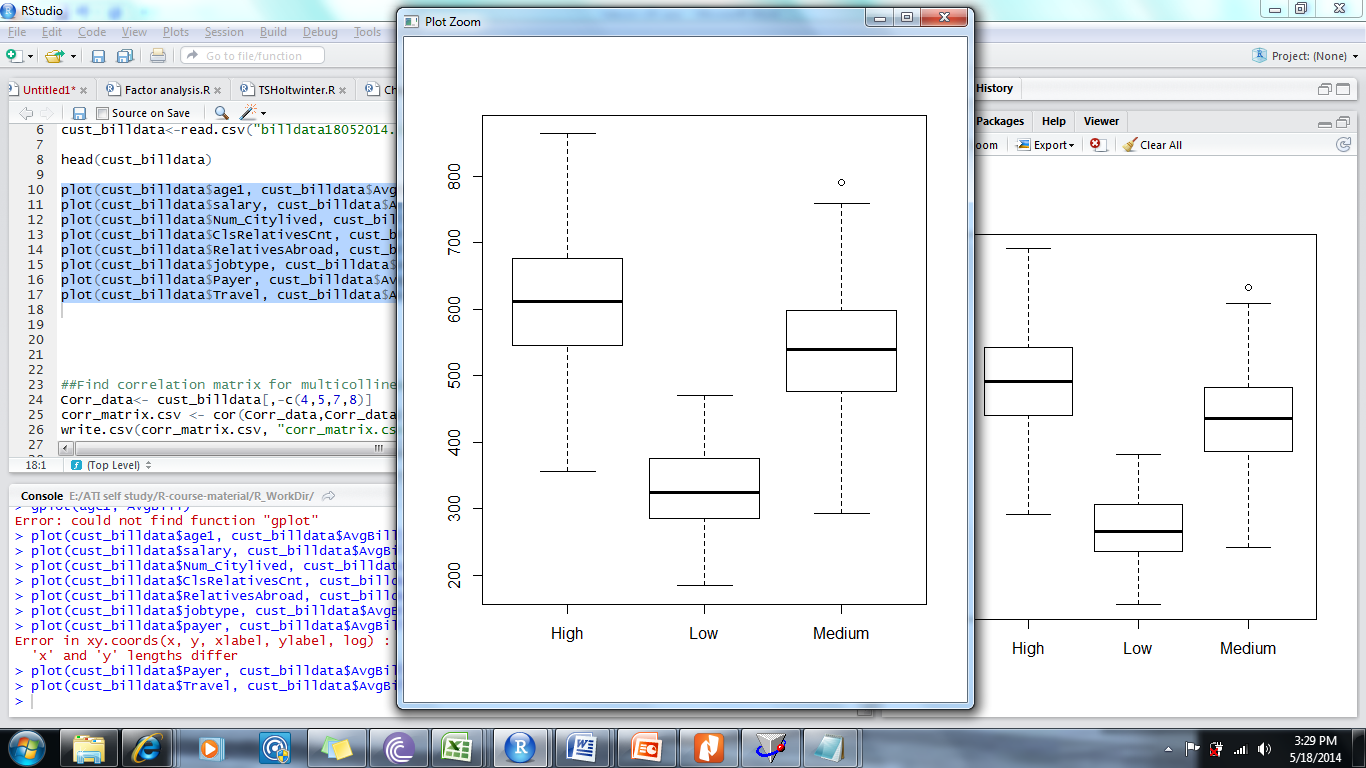
plot(cust\_billdata$jobtype, cust\_billdata$AvgBill)



plot(cust\_billdata$Payer, cust\_billdata$AvgBill)



plot(cust\_billdata$Travel, cust\_billdata$AvgBill)



##Find correlation matrix for multicollinearity check

Corr\_data<- cust\_billdata[,-c(4,5,7,8)]

corr\_matrix.csv <- cor(Corr\_data,Corr\_data)

write.csv(corr\_matrix.csv, "corr\_matrix.csv")

##split data for training and validation

set.seed(3)

train = sample(1:nrow(cust\_billdata),nrow(cust\_billdata)/2)

sample(train)

test = -train

training\_data = cust\_billdata[train,]

testing\_data = cust\_billdata[test,]

**##Run the code to generate model**

fit <- lm(AvgBill ~ age1+ salary +as.factor(jobtype)+Num\_Citylived+as.factor(Payer)+

ClsRelativesCnt+as.factor(Travel)+as.factor(RelativesAbroad), data=training\_data )

**Verify the results:**

summary(fit)

> summary(fit)

Call:

lm(formula = AvgBill ~ age1 + salary + as.factor(jobtype) + Num\_Citylived +

as.factor(Payer) + ClsRelativesCnt + as.factor(Travel) +

as.factor(RelativesAbroad), data = training\_data)

Residuals:

Min 1Q Median 3Q Max

-118.831 -27.058 1.209 27.665 106.449

Coefficients: (4 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.520e+02 2.818e+00 195.89 <2e-16 \*\*\*

age1 -1.286e+00 4.279e-02 -30.06 <2e-16 \*\*\*

salary 1.612e-03 2.404e-05 67.06 <2e-16 \*\*\*

as.factor(jobtype)Government -9.874e+01 1.498e+00 -65.92 <2e-16 \*\*\*

as.factor(jobtype)Other -1.545e+02 1.492e+00 -103.58 <2e-16 \*\*\*

as.factor(jobtype)Private -5.120e+01 1.477e+00 -34.67 <2e-16 \*\*\*

Num\_Citylived 3.148e+01 4.839e-01 65.06 <2e-16 \*\*\*

as.factor(Payer)Parents -1.964e+02 4.111e+00 -47.78 <2e-16 \*\*\*

as.factor(Payer)Self -7.085e+01 1.057e+00 -67.01 <2e-16 \*\*\*

ClsRelativesCnt NA NA NA NA

as.factor(Travel)Low NA NA NA NA

as.factor(Travel)Medium NA NA NA NA

as.factor(RelativesAbroad)Yes NA NA NA NA

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.97 on 4991 degrees of freedom

Multiple R-squared: 0.8655, Adjusted R-squared: 0.8653

F-statistic: 4015 on 8 and 4991 DF, p-value: < 2.2e-16

###Use following command to get the fitted values:

fitted(fit)

RAJ-# Use the below code to see the o/p of plot(fit) in matrix format

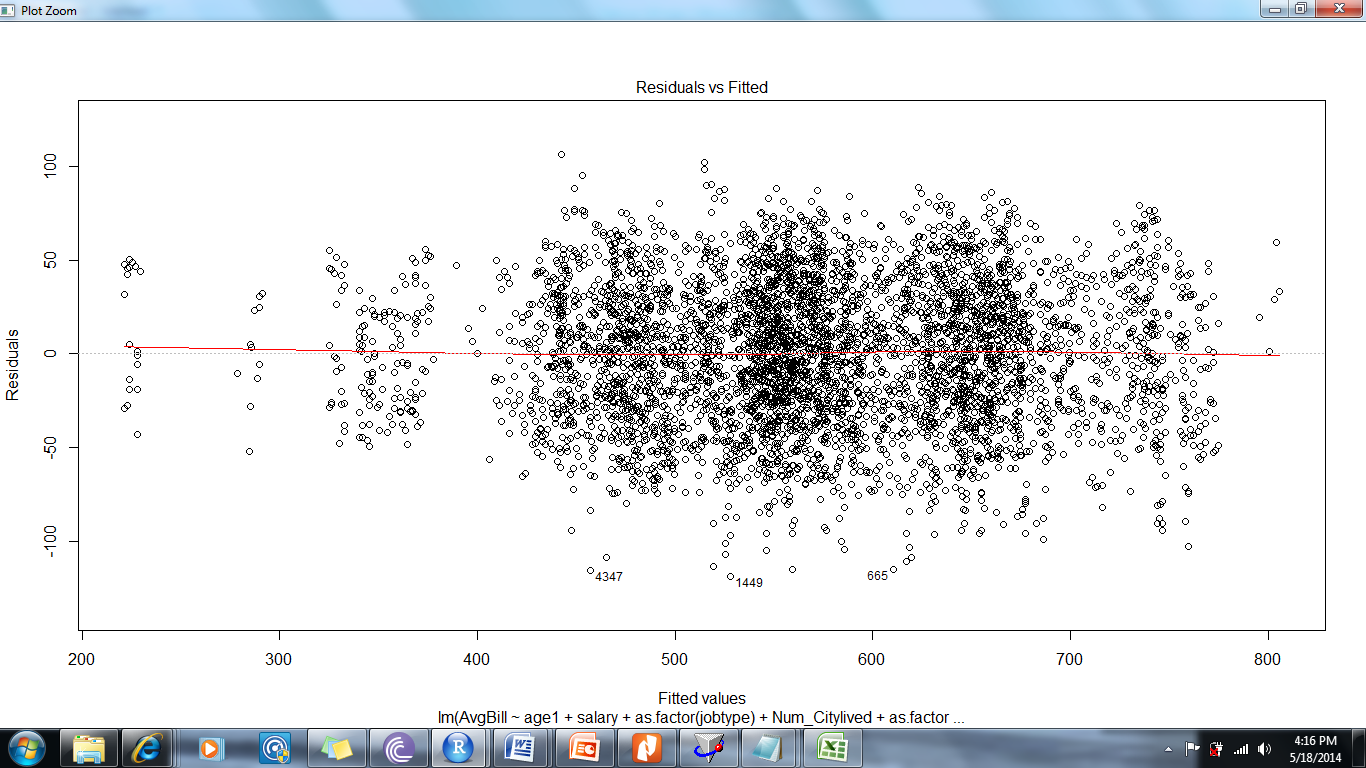
layout(matrix(c(1,2,3,4),2,2))

plot(fit)

http://data.library.virginia.edu/diagnostic-plots/

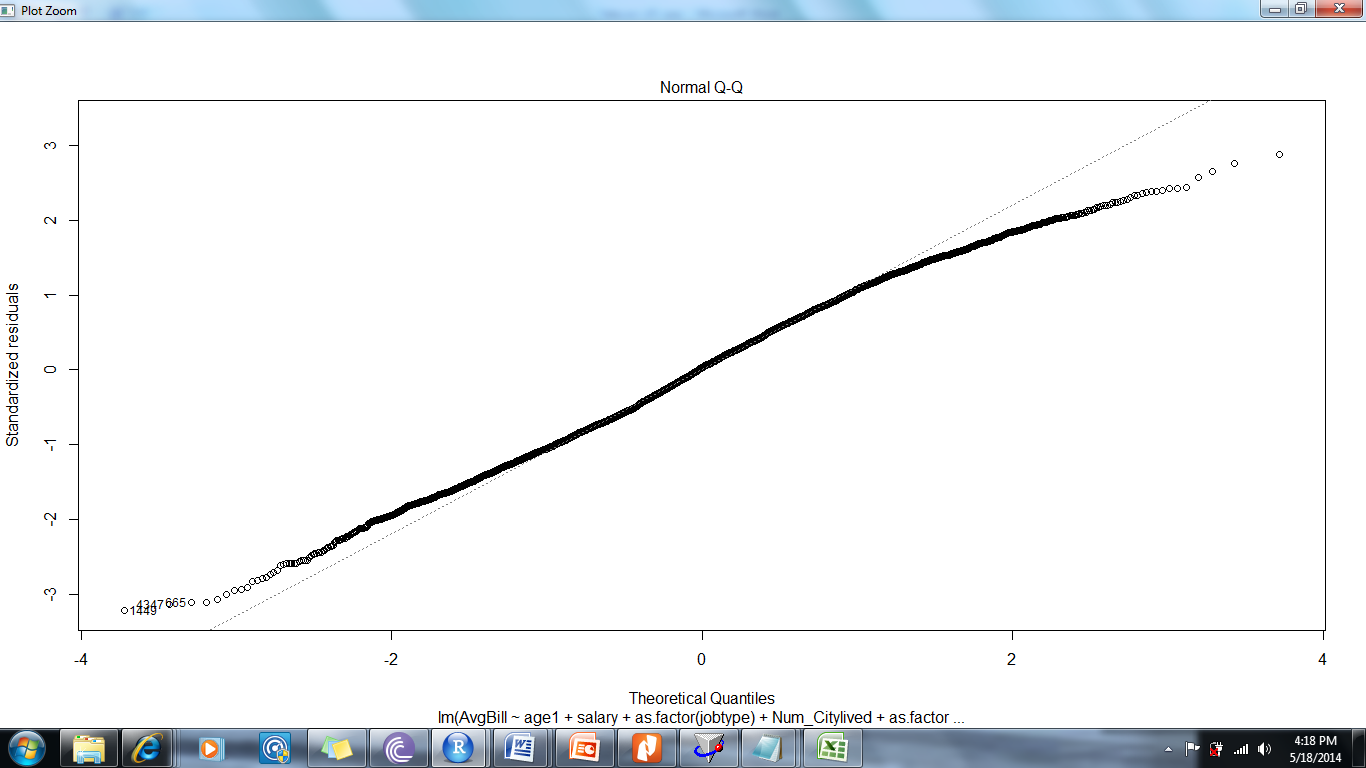
###Diagnostic plots:

Residual Vs Fitted values



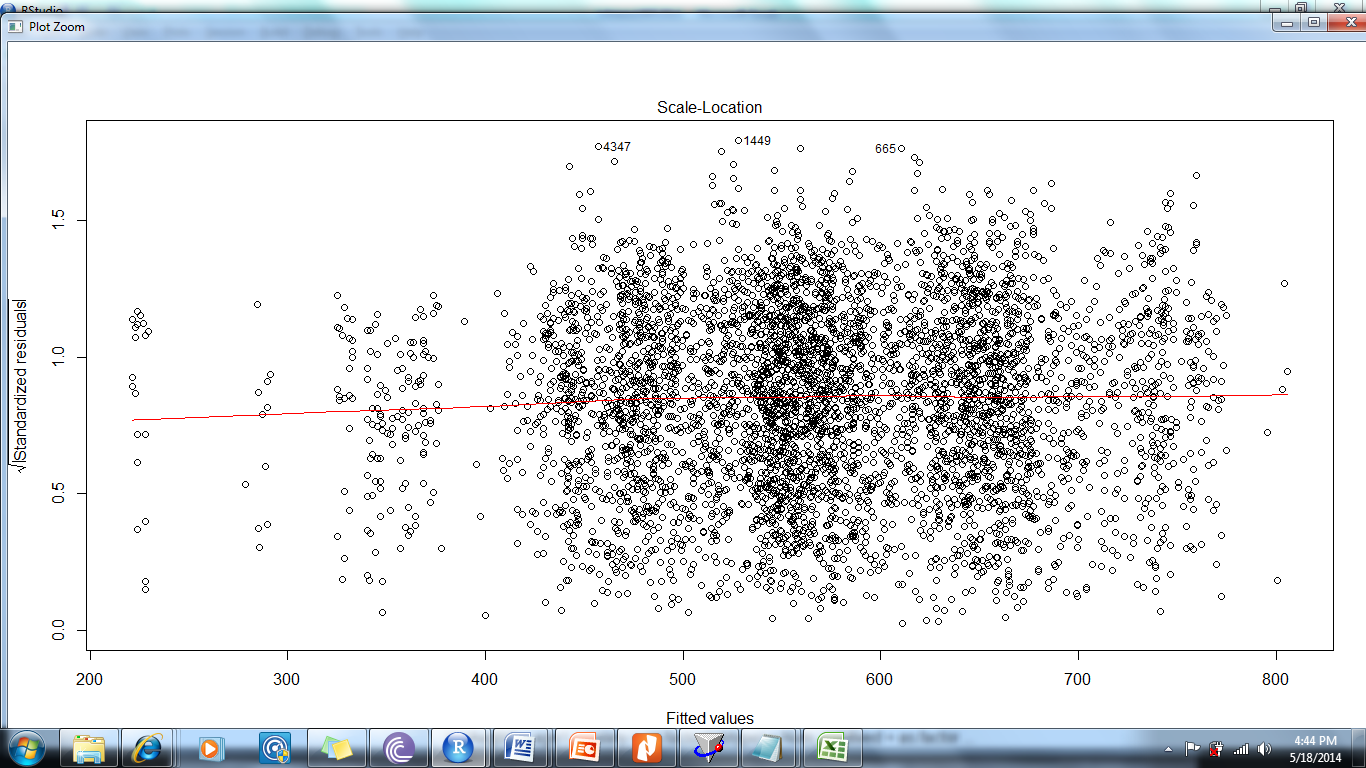
The plot shows no pattern in the residuals hence validates the assumptions of residual independence.

2) Normality:Q-Q plot



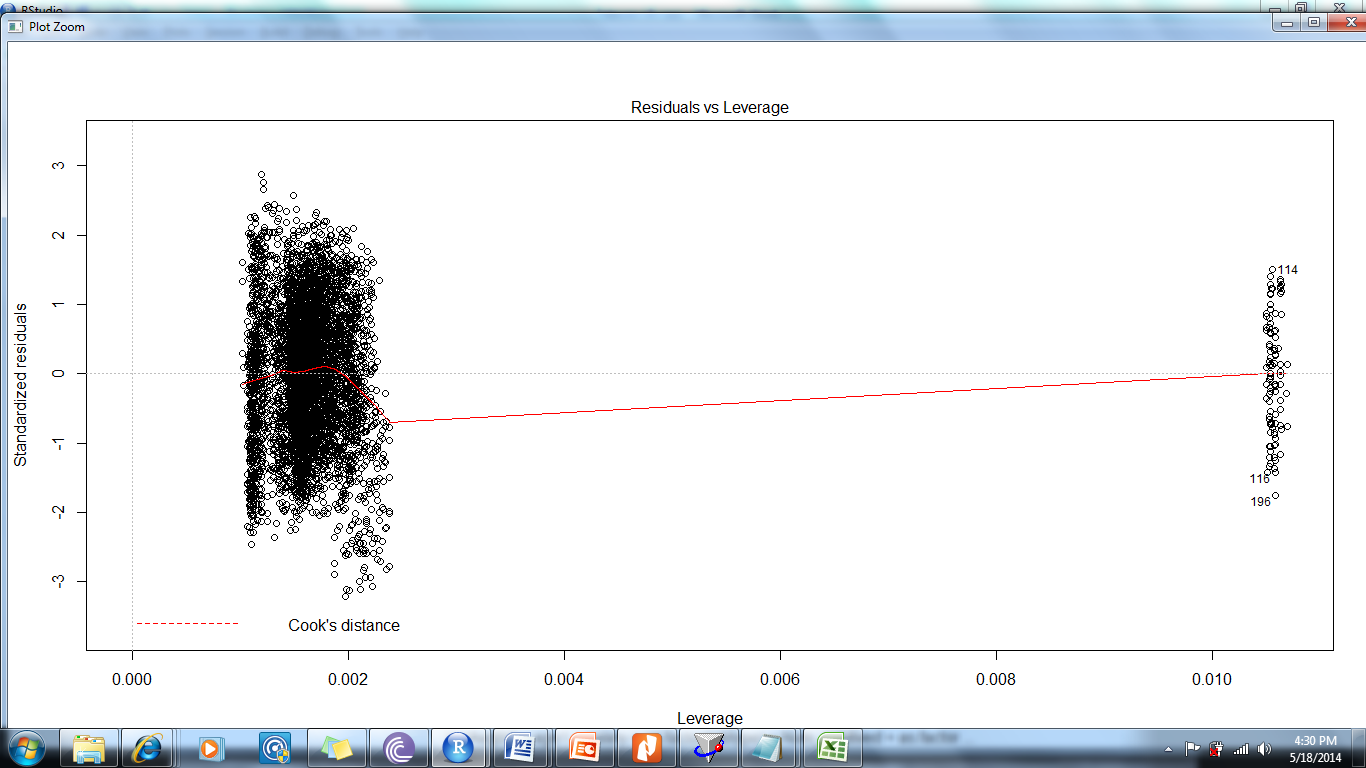
The line lies approximately on the normality curve hence validates the assumptions of normality.

**Scale Location plot:**



The ‘Scale-Location’ plot (Spread-Location or ‘S-L’ plot), takes the square root of the absolute residuals in order to diminish skewness .

**Residual Vs Leverage plot:**



The Cook's distance measures the effect of deleting a given observation. Data points with large residuals ([outliers](http://en.wikipedia.org/wiki/Outlier)) and/or high [leverage](http://en.wikipedia.org/wiki/Leverage_(statistics)) may distort the outcome and accuracy of a regression.

Cook's distance is used to indicate data points that are worth checking for validity.

From the plot above we can say that point 114, 116 and 196 needs to be revisited.

**Validation with Test data:**

Call:

lm(formula = AvgBill ~ age1 + salary + as.factor(jobtype) + Num\_Citylived +

as.factor(Payer) + ClsRelativesCnt + as.factor(Travel) +

as.factor(RelativesAbroad), data = testing\_data)

Residuals:

Min 1Q Median 3Q Max

-118.932 -25.675 -0.234 27.271 101.377

Coefficients: (4 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.518e+02 2.762e+00 199.77 <2e-16 \*\*\*

age1 -1.350e+00 4.219e-02 -32.01 <2e-16 \*\*\*

salary 1.656e-03 2.367e-05 69.96 <2e-16 \*\*\*

as.factor(jobtype)Government -1.007e+02 1.496e+00 -67.35 <2e-16 \*\*\*

as.factor(jobtype)Other -1.525e+02 1.471e+00 -103.69 <2e-16 \*\*\*

as.factor(jobtype)Private -5.196e+01 1.484e+00 -35.00 <2e-16 \*\*\*

Num\_Citylived 3.117e+01 4.814e-01 64.75 <2e-16 \*\*\*

as.factor(Payer)Parents -2.045e+02 3.993e+00 -51.22 <2e-16 \*\*\*

as.factor(Payer)Self -6.917e+01 1.052e+00 -65.74 <2e-16 \*\*\*

ClsRelativesCnt NA NA NA NA

as.factor(Travel)Low NA NA NA NA

as.factor(Travel)Medium NA NA NA NA

as.factor(RelativesAbroad)Yes NA NA NA NA

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.73 on 4991 degrees of freedom

Multiple R-squared: 0.8671, Adjusted R-squared: 0.8669

F-statistic: 4071 on 8 and 4991 DF, p-value: < 2.2e-16

Compare the coefficients, the Residual standard error and the R2 for the training and the testing data. The consistency across the samples highlights the stability of the model.

**Problem:**

Coefficients: (4 not defined because of singularities)

This appears when there is multicollinearity in the data.

To fix this create dummy variables and run code to get correlation matrix.

See the chart below as generated using CORR() function and then highlighted conditionally.



Based on above matrix run the model again after dropping variables like jobtype and payer

See the new results below:

> fit2 <- lm(AvgBill ~ age1+ salary +as.factor(Travel)+Num\_Citylived+as.factor(RelativesAbroad)

+ , data=training\_data )

> summary(fit2)

Call:

lm(formula = AvgBill ~ age1 + salary + as.factor(Travel) + Num\_Citylived +

as.factor(RelativesAbroad), data = training\_data)

Residuals:

Min 1Q Median 3Q Max

-169.555 -49.566 0.161 50.569 143.631

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.594e+02 4.469e+00 102.80 <2e-16 \*\*\*

age1 -1.342e+00 7.221e-02 -18.58 <2e-16 \*\*\*

salary 1.698e-03 4.054e-05 41.88 <2e-16 \*\*\*

as.factor(Travel)Low -1.969e+02 6.938e+00 -28.39 <2e-16 \*\*\*

as.factor(Travel)Medium -7.230e+01 1.784e+00 -40.53 <2e-16 \*\*\*

Num\_Citylived 2.656e+01 8.118e-01 32.72 <2e-16 \*\*\*

as.factor(RelativesAbroad)Yes 5.277e+01 1.772e+00 29.78 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 62.4 on 4993 degrees of freedom

Multiple R-squared: 0.6168, Adjusted R-squared: 0.6163

F-statistic: 1339 on 6 and 4993 DF, p-value: < 2.2e-16

Perform the same validation as above to validate the model across samples.

This model explains 61.68% of variance in the telephone bill.

**Complete code for reference:**

##Set working directory

setwd("E:/ATI self study/R-course-material/R\_WorkDir")

# Read data files for cust and their transactions in a file

cust\_billdata<-read.csv("billdata18052014.csv")

##inspect the imported data

head(cust\_billdata)

##plot Independent var against Dependent var to observe the distribution

plot(cust\_billdata$age1, cust\_billdata$AvgBill)

plot(cust\_billdata$salary, cust\_billdata$AvgBill)

plot(cust\_billdata$Num\_Citylived, cust\_billdata$AvgBill)

plot(cust\_billdata$ClsRelativesCnt, cust\_billdata$AvgBill)

plot(cust\_billdata$RelativesAbroad, cust\_billdata$AvgBill)

plot(cust\_billdata$jobtype, cust\_billdata$AvgBill)

plot(cust\_billdata$Payer, cust\_billdata$AvgBill)

plot(cust\_billdata$Travel, cust\_billdata$AvgBill)

##Find correlation matrix for multicollinearity check

Corr\_data<- cust\_billdata[,-c(4,9,14,18)]

corr\_matrix.csv <- cor(Corr\_data,Corr\_data)

write.csv(corr\_matrix.csv, "corr\_matrix.csv")

#split data for training and validation

set.seed(3)

train = sample(1:nrow(cust\_billdata),nrow(cust\_billdata)/2)

sample(train)

test = -train

training\_data = cust\_billdata[train,]

testing\_data = cust\_billdata[test,]

summary(training\_data)

##Save the training data

write.csv(training\_data, "trainingdata.csv")

##Run the code to generate model

fit <- lm(AvgBill ~ age1+ salary +as.factor(jobtype)+Num\_Citylived+as.factor(Payer)+

ClsRelativesCnt+as.factor(Travel)+as.factor(RelativesAbroad), data=training\_data )

fit2 <- lm(AvgBill ~ age1+ salary +as.factor(Travel)+Num\_Citylived+as.factor(RelativesAbroad)

, data=training\_data )

test\_fit <-lm(AvgBill ~ age1+ salary +as.factor(jobtype)+Num\_Citylived+as.factor(Payer)+

ClsRelativesCnt + as.factor(Travel) + as.factor(RelativesAbroad),

data=testing\_data )

##Generate the summary of the model

summary(fit1)

summary(fit2)

summary(test\_fit)

## Get the fitted values

fitted(fit)

##Generate the diagnostic plots

plot(fit)

plot(test\_fit)

testing\_data1=testing\_data[,c(1,2,3,4,9,13,14,18,21)]

run the model and store in different var say fit\_test

library(forecast)

accuracy(fit\_test$fitted.values,testing\_data$AvgBill)

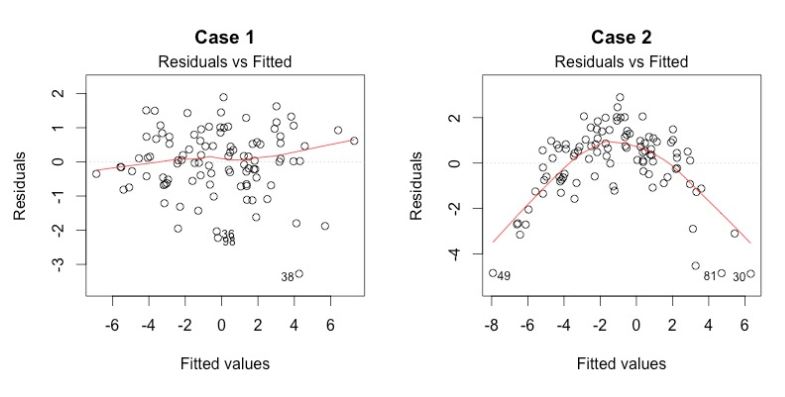
#StepwiseRegression  
library(MASS)  
fit <- lm(y~x1+x2+x3,data=mydata)  
step <- stepAIC(fit, direction="both")  
step$anova # display results

http://data.library.virginia.edu/diagnostic-plots/

**1. Residuals vs Fitted**

This plot shows if residuals have non-linear patterns. There could be a non-linear relationship between predictor variables and an outcome variable and the pattern could show up in this plot if the model doesn’t capture the non-linear relationship. If you find equally spread residuals around a horizontal line without distinct patterns, that is a good indication you don’t have non-linear relationships.

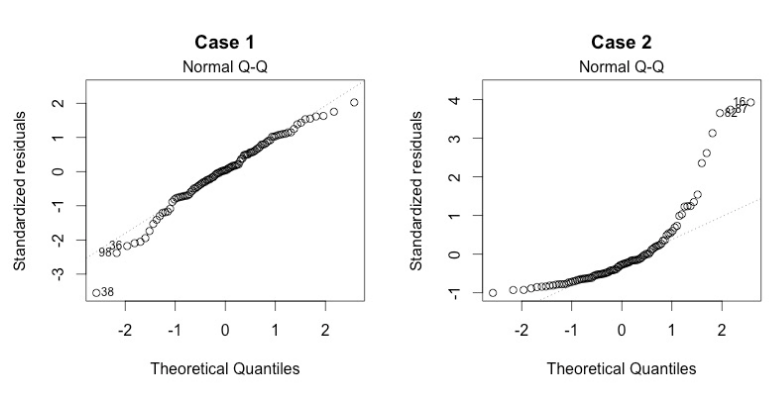
Let’s look at residual plots from a ‘good’ model and a ‘bad’ model. The good model data are simulated in a way that meets the regression assumptions very well, while the bad model data are not.



What do you think? Do you see differences between the two cases? I don’t see any distinctive pattern in Case 1, but I see a parabola in Case 2, where the non-linear relationship was not explained by the model and was left out in the residuals.

**2. Normal Q-Q**

This plot shows if residuals are normally distributed. Do residuals follow a straight line well or do they deviate severely? It’s good if residuals are lined well on the straight dashed line.

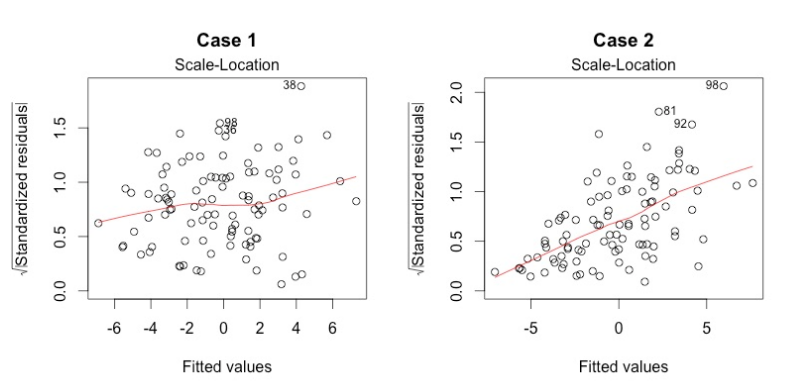


What do you think? Of course they wouldn’t be a perfect straight line and this will be your call. Case 2 definitely concerns me. I would not be concerned by Case 1 too much, although an observation numbered as 38 looks a little off. Let’s look at the next plot while keepig in mind that #38 might be a potential problem.

**3. Scale-Location**

It’s also called Spread-Location plot. This plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity). It’s good if you see a horizontal line with equally (randomly) spread points.

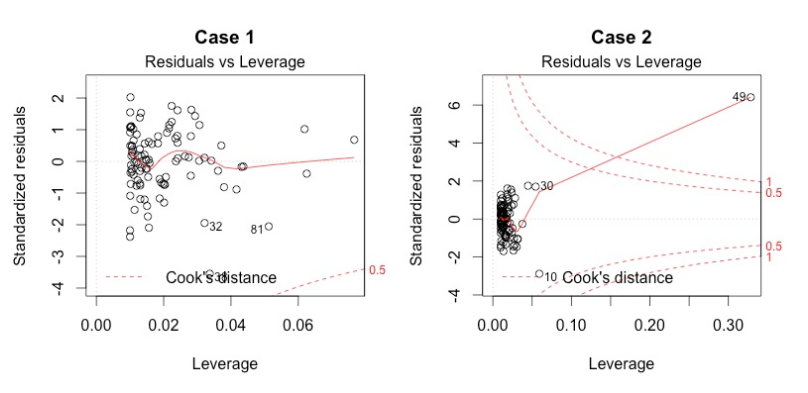
What do you think? In Case 1, the residuals appear randomly spread. Whereas, in Case 2, the residuals begin to spread wider along the x-axis as it passes around 5. Because the residuals spread wider and wider, the red smooth line is not horizontal and shows a steep angle in Case 2.



**4. Residuals vs Leverage**

This plot helps us to find influential cases (i.e., subjects) if any. Not all outliers are influential in linear regression analysis (whatever outliers mean). Even though data have extreme values, they might not be influential to determine a regression line. That means, the results wouldn’t be much different if we either include or exclude them from analysis. They follow the trend in the majority of cases and they don’t really matter; they are not influential. On the other hand, some cases could be very influential even if they look to be within a reasonable range of the values. They could be extreme cases against a regression line and can alter the results if we exclude them from analysis. Another way to put it is that they don’t get along with the trend in the majority of the cases.

Unlike the other plots, this time patterns are not relevant. We watch out for outlying values at the upper right corner or at the lower right corner. Those spots are the places where cases can be influential against a regression line. Look for cases outside of a dashed line, Cook’s distance. When cases are outside of the Cook’s distance (meaning they have high Cook’s distance scores), the cases are influential to the regression results. The regression results will be altered if we exclude those cases.



Case 1 is the typical look when there is no influential case, or cases. You can barely see Cook’s distance lines (a red dashed line) because all cases are well inside of the Cook’s distance lines. In Case 2, a case is far beyond the Cook’s distance lines (the other residuals appear clustered on the left because the second plot is scaled to show larger area than the first plot). The plot identified the influential observation as #49. If I exclude the 49th case from the analysis, the slope coefficient changes from 2.14 to 2.68 and R2 from .757 to .851. Pretty big impact!

The four plots show potential problematic cases with the row numbers of the data in the dataset. If some cases are identified across all four plots, you might want to take a close look at them individually. Is there anything special for the subject? Or could it be simply errors in data entry?