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 $\mathbf{r^2}$ 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. $R^2 = 0.75$ 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;

$Y = ß_0 + ß_1x + e$

Where

- y = the dependent variable
- x = the independent variable
- $ß_0$ = The y-intercept
- $ß_1$ = 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.

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.
H_0 → slope is zero i.e no relationship
H_a → slope is not zero i.e relationship exists
Test: T statistics

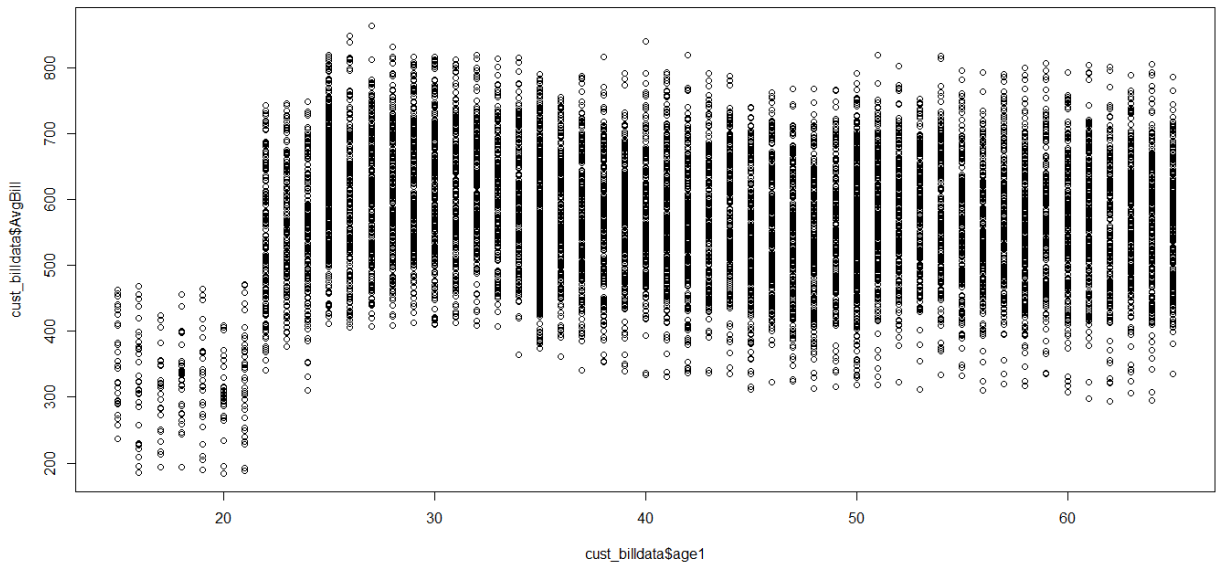
Case code and results

```
##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)
```

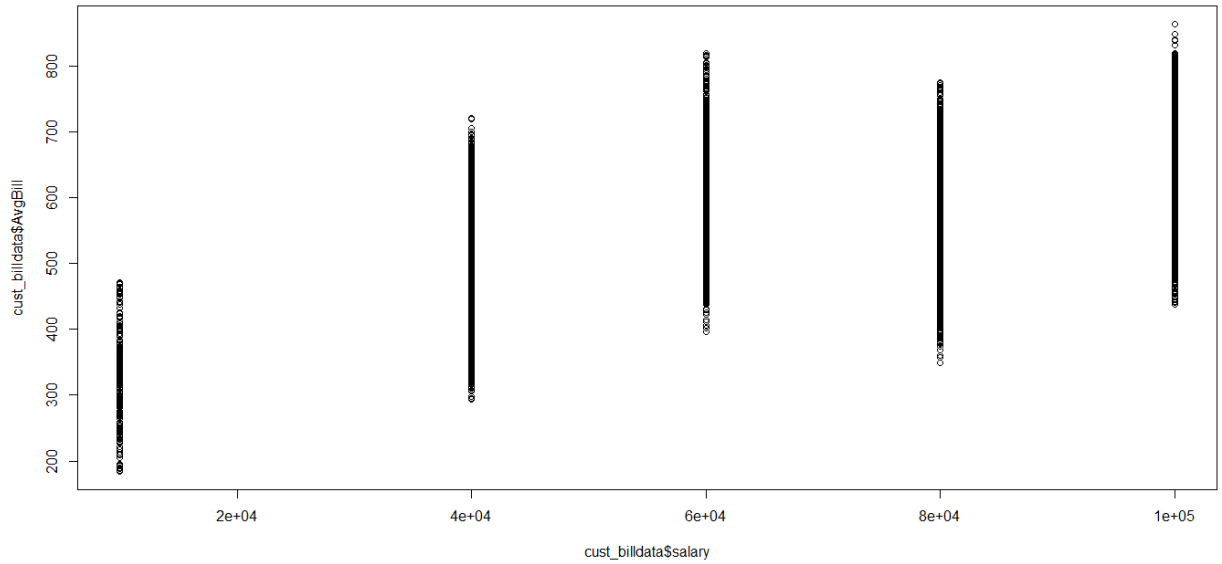
```
head(cust_billdata)
  age1 salary Num_Citylived    jobtype   Payer ClsRelativesCnt Travel
RelativesAbroad
1   21  10000             3      Other Parents              1    Low
No
2   21  10000             3      Other Parents              1    Low
No
3   16  10000             1 Government Parents              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)

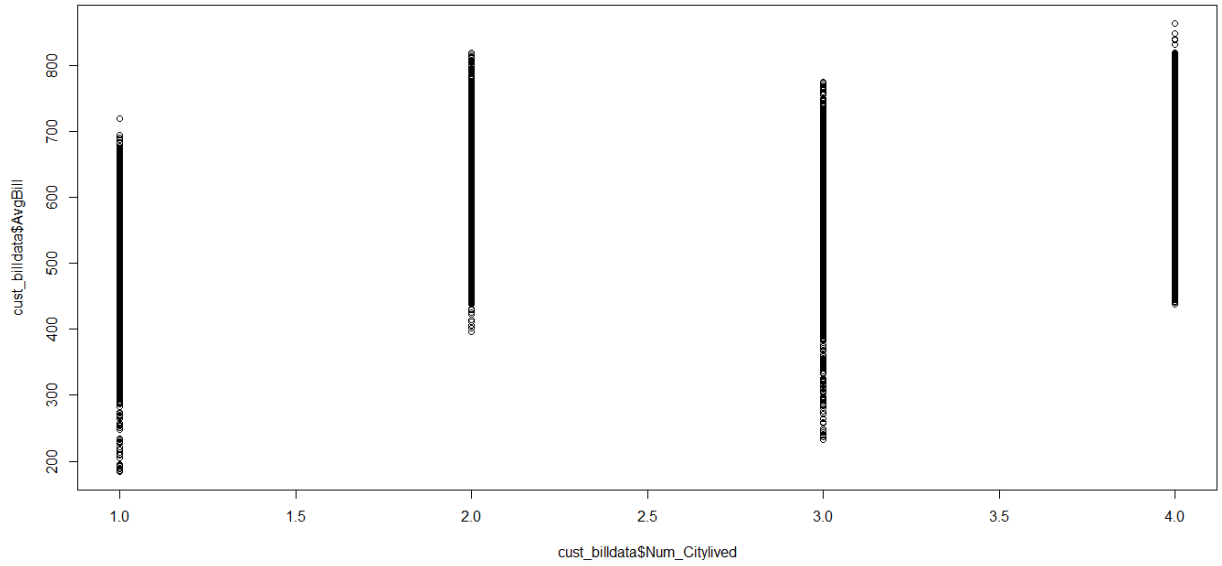
plot(cust_billdata$age1, cust_billdata$AvgBill)



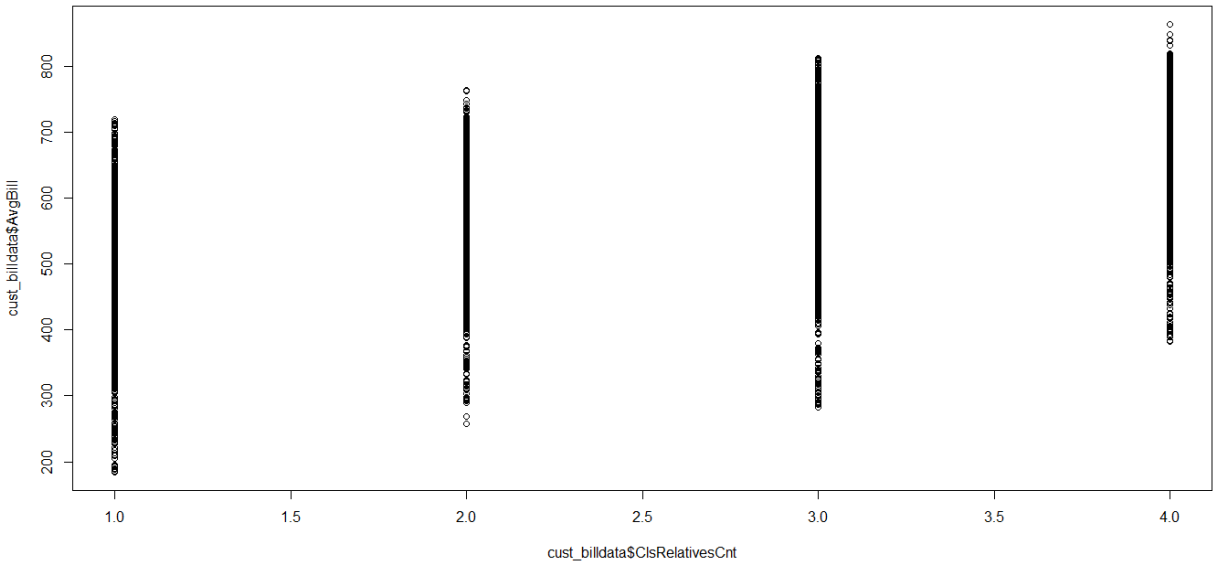
plot(cust_billdata$salary, cust_billdata$AvgBill)



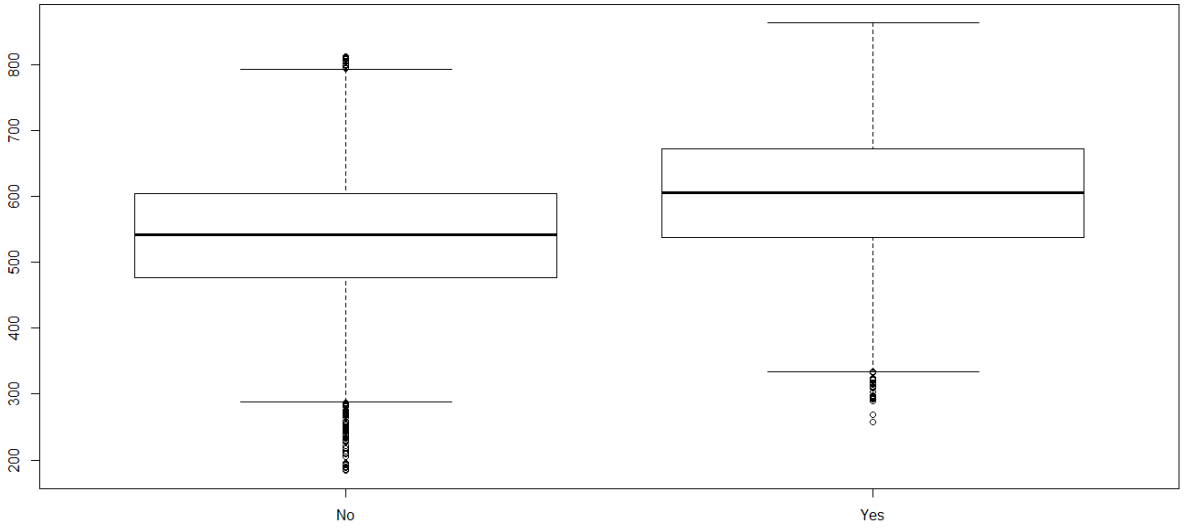
plot(cust_billdata$Num_Citylived, cust_billdata$AvgBill)



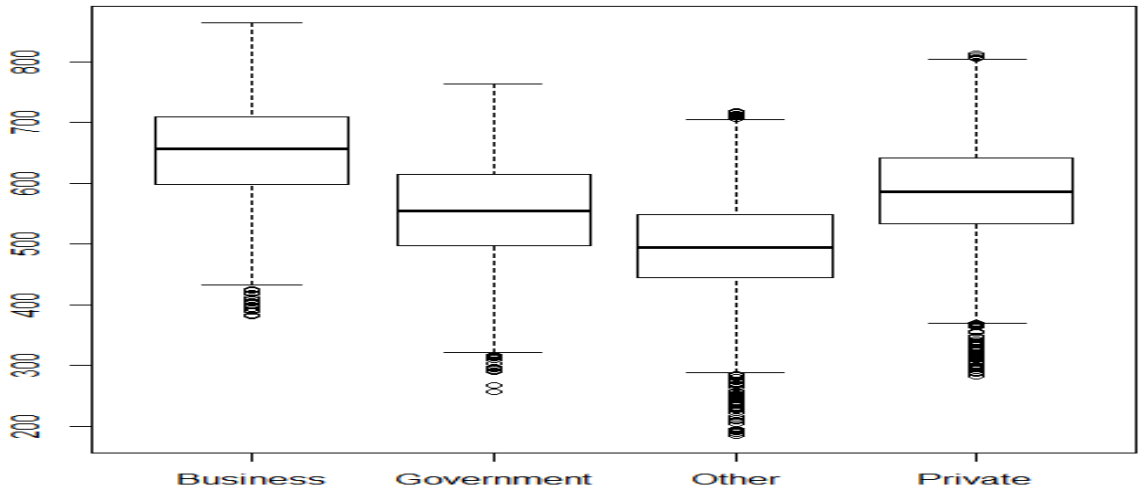
plot(cust_billdata$ClsRelativesCnt, cust_billdata$AvgBill)



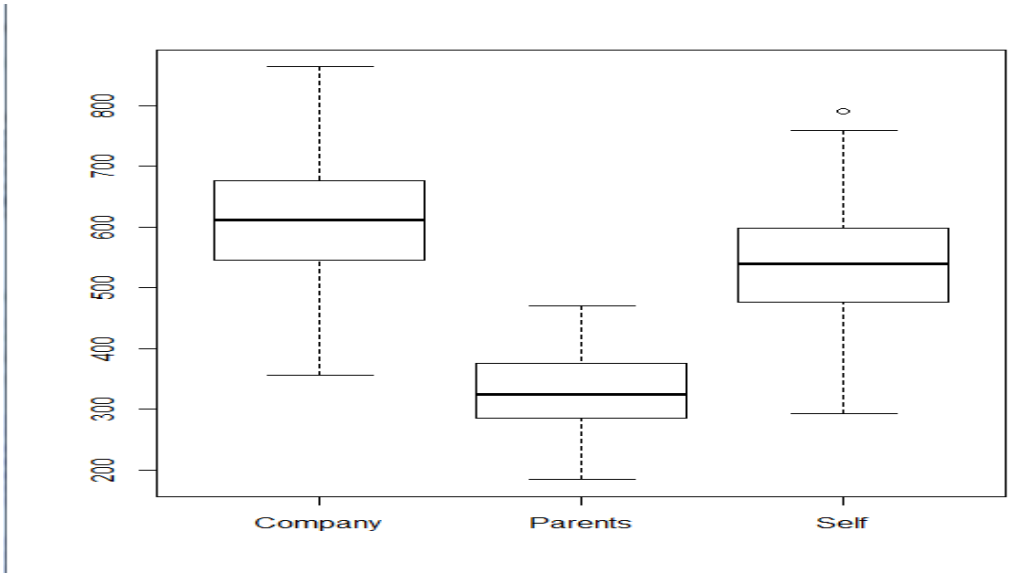
plot(cust_billdata$RelativesAbroad, cust_billdata$AvgBill)



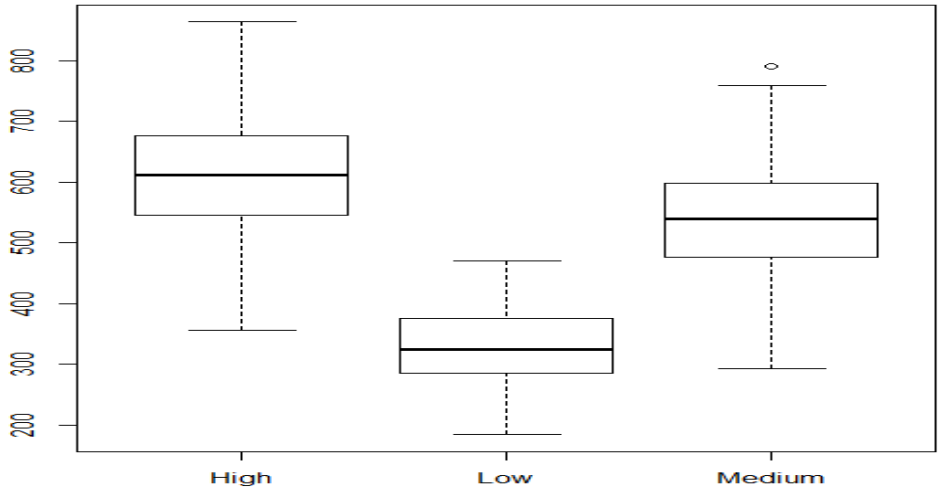
plot(cust_billdata$jobtype, cust_billdata$AvgBill)



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)

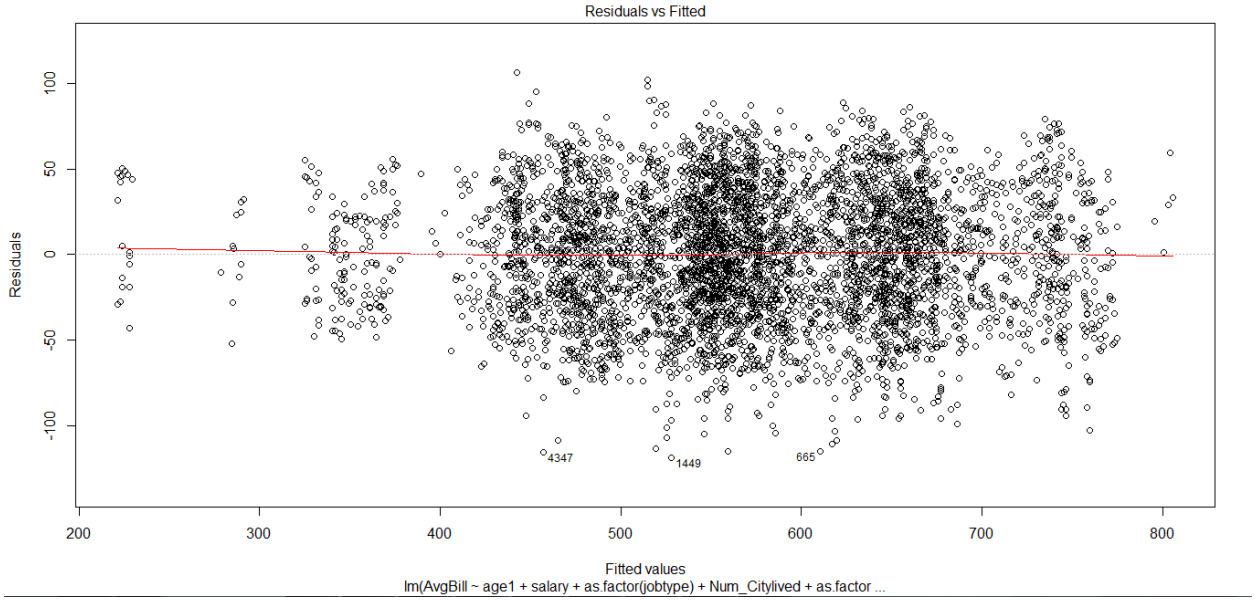
```
> summary(fit)

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)

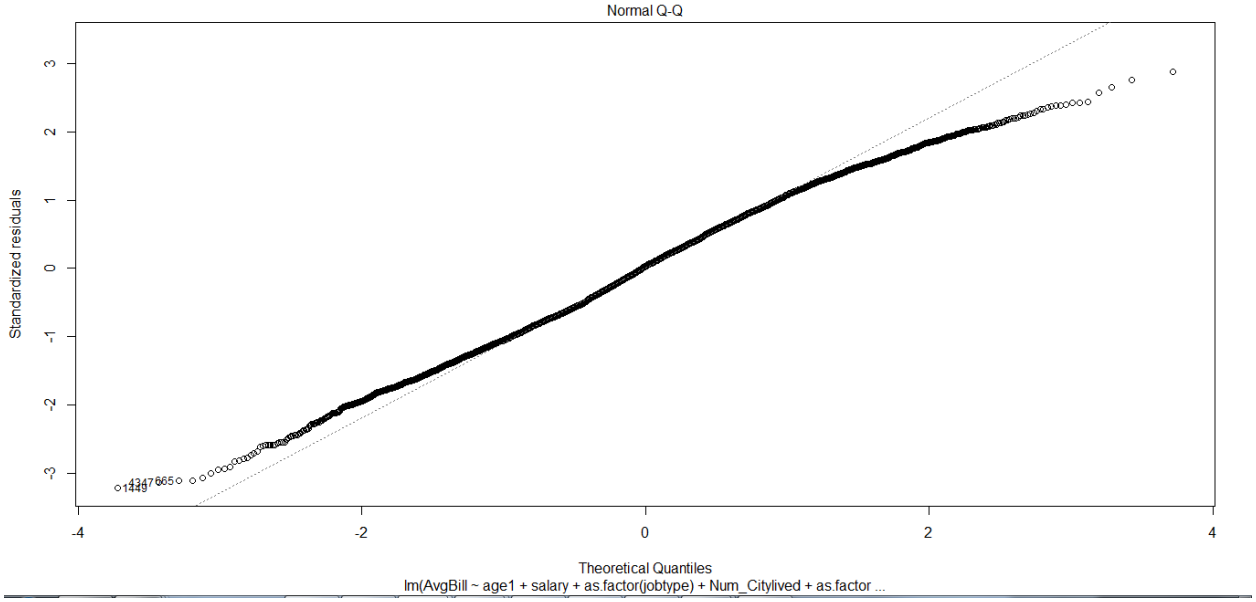
###Diagnostic plots:

Residual Vs Fitted values



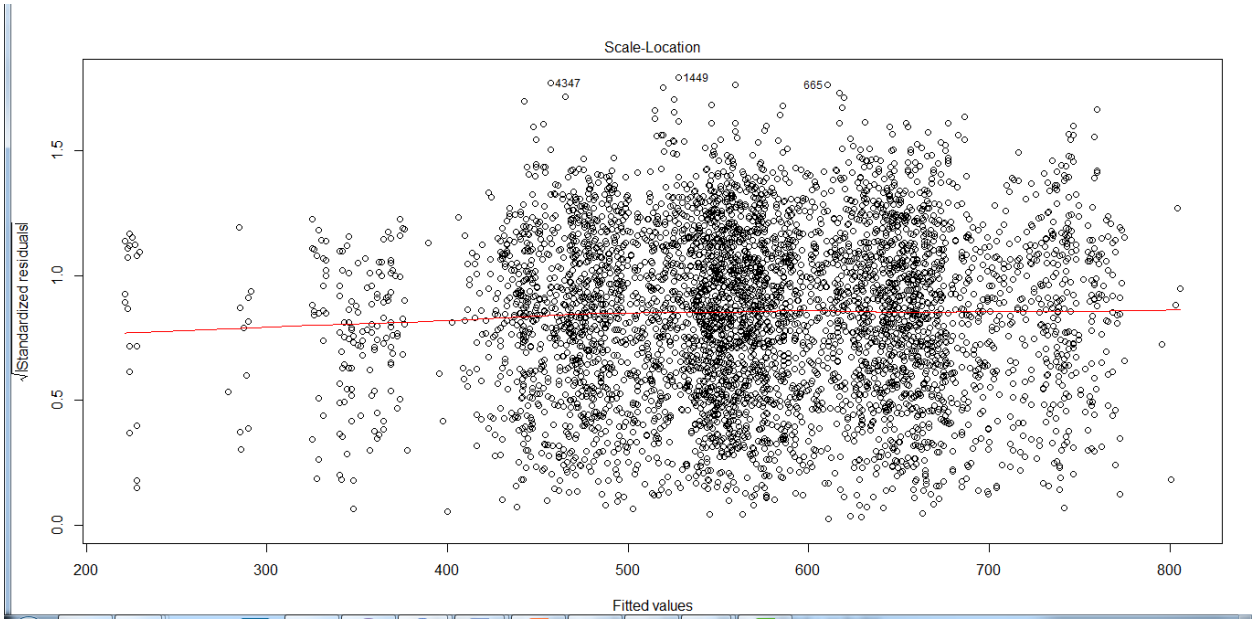
The plot shows no pattern in the residuals hence validates the assumptions of residual independence.

2) Normality:Q-Q plot



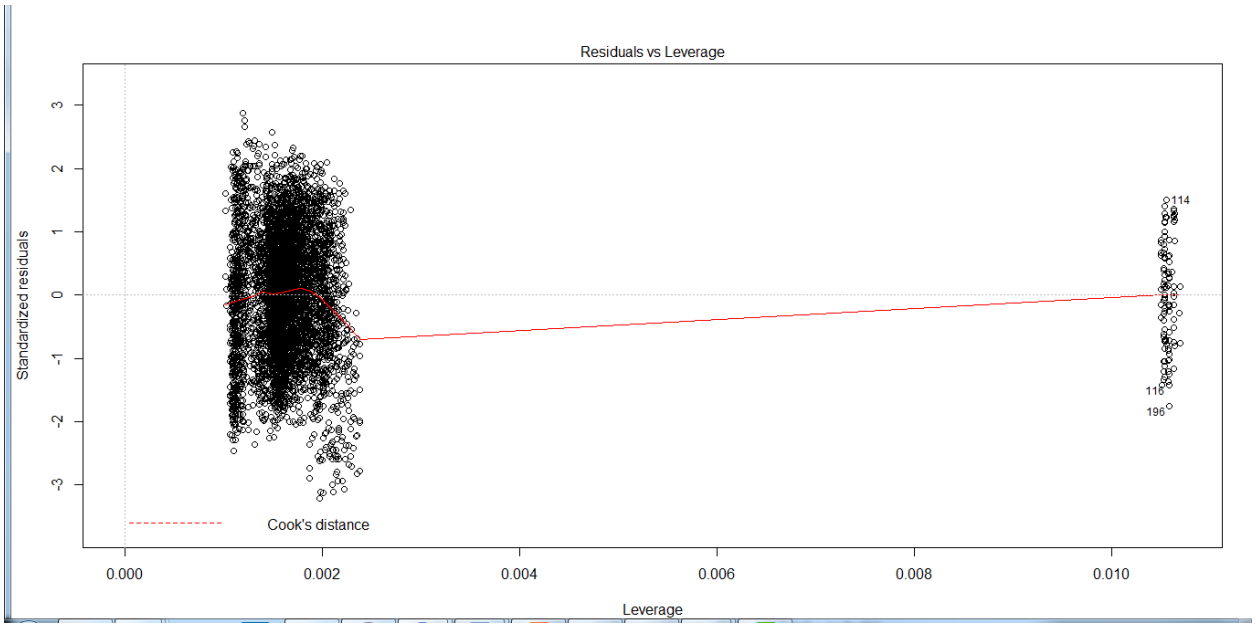
The line lies approximately on the normality curve hence validates the assumptions of normality.

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) and/or high leverage 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 R^2 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.

Variables	age1	salary	Num_Citylived	dv_jt_other	dv_jt_govt	dv_jt_biz	dv_jt_pvt	dv_payer_parents	dv_payer_self	dv_payer_company	ClsRelativesCnt	dv_travel_low	dv_travel_medium	dv_travel_high	dv_Rel_yes	dv_Rel_No
age1	1	0.164970167	0.044994579	-0.013974237	0.023152189	0.007843473	-0.016479011	-0.283223954	0.051949202	0.029120975	0.005036394	-0.283223954	0.051949202	0.029120975	0.026616806	-0.026616806
salary	0.164970167	1	0.212341589	-0.044129589	0.053402067	0.042849409	-0.050410127	-0.355799683	0.066103447	0.035741491	0.030417779	-0.355799683	0.066103447	0.035741491	0.082629368	-0.082629368
Num_Citylived	0.044994579	0.212341589	1	0.057308416	0.065906152	-0.011591008	-0.110659111	-0.048800621	-0.001639867	0.015600479	-0.074578138	-0.048800621	-0.001639867	0.015600479	0.046675238	-0.046675238
dv_jt_other	-0.013974237	-0.044129589	0.057308416	1	-0.333504511	-0.332509538	-0.34544217	0.01663441	0.00356307	-0.008319451	-0.781178865	0.01663441	0.00356307	-0.008319451	-0.571708167	0.571708167
dv_jt_govt	0.023152189	0.053402067	0.065906152	-0.333504511	1	-0.321439633	-0.333941712	-0.014578235	0.00223045	0.001942092	-0.246946637	-0.014578235	0.00223045	0.001942092	0.583347466	-0.583347466
dv_jt_biz	0.007843473	0.042849409	-0.011591008	-0.332509538	-0.321439633	1	-0.332945435	-0.006072802	-0.022377406	0.024097712	0.767208089	-0.006072802	-0.022377406	0.024097712	0.581607116	-0.581607116
dv_jt_pvt	-0.016479011	-0.050410127	-0.110659111	-0.34544217	-0.333941712	-0.332945435	1	0.00365863	0.016201774	-0.017236111	0.27063095	0.00365863	0.016201774	-0.017236111	-0.572457636	0.572457636
dv_payer_parents	-0.283223954	-0.355799683	-0.048800621	0.01663441	-0.014578235	-0.006072802	0.00365863	1	-0.140483897	-0.145723787	-0.009731684	1	-0.140483897	-0.145723787	-0.017732483	0.017732483
dv_payer_self	0.051949202	0.066103447	-0.001639867	0.00356307	0.00223045	-0.022377406	0.016201774	-0.140483897	1	-0.959042294	-0.01222956	-0.140483897	1	-0.959042294	-0.017277974	0.017277974
dv_payer_company	0.029120975	0.035741491	0.015600479	-0.008319451	0.001942092	0.024097712	-0.017236111	-0.145723787	-0.959042294	1	0.015004438	-0.145723787	-0.959042294	1	0.022338019	-0.022338019
ClsRelativesCnt	0.005036394	0.030417779	-0.074578138	-0.781178865	-0.246946637	0.767208089	0.27063095	-0.009731684	-0.01222956	0.015004438	1	-0.009731684	-0.01222956	0.015004438	0.445925426	-0.445925426
dv_travel_low	-0.283223954	-0.355799683	-0.048800621	0.01663441	-0.014578235	-0.006072802	0.00365863	1	-0.140483897	-0.145723787	-0.009731684	1	-0.140483897	-0.145723787	-0.017732483	0.017732483
dv_travel_medium	0.051949202	0.066103447	-0.001639867	0.00356307	0.00223045	-0.022377406	0.016201774	-0.140483897	1	-0.959042294	-0.01222956	-0.140483897	1	-0.959042294	-0.017277974	0.017277974
dv_travel_high	0.029120975	0.035741491	0.015600479	-0.008319451	0.001942092	0.024097712	-0.017236111	-0.145723787	-0.959042294	1	0.015004438	-0.145723787	-0.959042294	1	0.022338019	-0.022338019
dv_Rel_yes	0.026616806	0.082629368	0.046675238	-0.571708167	0.583347466	0.581607116	-0.572457636	-0.017732483	-0.017277974	0.022338019	0.445925426	-0.017732483	-0.017277974	0.022338019	1	-1
dv_Rel_No	-0.026616806	-0.082629368	-0.046675238	0.571708167	-0.583347466	-0.581607116	0.572457636	0.017732483	0.017277974	-0.022338019	-0.445925426	0.017732483	0.017277974	-0.022338019	-1	1

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)
```