

THE IVY PROFESSIONAL SCHOOL

AN ANALYSIS ON

“*CUSTOMER LIFETIME VALUE*“

OF AN AUTO INSURANCE COMPANY

PROJECT REPORT

*Submitted by*

NAME OF THE CANDIDATE

AKASH MEHTA

SUBMISSION DATE

APRIL 2018

CONTENTS : Page No.

1. Multiple Linear Regression…………………………………… 3

2. Background…………………………………………………… 3

3. The Task………………………………………………………. 3

4. Setting Up the R Model by loading the required libraries……. 3

5. The Data………………………………………………………. 4

6. Prepare For Modelling………………………………………... 6

6.1. Conversion into Factor…………………………………… 7

6.2. Outliers Removal………………………………………… 8

6.3. Treating Missing Values…………………………………. 11

6.4. Dataset Division………………………………………….. 12

7. Build Linear Model……………………………………………. 13

8. Linear Regression Diagnostics………………………………… 14

8.1. The P-Value: Checking for statistical significance………. 15

9. Linear Model Analysis…………………………………………. 16

9.1. Residuals………………………………………………….. 17

9.2. Estimate Coefficient……………………………………… 17

9.3. Standard Error…………………………………………….. 17

9.4. R-Square and Adj R-Square………………………………. 17

10. How to know if the model is best fit for your data…………… 18

11. Predicting Linear Model……………………………………… 19

11.1. MAPE…………………………………………………… 20

12. Assumptions of Linear Regression……................................... 21

13. Regression Analysis Output………………………………….. 24

13.1. Multiple R………………………………………………. 24

13.2. R Squared……………………………………………….. 24

13.3. Adj R Square……………………………………………. 24

13.4. MAPE…………………………………………………… 25

13.5. Assumptions…………………………………………….. 25

14. The Significant Variables and their significance…………….. 28

15. Business Interpretation……………………………………….. 30

# Multiple Linear Regression

**Key ideas**: Outlier Identification, Missing Value Removal, Splitting Dataset into Development and Validation, Stepwise regression on development dataset, Regression assumptions, residuals, Interpreting model coefficients, Prediction on validation.

1. **Background**

An auto insurance company wants to know how CLV is being affected by all the factors like Monthly\_Premium\_Auto, Total\_Claim\_Amount, Sales\_Channel, etc. And wants to predict future CLV for those customers.

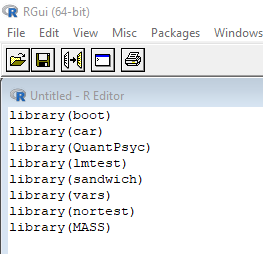
1. **The Task**

We want to build a model that allows the insurance company to predict customer lifetime value for all given customer. In our dataset,

we have customer lifetime value which is “*the present value of the future cash flows attributed to the customer during his/her entire relationship with the company.”*

The resulting model will be used to forecast CLV and guide the insurance in future marketing campaigns.

1. **Setting up the R model by loading the required libraries:**

****

1. **The Data Insurance\_Marketing-Customer-Value-Analysis.xlsx**

The data set contains information on 9,134 auto insurance customers:

Customer These are the unique customer id.

State The states to which customers belong.

Customer\_Lifetime The customer’s lifetime value.

\_Value

Response Customer’s willingness to provide the firm

with constructive feedback and suggestion.

Coverage Insurance against accident which covers

costs of repairs, legal fees and medical

coverage.

Education Education qualification of insurance holder.

Effective\_To\_Date The date that a policy becomes active.

EmploymentStatus Whether the customer is employed or not.

Gender Gender of customers.

Income Income of customers.

Location\_Code The part of the state the customers belongs.

Marital\_Status Whether the customer is married, single or

divorced.

Monthly\_Premium Premium deposited by customers on a

\_Auto monthly basis.

Months\_Since Number of months passed out since last

\_Last\_Claim claim.

Months\_Since Number of months after the start date of

\_Policy\_Inception policy.

Number\_of Complaints against claim.

\_Open\_Complaints

Number\_of\_Policies Number of policies manage by a customer.

Policy\_Type Different types of auto insurance policy.

Policy Sub-Policies within policies.

Renew\_Offer\_Type Different offers for policy renewal.

Sales\_Channel Medium through which the policies had

been sold to customers.

Total\_Claim\_Amount Total amount of benefit taken from policies.

Vehicle\_Class Different types of vehicles based on its

properties.

Vehicle\_Size Size of vehicles.

1. **Prepare for Modelling**

We begin by finding the structure of the given dataset. Here, we use *str()* function .We check if there are any variables whose data-type needs conversion into another i.e., we basically convert *char* data-type which are categorical variables into a *factor* data-type.

The code str(data) will result in following output:- ****

Now, we need to convert those variables into a factor data-type.

* 1. **Conversion into Factor**

We use *as.factor* () function to implement the desire conversion.

Code :-

#converting Variables into factor

data$State <- as.factor(data$State)

data$Response <- as.factor(data$Response)

data$Coverage <- as.factor(data$Coverage)

data$Education <- as.factor(data$Education)

data$EmploymentStatus <- as.factor(data$EmploymentStatus)

data$Gender <- as.factor(data$Gender)

data$Location\_Code <- as.factor(data$Location\_Code)

data$Marital\_Status <- as.factor(data$Marital\_Status)

data$Number\_of\_Open\_Complaints <- as.factor(data$Number\_of\_Open\_Complaints)

data$Number\_of\_Policies <- as.factor(data$Number\_of\_Policies)

data$Policy\_Type <- as.factor(data$Policy\_Type)

data$Policy <- as.factor(data$Policy)

data$Renew\_Offer\_Type <- as.factor(data$Renew\_Offer\_Type)

data$Sales\_Channel <- as.factor(data$Sales\_Channel)

data$Vehicle\_Class <- as.factor(data$Vehicle\_Class)

data$Vehicle\_Size <- as.factor(data$Vehicle\_Size)

Note: - We don’t use Customer ids in our model creation because it is an insignificant variable.

* 1. **Outliers Removal**

Here, we are concern with the extreme values which may hamper our

model and prediction .So, we must remove those values before

proceeding towards model creation.

To do so we use the concept of *boxplot*() function, which highlights

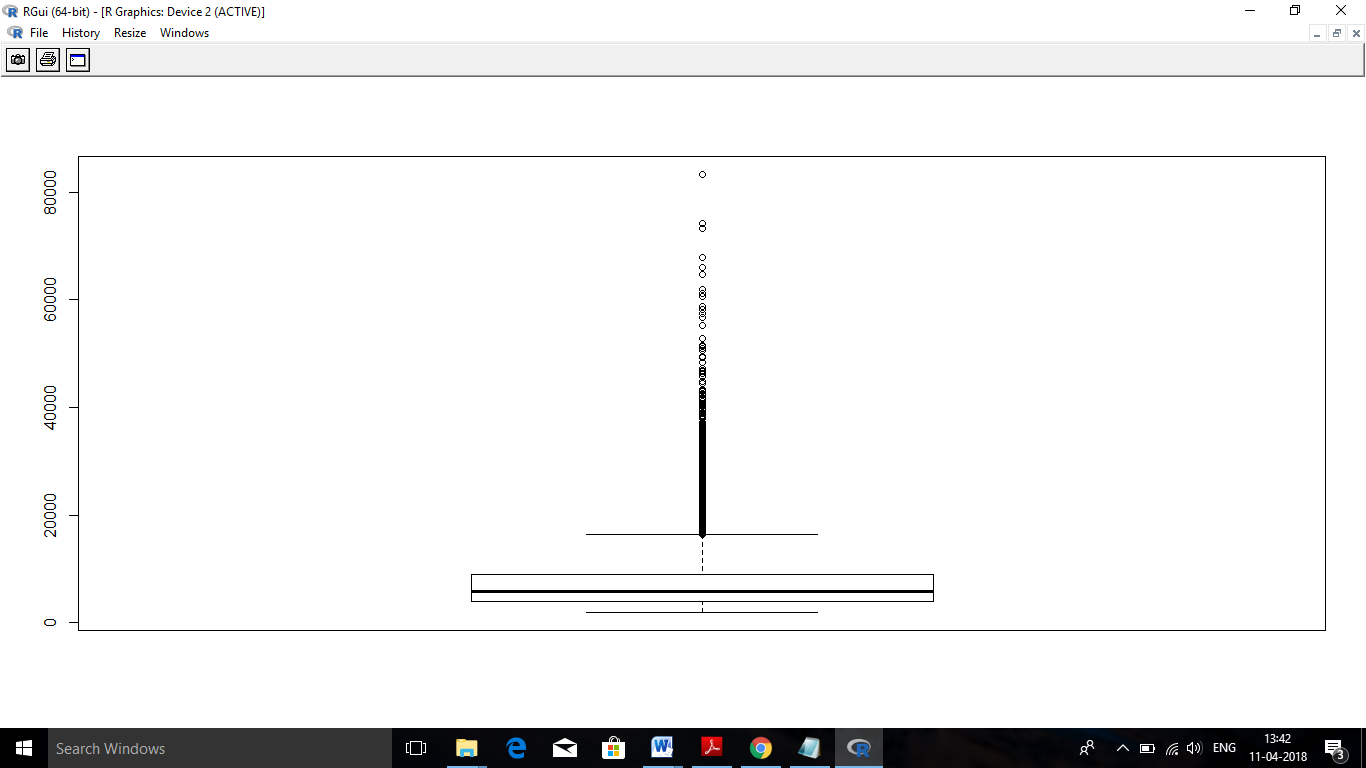
the outliers. And we also use *quantile*() function to get the extreme

values. And then we filter the variables one by one.

Code:-

**Boxplot**

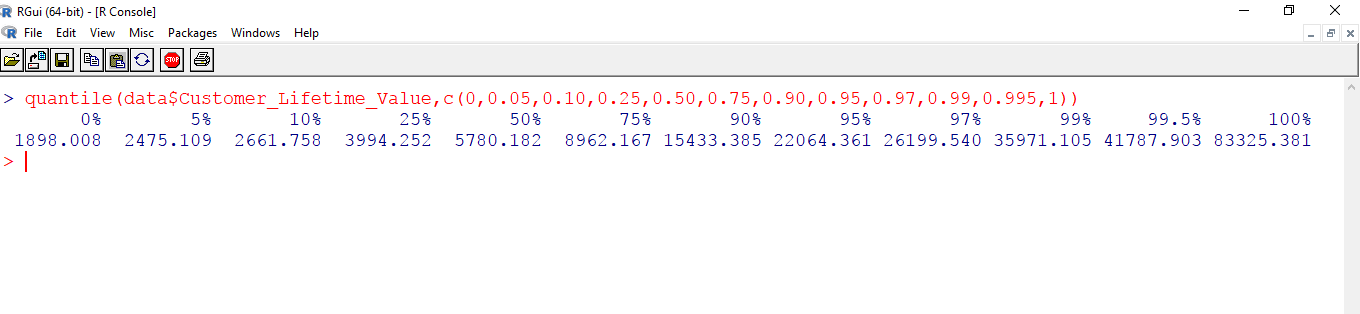
boxplot (data$Customer\_Lifetime\_Value)



Note:- Black dots are outliers.

**Quantiles**

quantile(data$Customer\_Lifetime\_Value,c(0,0.05,0.10,0.25,0.50,0.75,0.90,0.95,0.97,0.99,0.995,1))



These are quantiles from 0% to 100% along with values.

data <- data[data$Customer\_Lifetime\_Value <41787.903,]

By using the above code we remove the outliers.

And we will repeat these steps in sequence until the outliers are being removed from the data.

boxplot (data$Customer\_Lifetime\_Value)

quantile(data$Customer\_Lifetime\_Value,c(0,0.05,0.10,0.25,0.50,0.75,0.90,0.95,0.97,0.99,0.995,1))

data <- data[data$Customer\_Lifetime\_Value <35857.788,]

boxplot(data$Customer\_Lifetime\_Value)

quantile(data$Customer\_Lifetime\_Value,c(0,0.05,0.10,0.25,0.50,0.75,0.90,0.95,0.97,0.99,0.995,1))

data <- data[data$Customer\_Lifetime\_Value <15002.559,]

boxplot(data$Customer\_Lifetime\_Value)

boxplot(data$Income)

boxplot(data$Monthly\_Premium\_Auto)

quantile(data$Monthly\_Premium\_Auto,c(0,0.05,0.10,0.25,0.50,0.75,0.90,0.95,0.97,0.99,0.995,1))

data <- data[data$Monthly\_Premium\_Auto <226,]

boxplot(data$Monthly\_Premium\_Auto)

quantile(data$Monthly\_Premium\_Auto,c(0,0.05,0.10,0.25,0.50,0.75,0.90,0.95,0.97,0.99,0.995,1))

data <- data[data$Monthly\_Premium\_Auto <170,]

boxplot(data$Monthly\_Premium\_Auto)

boxplot(data$Months\_Since\_Last\_Claim)

boxplot(data$Months\_Since\_Policy\_Inception)

boxplot (data$Effective\_To\_Date)

boxplot(data$Total\_Claim\_Amount)

quantile(data$Total\_Claim\_Amount,c(0,0.05,0.10,0.25,0.50,0.75,0.90,0.95,0.97,0.99,0.995,1))

data <- data[data$Total\_Claim\_Amount <1232,]

boxplot(data$Total\_Claim\_Amount)

quantile(data$Total\_Claim\_Amount,c(0,0.05,0.10,0.25,0.50,0.75,0.90,0.95,0.97,0.99,0.995,1))

data <- data[data$Total\_Claim\_Amount <1140,]

boxplot(data$Total\_Claim\_Amount)

quantile(data$Total\_Claim\_Amount,c(0,0.05,0.10,0.25,0.50,0.75,0.90,0.95,0.97,0.99,0.995,1))

data <- data[data$Total\_Claim\_Amount <930,]

boxplot(data$Total\_Claim\_Amount)

* 1. **Treating Missing Values**

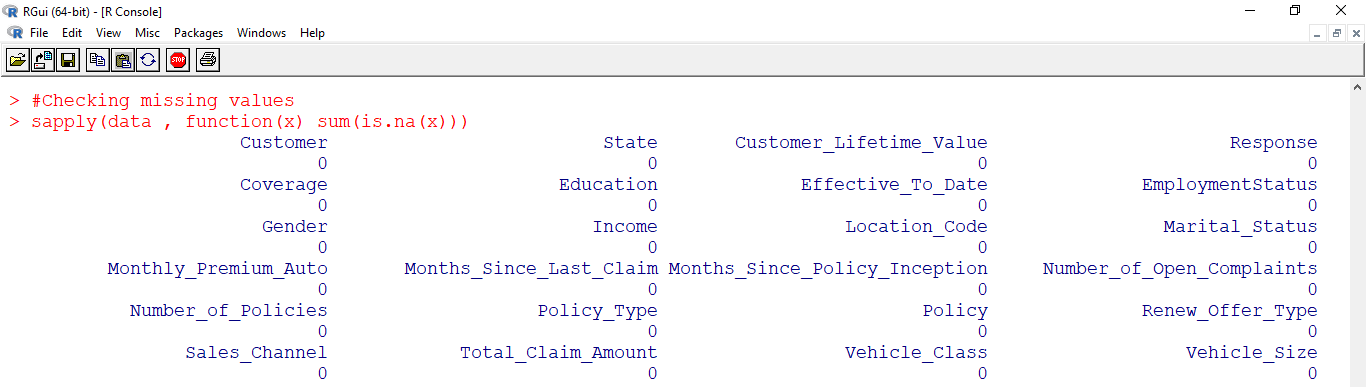
After removing the outliers we would find out if there are any

missing values in the deduced dataset.

The code below finds out missing value present in the dataset:

#Checking missing values

*sapply(data , function(x) sum (is.na(x)))*



And the following code will remove all the missing values.

#Removing Missing values

*data <- na.omit(data)*

* 1. **Dataset Division**

This is the final step before we begin to create our model.

In order to test the efficiency of the model created we divide our

dataset into two parts: Development and Validation.

* The development part is created by taking 80% of data from

the original dataset. We create our model using development

dataset.

* The remaining 20% of data is used to create validation dataset.

This dataset is used for prediction .

We have used *sample()* functionto create following dataset.

*set.seed()* function is used to randomly set a starting value for sample

function.

The following code will perform the dataset division process:

#Dividing my dataset into two set: - Development and Validation

set.seed (2)

train <- sample(1:nrow(data), floor(0.80\*nrow(data)))

test <- -train

development <- data[train,]

validation <- data[test,]

The below code checks the percentage of data in each dataset:

(nrow(development)/nrow(data))\*100

(nrow(validation)/nrow(data))\*100

## Build Linear Model

Objective now is building a linear model and sees how well this

model fits the observed data. Let’s see the syntax for building the

linear model. The function used for building linear models is *lm*().

The *lm*() function takes in two main arguments, namely:

1) Formula 2) Data . The data is typically a *data.frame* and the

formula is an object of class formula. But the most common

convention is to write out the formula directly in place of the

argument as written below.

#Creating Model

fit <- lm(Customer\_Lifetime\_Value ~ State +Response +Coverage

+Education +Effective\_To\_Date +EmploymentStatus +Gender

+Income +Location\_Code +Marital\_Status

+Monthly\_Premium\_Auto +Months\_Since\_Last\_Claim

+Months\_Since\_Policy\_Inception

+Number\_of\_Open\_Complaints +Number\_of\_Policies

+Policy\_Type +Policy +Renew\_Offer\_Type +Sales\_Channel

+Total\_Claim\_Amount +Vehicle\_Class +Vehicle\_Size

, data=development)

Here, predictors are State, Response, Coverage, Education, etc.

response variable is Customer\_Lifetime\_Value.

## Linear Regression Diagnostics

To make my model more robust, we will try eliminating all those

predictor variables which will be less significant in my model.

Is this enough to actually use this model?

NO! Before using a regression model, you have to ensure that it is

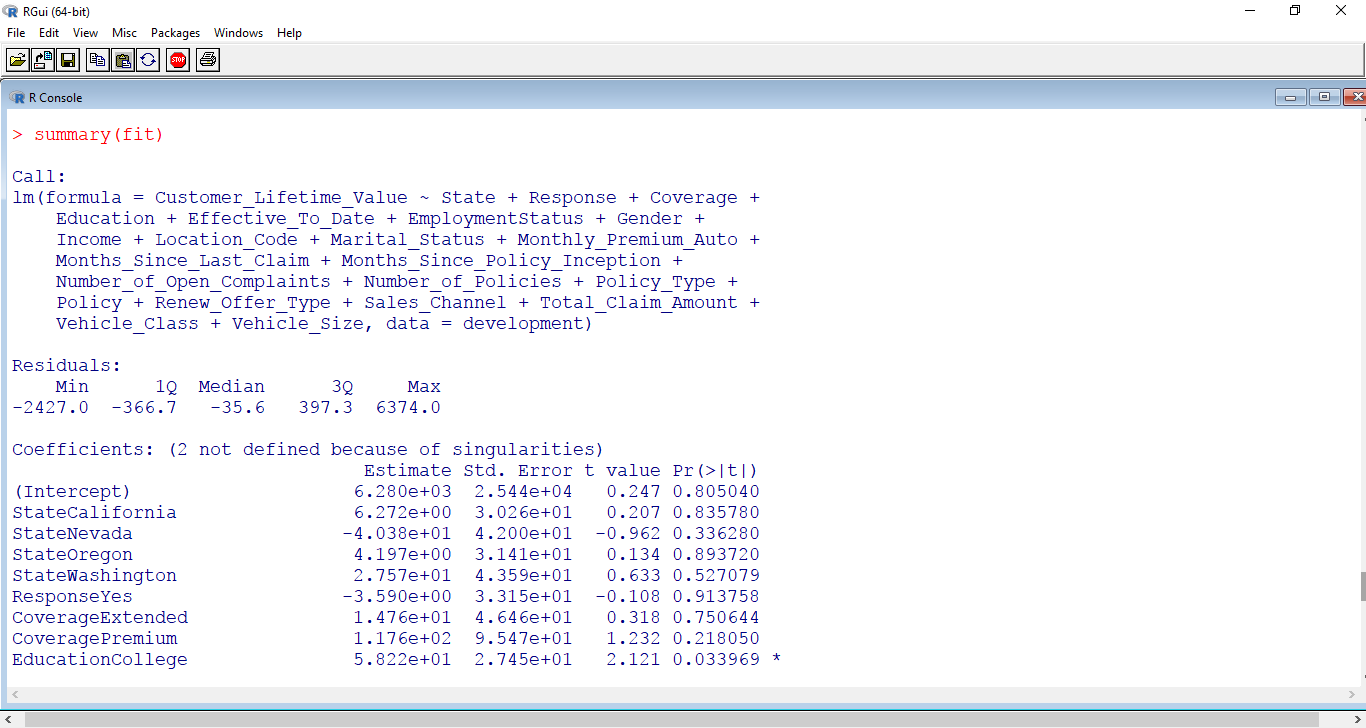
statistically significant.

How do you ensure this? Let’s begin by printing the summary

statistics for *fit*(my model name).

#To print all the descriptive statistics for all predictor variable

summary(fit)



## The p Value: Checking for statistical significance

The summary statistics above tells us a number of things. One of them is the model p-Value (bottom last line) and the p-Value of individual predictor variables (extreme right column under ‘Coefficients’). The p-Values are very important because, we can consider a linear model to be statistically significant only when both these p-Values are less that the pre-determined statistical significance level, which is ideally 0.05. This is visually interpreted by the significance stars at the end of the row. *The more the stars beside the variable’s p-Value, the more significant the variable.*

The less significant variables are removed from the model one by one.

And finally we would have the final linear model with all predictor variables, with stars besides them.

The following is the final linear model after removing insignificant variables.

fit <- lm(Customer\_Lifetime\_Value ~ Education

+I(EmploymentStatus=="Unemployed")

+Gender +Income +Marital\_Status +Monthly\_Premium\_Auto

+Months\_Since\_Last\_Claim +Months\_Since\_Policy\_Inception

+Number\_of\_Open\_Complaints +Number\_of\_Policies

+I(Policy\_Type=="Personal Auto")

+I(Policy=="Corporate L3")

+I(Renew\_Offer\_Type=="Offer2"),data=development)

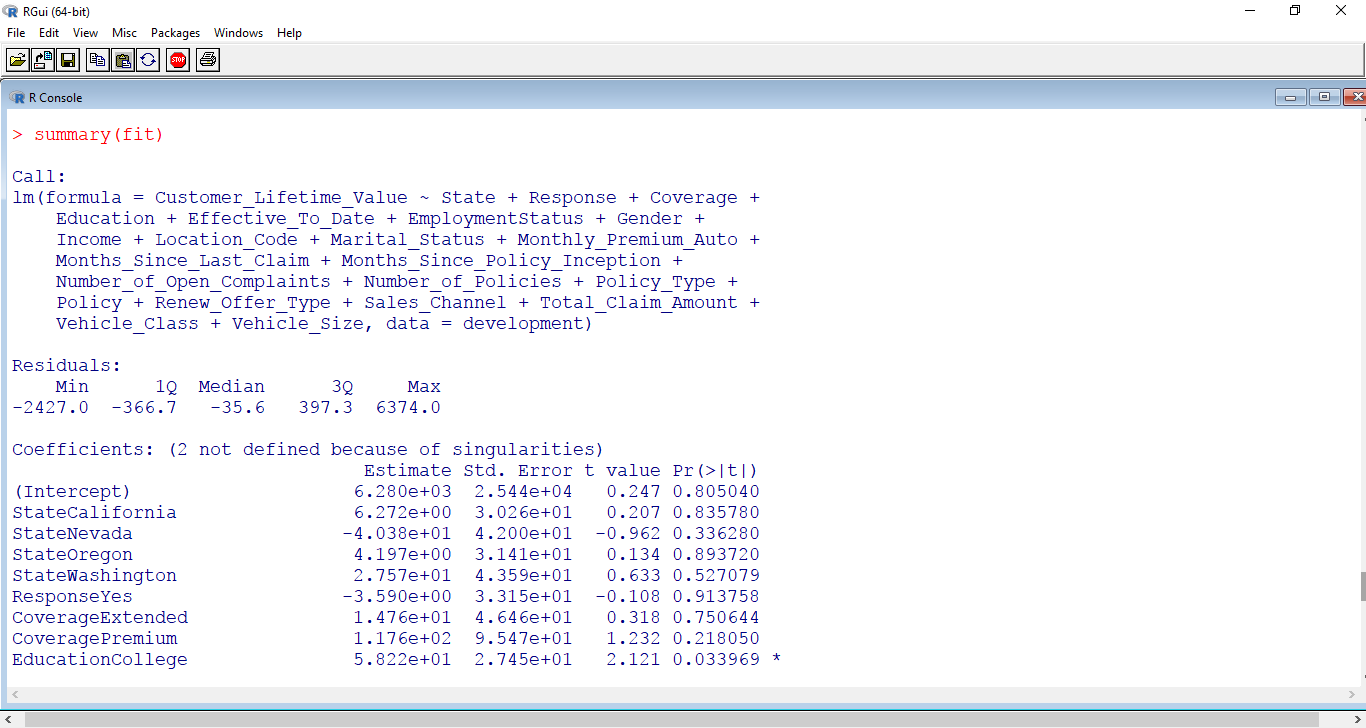
Note: - We have removed predictor variables: Response, State, Coverage, Effective\_ To\_Date , EmploymentStatus, Policy, Vehicle\_Class, Policy\_Type, Renew\_Offer\_Type, Sales\_Channel,

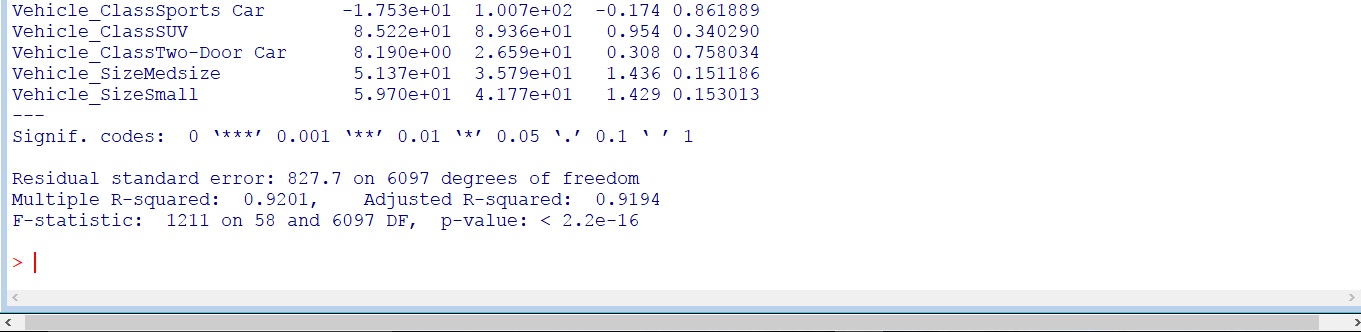
Total\_Claim\_Amount, Location\_Code, Vehicle\_Size,

I(Policy=="Corporate L2"), I(EmploymentStatus=="Employed")

## Linear model analysis

Summary statistics are very useful to interpret the key components of the linear model output.





* 1. **Residuals**

Residuals show if the predicted response values are close or not to

the response values that the model predicts.

* 1. **Estimate coefficient**

The first row is the intercept and represents in the current case the

expected Customer\_Lifetime\_Value value of -2.203e+03 percent

when the average value of all predictors are considered across the

dataset.

* 1. **Standard error**

Standard error measures how the coefficient estimates can vary

from the actual average value of the response variable (i.e. if the

model is run more times).

Note: ideally, having a low number is the best situation in

regression analysis. Standard error is then also used to set

confidence intervals.

## R-Squared and Adj R-Squared

For the simple linear regression, R-squared is the square of the

correlation between two variables. Its value can vary between 0 and

1: *a value close to 0 means that the regression model does not*

*explain the variance in the response variable, while a number close*

*to 1 that the observed variance in the response variable is well*

*explained.*

High value of R-squared does not necessarily indicate if a regression

model provides an adequate fit to data. A good model could show a

low R-squared value, while, on the other hand, a biased model could

have a high R-squared value.

Note: R-squared value tends to increase as more variables are included in the model. So, adjusted R-squared is the preferred measure as it adjusts for the number of variables considered .

## How to know if the model is best fit for your data?

The most common metrics to look at while selecting the model are:

| **STATISTIC** | **CRITERION** |
| --- | --- |
| R-Squared | Higher the better *(> 0.70)* |
| Adj R-Squared | Higher the better |
| F-Statistic | Higher the better |
| Std. Error | Closer to zero the better |
| t-statistic | Should be greater 1.96 for p-value to be less than 0.05 |
| AIC | Lower the better |
| BIC | Lower the better |
| Mallows cp | Should be close to the number of predictors in model |
| MAPE (Mean absolute percentage error) | Lower the better |
| MSE (Mean squared error) | Lower the better |
| Min\_Max Accuracy => mean(min(actual, predicted)/max(actual, predicted)) | Higher the better |

## Predicting Linear Models

Once the model has been improved, it is possible to

run predictive analytics, the real goal of the regression analysis.

Dataset can be split into **training** (development)

and **testing** (validation) dataset, and the test one will be used to

evaluate the model comparing the predicted response with the actual

response value.

As we have already divided our dataset into *development* and

*validation.* And had created our model using development dataset.

Now, we can predict our model using validation dataset.

## Get the predicted on validation

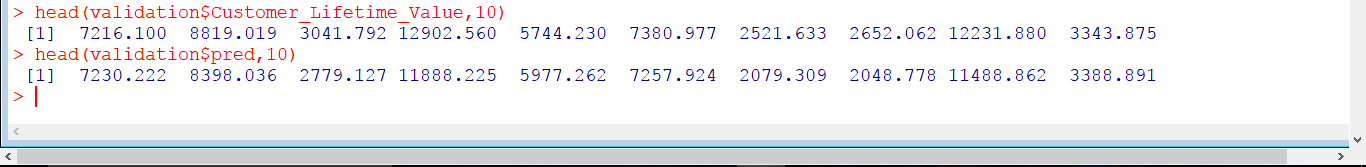
validPred <- predict(fit, validation)

validation$pred <- validPred

The following code will show my actual and predicted result:

head(validation$Customer\_Lifetime\_Value)

head(validation$pred)



* 1. **MAPE**

**MAPE** means *Mean Absolute Percentage Error*.

It help us to find how different my prediction are from the actual

value. It ranges from 0 to 1.

MAPE tells us the average percentage deviation of predicted value

from the actual values.

*Lower the MAPE better the model.*

The below code is used to find out the MAPE value:

#Calculating MAPE on Validation

attach(validation)

(sum((abs(Customer\_Lifetime\_Value- pred))

/Customer\_Lifetime\_Value))/nrow(validation)

# Assumptions of Linear Regression

Building a linear regression model is only half of the work. In order

to actually be usable in practice, the model should conform to the

assumptions of linear regression.

## Assumption 1

#### *No perfect multicollinearity between predictor variables.*

Assumption says that there should be zero correlation among all the

predictor variables.

***How to check?***

Using Variance Inflation factor (VIF).

VIF is a metric computed for every predictor variable that goes into a linear model. If the VIF of a variable is high, it means the information in that variable is already explained by other predictor variables present in the given model, which means, more redundant is that variable. So, lower the VIF (<2) the better. VIF for a predictor variable calculated as:

VIF=1/(1−Rsq)

where, Rsq is the Rsq term for the model with given X as response against all other predictors that went into the model as predictors.

Practically, if two of the predictor′s have high correlation, they will likely have high VIFs. *Generally, VIF for an*predictor *variable should be less than 4 in order to be accepted as not causing multi-*

*collinearity. The cut off is kept as low as 2, if you want to be strict about your*predictor*variables.*

# Checking multicollinearity

*vif(fit)*

*# should be within 2. If it is greater than 10 then serious problem*

#### How to rectify?

Two ways:

1. Either iteratively remove the predictor var with the highest VIF
2. See correlation between all variables and keep only one of all

highly correlated pairs.

* **Assumption 2**

#### *Normality of residuals*

The residuals should be normally distributed. If the maximum

likelihood method (not OLS) is used to compute the estimates, this

also implies the Response and the Predictors are also normally

distributed.

This can be checked using *Anderson-Darling* Test.

resids <- fit$residuals

ad.test(resids)

## Normality testing Null hypothesis is data is normal.

#get Anderson-Darling test for normality

* **Assumption 3**

Homoscedasticity of residuals or equal variance

Homoscedasticity describes a condition where residuals are same

across all the predictor variables.

In other words, this assumption means that the **variance** around

the **regression line** is the same for all values of the predictor

variable (X).

This can be checked using *Breusch-Pagen* Test.

bptest(fit)

# Breusch-Pagan test

# Null hypothesis -> error is homogeneous (p value should be more

than 0.05)

* **Assumption 4**

#### *No autocorrelation of residuals*

Its says that one residual value should not be correlated to another

residual value i.e., there shouldn’t be a serial correlation among

residuals.

This can be checked using *Darwin-Watson* Test.

dwt(fit)

# Null H0: residuals from a linear regression are uncorrelated.

value should be close to 2.

#Less than 1 and greater than 3 -> concern

#Should get a high p value

## REGRESSION ANALYSIS OUTPUT

These are the “*Goodness of Fit*” measures. They tell you how well the calculated linear regression equation fits your data.

* 1. **Multiple R**

This is the *correlation coefficient*. It tells you how strong the linear relationship is. For example, a value of 1 means a perfect positive relationship and a value of zero means no relationship at all. It is the square root of r squared.

* 1. **R squared**

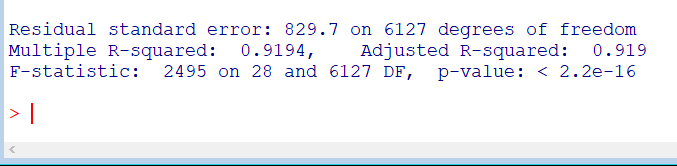
This is r2, the [Coefficient of Determination](http://www.statisticshowto.com/what-is-a-coefficient-of-determination/). It tells you how many points fall on the regression line. In Our Model, 91% means that 91% of the variations of y-values around the mean are explained by the x-values. In other words*, 91% of the values fit the model.*

* 1. **Adjusted R square**

The adjusted R-square adjusts for the number of terms in a model. You’ll want to use this instead of R squared if you have more than one

predictor variable. In our model, *the adjusted R square comes around*

*91.9%.*



So, we can conclude that my model is good enough to run and predict

on the dataset containing same predictors and response variable.

* 1. **MAPE:**

**MAPE** means *Mean Absolute Percentage Error*.

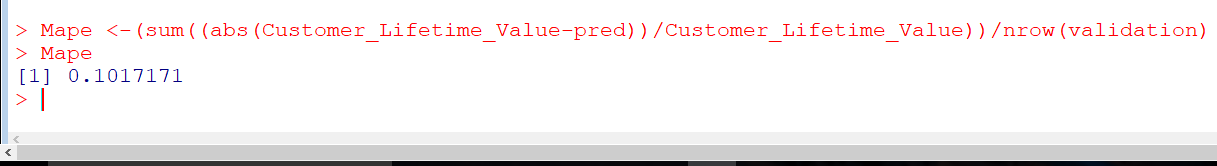
It help us to find how different my prediction are from the actual

value. It ranges from 0 to 1.

MAPE tells us the average percentage deviation of predicted value

from the actual values.

*Lower the MAPE better the model.*

****

Here, my MAPE is around 10% which very much acceptable.

* 1. **Assumptions :**
     1. Assumption 1

No perfect multicollinearity between predictor variables.

*Generally, VIF for an*predictor *variable should be less than 4 in*

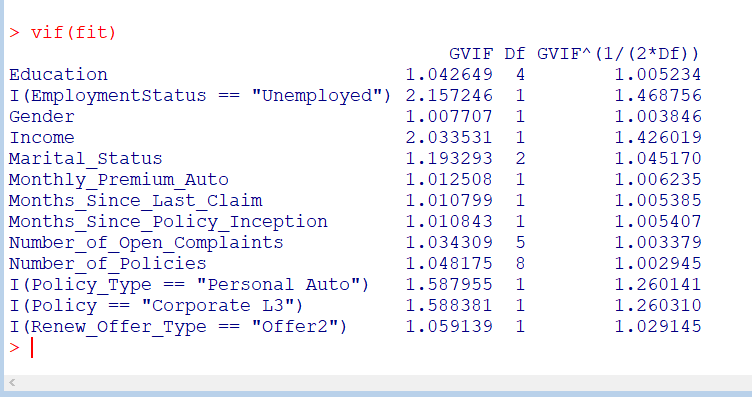
*order to be accepted as not causing multi-collinearity. The cut off*

*is kept as low as 2, if you want to be strict about your* predictor

*variables.*

In our model, *multicollinearity between independent variables was*

*negligible.*

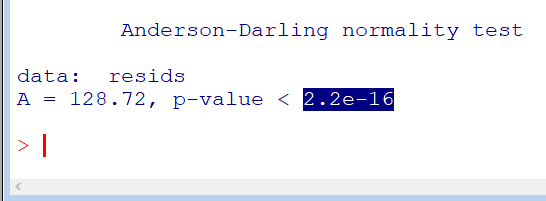
****

* + 1. **Assumption 2**

Normality of residuals.

H0: Data is Normally Distributed.

H1: Data is not Normally Distributed.



My p-value is very low .It means I can’t accept the null hypothesis.

To conclude, data is not normally distributed.

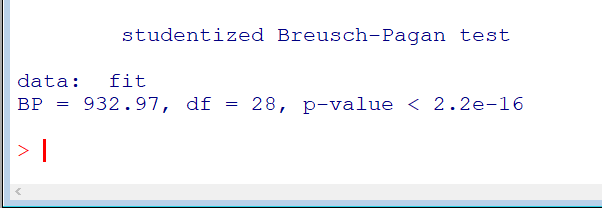
Hence, this assumption is violated.

* + 1. **Assumption 3**

Homoscedasticity of residuals or equal variance

H0: Data is Homoscedastic

H1: Data is Heteroscedastic

****

My p-value is very low .It means I can’t accept the null hypothesis.

To conclude, data is Heteroscedastic.

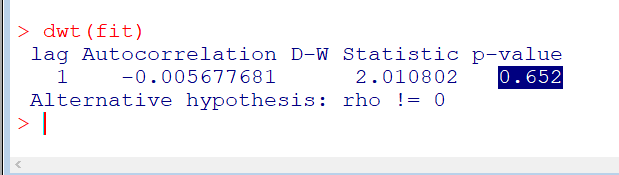
Hence, this assumption is violated.

* + 1. **Assumption 4**

No autocorrelation of residuals.

H0: Data is not serially correlated.

H1: Data is serially correlated.



My p-value is high .It means I can’t reject the null hypothesis.

To conclude, Data is not serially correlated.

Hence, this assumption is accepted.

Although two of my assumptions (Normality of residuals and Homoscedasticity of residuals or equal variance) are violated still

I can say that as I have removed multicollinearity so my model is good enough.

1. **The significant variables and their significance:**

These are the variables that are significant as per our

model and the image below shows us the relationship of these

variables with the dependent variable.

The following are the significant variables with a positive relationship with customer lifetime value:

* The customers who are well educated ,at least with a High school degree.(Variable:-Education)
* The customers with good earning i.e., their income is sufficient

to afford premium.(Variable:-Income)

* The customers who are married.(Variable:-Marital\_Status)
* Those who are paying a monthly auto premium(Variable:-

Monthly\_Premium\_Auto)

* Those who are having at least two insurance policies.(variable:- Number\_of\_Policies)

These are the significant variables with a negative relationship with

the dependent variable:

* The Customers who are unemployed.(Variable:- Employment\_

Status)

* The Customers who are male.(Variable:-Gender)
* The Customers who are single.(Variable:- Marital\_Status)
* The Customers who have not claim their insurance since months

ago. (Variable:-Months\_Since\_Last\_Claim)

* The Customers who provides open complaints regarding their

insurance.(Variable:-Number\_of\_Open\_Complaints)

* The Customers who have a Personal Auto policy type.

(Variable:-Policy\_Type)

* The Customers with Corporate L3 policy .(Variable:-Policy)
* The Customers with Renew offer type of Offer 2.

(Variable:-Renew\_Offer\_Type)

1. **Business Interpretation**

* **On the Basis of Education Qualification**

*The insurance company should focus more on the customers who*

*are at least a high school pass out*. Here, according to my model ,

which is considering Bachelor as a base . We compare the rest of

items with it.

So, a college degree holder’s CLV increases 5 units more in

comparison to a bachelor degree holder.

Similarly, a Doctorate’s CLV increases 2 units more in

comparison to a bachelor degree holder.

And ,a Master degree holder’s CLV increases 1 units more in

comparison to a bachelor degree holder.

Finally, a high school pass out’s CLV increases 6 units more in

comparison to a bachelor degree holder.

* **On the Basis of Income**

*Insurance company should target customers with a good income.*

According to my model, if there is an increment in income by 1 dollar then the CLV value increases by 4 units on an average.

* **On the Basis of Marital Status**

*Insurance company should pay more attention towards married*

*customers.* *On contrary, customers who are single have negative impact on the CLV.* Keeping them in mind , new policies should be brought into market.

As per my model, we are considering Divorced as base.

So, for every increment in number of married customers there is an increase in CLV by 7 units in comparison to a divorced one.

On the other hand, for every increment in number of customer who are single there is a decrement in CLV by 8 units in comparison to a divorced one.

* **On the Basis of Monthly Premium**

*Company should consider more those customers who are paying premium on monthly basis.*

As per my model, for every increment in number of customers

who are paying monthly premium there is an increase in CLV

by 6 units.

* **On the Basis of Number of Policies**

*Those customers ,who hold two or more number of policies are the one to stay for a long time with the company.* So, it’s company’s responsibility to provide them with better treatment

and more appealing plans.

To conclude, for every increment in number of customers with

2 policies there is an increment in CLV by 6 units and for

customers holding 3 to 9 policies there is an increment in CLV

by 3 units on an average in comparison to customer with only one

policy.

* **On the Basis of Employment Status**

*The company need not consider those customers who are unemployed.* According to my model, we are considering Disabled as our base and evaluating the rest in comparison to it.

To conclude, for an increment in number of customers who are

Unemployed there is a decrement in CLV by 2 units in comparison to a disabled customer.

* **On the Basis of Gender**

*When considering customers on the basis of gender they tends to have a negative relation with CLV.*

As per my model, taking Female as base we evaluate the rest.

We found that for every increment in male customer there is a decrement in CLV by 5 unit in comparison to a female customer.

* **On the Basis of Month Since Last Claim**

*Those customers who aren’t claiming their insurance for a considerable amount of time are the one who might not stick to the same company for long duration.*

According to my model, for every increment in months for last claim there is a decrement in CLV by 2 units.

* **On the Basis of Month Since Policy Inception**

*The customers who have come a long stride forward with policy’s effective date tends to have negative impact on CLV.*

As per my model, for every increment in months since policy inception there is a decrement in CLV by 6 units.

* **On the Basis of Number of Open Complaints**

*The most important aspect of any company is to solve their customer’s grievances. The customers with a lot of complaints*

*forms a negative relation with company. If their issue isn’t resolved then they might not further buy any other insurance claim.*

As per my model, we are considering 0 Complaint as base for

evaluating the rest in comparison to it.

For every increment in number of customers with 1 complaint

there is a decrement in CLV by 6 units in comparison to 0

complaint.

Similarly, for every increment in number of customers with 2

complaint there is a decrement in CLV by 1 units in comparison to 0 complaint.

Similarly, for every increment in number of customers with 3 complaint there is a decrement in CLV by 3 units in comparison to 0 complaint.

Similarly, for every increment in number of customers with 4 complaint there is a decrement in CLV by 3 units in comparison to 0 complaint.

And, for every increment in number of customers with 5 complaint there is a decrement in CLV by 6 units in comparison to 0 complaint.

* **On the Basis of Policy Type**

*The customers with policy type of* ***personal auto*** *forms a negative relation with CLV.*

According to my model, we are considering Corporate Auto as our base and evaluating the rest in comparison to it.

So, for every increment in policy type personal auto there is a

decrement in CLV by 6 unit in comparison to Corporate Auto.

* **On the Basis of Policy**

*The Customers with Corporate L3 Policy shows negative relationship to CLV. So, company should reconsider their terms and condition .And come up with a better version of this policy.*

As per my model, we are considering Corporate L1as base for

evaluating the rest in comparison to it.

For every increment in corporate L3 policy there is a decrement in CLV by 7 units in comparison to corporate L1.

* **On the Basis of Renew offer Type**

*The customers who renew their offer special offer 2 are the one who are having a negative relationship with the company. This may be with some terms and clauses which they find inconvenient*

*in the renewed offer.*

According to my model, we are considering Offer 1as our base and evaluating the rest in comparison to it.

For every increment in renew Offer 2 policy there is a decrement in CLV by 5 units in comparison to Offer1.