Individual Assignment #2

AD699: Data Mining for Business Analytics

Boston University

Spring 2020

Yuqi Zheng

Simple Linear Regression:

1.

```
lendingclub<-read.csv("lendingclub.csv")

dim(lendingclub)

View(lendingclub)
```

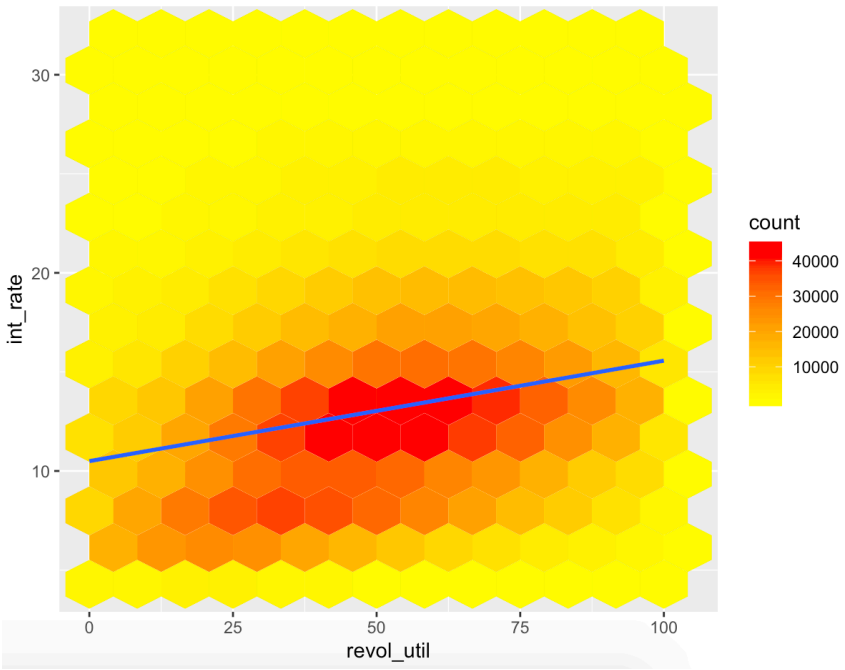
By the way, we could check the dimension of this file too.

2.

```
lendingclub<-filter(lendingclub,revol_util <=100)

range(lendingclub$revol_util)
```

3.

```
ggplot(lendingclub,aes(x=revol_util,y=int_rate))+geom_hex(bins=12)+scale_fill_gradient(low="
black",high="red")+geom_smooth(method="lm")
```



Graph 1

The revolving utilization ratio is also known as the debt-to-limit ratio or credit utilization ratio. Above graph shows the relationship between revol_util and int_rate. We could see there are highest density around 50 revol_util with 13% int_rate for 40000 people in the lending club, which means a majority of people in lending club have around 13% personal loan rate with using half of total credit availability. And the int_rate best-fit line is increasing when revol_util increase. It makes sense because lower personal interest loan rate means the people have more cash flow so that they will use less credits too.

4.

```
> cor(lendingclub$revol_util,lendingclub$int_rate)
[1] 0.2588485
```

The correlation between revol_util and int_rate is 0.2588485.

5.

```
dim(lendingclub)
set.seed(480)
club <- sample_n(lendingclub, 2131401)
View(club)
2131401*0.6
train <- slice(club, 1:1278841)
valid <- slice(club, 1278842:2131401)
```

6.

```
options(scipen=999)
linearClubTrain<-lm(int_rate~revol_util, data=train)
linearClubTrain
```

summary(linearClubTrain)

```
> options(scipen=999)
> linearClubTrain<-lm(int_rate~revol_util, data=train)
> linearClubTrain

Call:
lm(formula = int_rate ~ revol_util, data = train)

Coefficients:
(Intercept)   revol_util
    10.4982       0.0506

> summary(linearClubTrain)

Call:
lm(formula = int_rate ~ revol_util, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2229  -3.4450  -0.6048   2.5903  20.4918

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) 10.4981788  0.0093625  1121.3 <0.0000000000000002 ***
revol_util   0.0505990  0.0001673   302.5 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.627 on 1278839 degrees of freedom
Multiple R-squared:  0.06677,   Adjusted R-squared:  0.06677
F-statistic: 9.15e+04 on 1 and 1278839 DF,  p-value: < 0.00000000000000022
```

Graph 2

7.

Normally the higher R square number, the better the model fits to the data. Adding more variables will increase the R square, because adding more variables might lead to overfit the model. It will lead us to have more predicts for our dataset. In addition, trying many different variables in data mining process will introduce a variety of problems, including misleading coefficients and inflated R-squared value.

8.

range(train$revol_util)

df_pred<-predict(linearClubTrain,data.frame(revol_util=25))

df_pred

```
> range(train$revol_util)
[1]   0 100
> df_pred<-predict(linearClubTrain,data.frame(revol_util=25))
> df_pred
       1
11.76315
```

Graph 3

The predict outcome is 11.76315.

9.

```
library(forecast)
predTrain <- predict(linearClubTrain, train)
accuracy(predTrain, train$int_rate)
predValid <- predict(linearClubTrain, valid)
accuracy(predValid, valid$int_rate)
```

```
> library(forecast)
> predTrain <- predict(linearClubTrain, train)
> accuracy(predTrain, train$int_rate)
                         ME     RMSE      MAE       MPE     MAPE
Test set -0.000000000007773284 4.627466 3.623806 -13.42498 32.12573
>
>
> predValid <- predict(linearClubTrain, valid)
> accuracy(predValid, valid$int_rate)
                 ME     RMSE      MAE       MPE     MAPE
Test set 0.00887468 4.633069 3.626284 -13.35531 32.09116
```

Graph 4

Comparing to the training set and the validation set, the MAE of both sets looks steady and the RMSE of validation set is also similar to each other. All of which suggest that my model has the nearly same performance against training set and validation set, and it is able to make a rather accurate prediction.

Multiple Linear Regression:

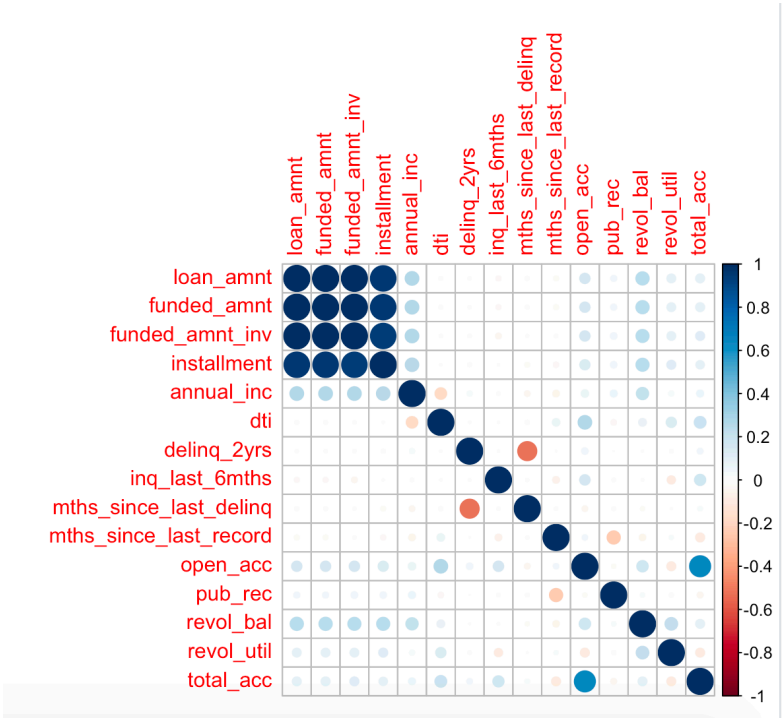
1.

```
library(corrplot)

names(train)

numbers<-train[,c(2:4,7,10,13:22)]

traincor <- cor(numbers,use='complete.obs')

traincor

corrplot(traincor)

df_new<-train[,c(2:4,6,7,18,22)]

cor(df_new,use='complete.obs')

train2<-train[,-c(1,3,4,7,18)]

names(train2)
```

```
> library(corrplot)
> names(train)
 [1] "member_id"             "loan_amnt"             "funded_amnt"           "funded_amnt_inv"       "term"
 [6] "int_rate"              "installment"           "emp_length"            "home_ownership"        "annual_inc"
[11] "verification_status"   "purpose"               "dti"                   "delinq_2yrs"           "inq_last_6mths"
[16] "mths_since_last_delinq" "mths_since_last_record" "open_acc"             "pub_rec"               "revol_bal"
[21] "revol_util"            "total_acc"
> numbers<-train[,c(2:4,7,10,13:22)]
>
> traincor <- cor(numbers,use='complete.obs')
> traincor
                          loan_amnt funded_amnt funded_amnt_inv  installment  annual_inc          dti   delinq_2yrs inq_last_6mths
loan_amnt              1.00000000  0.99988395     0.993768095  0.953323730  0.26824569 -0.015929064  0.0106483403   -0.030271737
funded_amnt            0.99988395  1.00000000     0.994038912  0.953436968  0.26821793 -0.015837714  0.0106745461   -0.030631290
funded_amnt_inv        0.99376809  0.99403891     1.000000000  0.946340950  0.26673056 -0.011327964  0.0125351541   -0.041582081
installment            0.95332373  0.95343697     0.946340950  1.000000000  0.25898469 -0.004660920  0.0170501827   -0.004129043
annual_inc             0.26824569  0.26821793     0.266730562  0.258984686  1.00000000 -0.188328618  0.0328452191    0.014043050
dti                   -0.01592906 -0.01583771    -0.011327964 -0.004660920 -0.18832862  1.000000000 -0.0046736937   -0.004651195
delinq_2yrs            0.01064834  0.01067455     0.012535154  0.017050183  0.03284522 -0.004673694  1.0000000000    0.005444113
inq_last_6mths        -0.03027174 -0.03063129    -0.041582081 -0.004129043  0.01404305 -0.004651195  0.0054441132    1.000000000
mths_since_last_delinq -0.01743892 -0.01732636    -0.009344794 -0.027110437 -0.04065715  0.019847699 -0.5273623856    0.010588350
mths_since_last_record -0.02965557 -0.02948069    -0.012909071 -0.034770349 -0.05764912  0.076986674  0.0014002261   -0.060527613
open_acc               0.15013156  0.15019695     0.151751837  0.145567542  0.07570874  0.264668515  0.0565111356    0.154080094
pub_rec                0.04900922  0.04922729     0.058646166  0.054653037  0.06540075 -0.038168552 -0.0001772762   -0.013079980
revol_bal              0.24935734  0.24915895     0.240836043  0.242862840  0.21341230  0.081490341  0.0026402369   -0.006261439
revol_util             0.10550225  0.10535895     0.104835598  0.112631360  0.03384890  0.146884536 -0.0128871465   -0.092406475
total_acc              0.10965339  0.10970577     0.112371464  0.098415188  0.06548165  0.197122140  0.0445262580    0.177272667
                       mths_since_last_delinq mths_since_last_record    open_acc       pub_rec    revol_bal   revol_util   total_acc
loan_amnt                      -0.017438918           -0.029655571  0.15013156  0.0490092211  0.249357344  0.105502253  0.10965339
funded_amnt                    -0.017326362           -0.029480686  0.15019695  0.0492272939  0.249158946  0.105358955  0.10970577
funded_amnt_inv                -0.009344794           -0.012909071  0.15175184  0.0586461663  0.240836043  0.104835598  0.11237146
installment                    -0.027110437           -0.034770349  0.14556754  0.0546530367  0.242862840  0.112631360  0.09841519
annual_inc                     -0.040657152           -0.057649119  0.07570874  0.0654007492  0.213412304  0.033848898  0.06548165
dti                             0.019847699            0.076986674  0.26466852 -0.0381685520  0.081490341  0.146884536  0.19712214
delinq_2yrs                    -0.527362386            0.001400226  0.05651114 -0.0001772762  0.002640237 -0.012887146  0.04452626
inq_last_6mths                  0.010588350           -0.060527613  0.15408009 -0.0130799796 -0.006261439 -0.092406475  0.17727267
mths_since_last_delinq          1.000000000           -0.012169305 -0.04146804  0.0151054616 -0.020364320  0.002194953  0.02696689
mths_since_last_record         -0.012169305            1.000000000  0.04022205 -0.2477989009 -0.053508015  0.020724753 -0.10986203
open_acc                       -0.041468044            0.040222050  1.00000000 -0.0212609489  0.183338134 -0.109226177  0.64619306
pub_rec                         0.015105462           -0.247798901 -0.02126095  1.0000000000  0.024102364  0.019719609 -0.04421030
revol_bal                      -0.020364320           -0.053508015  0.18333813  0.0241023644  1.000000000  0.226388708  0.10833819
revol_util                      0.002194953            0.020724753 -0.10922618  0.0197196091  0.226388708  1.000000000 -0.10772067
total_acc                       0.026966893           -0.109862031  0.64619306 -0.0442102986  0.108338191 -0.107720671  1.00000000
```

```
> cor(df_new,use='complete.obs')
              loan_amnt funded_amnt funded_amnt_inv     int_rate installment     open_acc   total_acc
loan_amnt     1.0000000   0.9997331       0.9989223  0.100399522   0.9467392  0.187975055  0.20721420
funded_amnt   0.9997331   1.0000000       0.9992493  0.100382973   0.9471155  0.188173351  0.20714674
funded_amnt_inv 0.9989223 0.9992493       1.0000000  0.100395522   0.9461679  0.188391158  0.20706232
int_rate      0.1003995   0.1003830       0.1003955  1.000000000   0.1230089 -0.008214117 -0.03800491
installment   0.9467392   0.9471155       0.9461679  0.123008931   1.0000000  0.176028389  0.18506368
open_acc      0.1879751   0.1881734       0.1883912 -0.008214117   0.1760284  1.000000000  0.71617524
total_acc     0.2072142   0.2071467       0.2070623 -0.038004909   0.1850637  0.716175242  1.00000000
>
>
> train2<-train[,-c(1,3,4,7,18)]
> names(train2)
 [1] "loan_amnt"             "term"                  "int_rate"              "emp_length"            "home_ownership"
 [6] "annual_inc"            "verification_status"   "purpose"               "dti"                   "delinq_2yrs"
[11] "inq_last_6mths"        "mths_since_last_delinq" "mths_since_last_record" "pub_rec"              "revol_bal"
[16] "revol_util"            "total_acc"
```

Graph 5



Graph 6

Firstly, I select “loan_amnt”, “funded_amnt”, “funded_amnt_inv”, “int_rate”, “installment”, “open_acc” and “total_acc” from the training set I created before as the variables that I want to further examine. Secondly I use “cor()” function to see the correlations among these variables and create a correlation matrix to see the results. Then I find out four pairs of variables are strongly correlated to each other (the values of correlation are greater than 0.9), and those four pairs are coming from “loan_amnt”, “funded_amnt”, “funded_amnt_inv” and “installment”. Also, the “open_acc” and “total_occ” have a strong correlation. To avoid overfitting issue, we

only keep two variables in these six variables. At last I created a dataset without “member_id”, “funded_amnt”, funded_amnt_inv”, “installment” and “open_acc”.

2.

```
Club1<-lm(int_rate~.,train2)
summary(Club1)
```

Dummy variables are numeric variables that represent categorical data, such as “Private” in College dataset. They can take on only two mutually exclusive values and use 1 to represent presence and 0 for absence. It allows us to treat categorical data as numeric data and easily interpret regression results.

3.

```
Club1Step <- step(Club1, direction = "backward")
summary(Club1Step)
View(train2)
```

4.

As the results shown as Graph 7 which the results from previous steps, I decide to remove all the category called “emp_length years” and “home_ownership” due to large p_values.

As I believe it would be better to see how purpose effect on int_rate, how the annual income effect on int_rate and if it true that the more annual income, the lower int_rate of customer in the lending club.

```
> summary(Club1Step)

Call:
lm(formula = int_rate ~ loan_amnt + term + emp_length + home_ownership +
    annual_inc + verification_status + purpose + dti + delinq_2yrs +
    inq_last_6mths + mths_since_last_delinq + mths_since_last_record +
    pub_rec + revol_bal + revol_util + total_acc, data = train2)

Residuals:
     Min       1Q   Median       3Q      Max
-27.9121  -2.5694  -0.4386   2.0425  22.8825

Coefficients:
                                      Estimate    Std. Error t value             Pr(>|t|)
(Intercept)                       10.6870822562  0.5574967209  19.170 < 0.0000000000000002 ***
loan_amnt                          0.0000379589  0.0000017678  21.473 < 0.0000000000000002 ***
term 60 months                     3.9935561898  0.0304697394 131.066 < 0.0000000000000002 ***
emp_length1 year                  -0.0057521064  0.0685191787  -0.084             0.933097
emp_length10+ years               -0.1222115865  0.0511862516  -2.388             0.016961 *
emp_length2 years                  0.0381705045  0.0630177799   0.606             0.544709
emp_length3 years                  0.0648122886  0.0640713562   1.012             0.311749
emp_length4 years                  0.0413670834  0.0680491501   0.608             0.543255
emp_length5 years                 -0.0241618972  0.0673582932  -0.359             0.719815
emp_length6 years                  0.0595601980  0.0731293831   0.814             0.415389
emp_length7 years                 -0.0287213169  0.0754532948  -0.381             0.703464
emp_length8 years                 -0.1188968516  0.0748592374  -1.588             0.112228
emp_length9 years                 -0.0001547515  0.0794880635  -0.002             0.998447
emp_lengthn/a                     -0.1467236377  0.0652491420  -2.249             0.024536 *
home_ownershipMORTGAGE            -1.7092016949  0.5422622969  -3.152             0.001622 **
home_ownershipNONE                -1.8960657818  2.6818326116  -0.707             0.479566
home_ownershipOTHER                0.0143414343  2.6822340914   0.005             0.995734
home_ownershipOWN                 -1.2052683439  0.5431505508  -2.219             0.026487 *
home_ownershipRENT                -1.1192835581  0.5423254325  -2.064             0.039034 *
annual_inc                        -0.0000020116  0.0000001512 -13.302 < 0.0000000000000002 ***
verification_statusSource Verified  0.8600220860  0.0298367917  28.824 < 0.0000000000000002 ***
verification_statusVerified        1.9301113565  0.0333849506  57.814 < 0.0000000000000002 ***
purposecredit_card                -1.6510541301  0.1173957177 -14.064 < 0.0000000000000002 ***
purposedebt_consolidation         -0.1146402372  0.1150869893  -0.996             0.319195
purposeeducational                -1.6137665124  0.6076463040  -2.656             0.007914 **
purposehome_improvement            0.1330661891  0.1210751985   1.099             0.271754
purposehouse                       2.0941681716  0.1820417845  11.504 < 0.0000000000000002 ***
purposemajor_purchase              0.2439971465  0.1395360317   1.749             0.080358 .
purposemedical                     0.9410288270  0.1528925603   6.155  0.00000000075451007 ***
purposemoving                      1.4747338949  0.1878680658   7.850  0.00000000000000421 ***
purposeother                       1.3478923975  0.1231932319  10.941 < 0.0000000000000002 ***
purposerenewable_energy            1.5261803136  0.4331980994   3.523             0.000427 ***
purposesmall_business              2.1150482081  0.1522597334  13.891 < 0.0000000000000002 ***
purposevacation                    0.7585604849  0.1839478233   4.124  0.00003730172525897 ***
purposewedding                     0.9596128647  0.5482062780   1.750             0.080042 .
dti                                0.0836165382  0.0015619723  53.533 < 0.0000000000000002 ***
delinq_2yrs                        0.0995469326  0.0135646188   7.339  0.00000000000021734 ***
inq_last_6mths                     0.8933388904  0.0113410976  78.770 < 0.0000000000000002 ***
mths_since_last_delinq            -0.0077914458  0.0006184442 -12.598 < 0.0000000000000002 ***
mths_since_last_record             0.0056372965  0.0004619880  12.202 < 0.0000000000000002 ***
pub_rec                            0.0310191764  0.0127091574   2.441             0.014661 *
revol_bal                         -0.0000168726  0.0000007629 -22.116 < 0.0000000000000002 ***
revol_util                         0.0243176959  0.0005788692  42.009 < 0.0000000000000002 ***
total_acc                         -0.0210053161  0.0010932826 -19.213 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.714 on 95888 degrees of freedom
  (1182909 observations deleted due to missingness)
Multiple R-squared:  0.3343,    Adjusted R-squared:  0.3341
```

Graph 7

5.

```
train3<-train2[,-c(4,5)]

Club2<-lm(int_rate~.,train3)

summary(Club2)

Club2Step2 <- step(Club2, direction = "backward")

summary(Club2Step2)
```

```
> summary(Club2)

Call:
lm(formula = int_rate ~ ., data = train3)

Residuals:
     Min       1Q   Median       3Q      Max
-27.7809  -2.5786  -0.4371   2.0427  24.0031

Coefficients:
                                      Estimate    Std. Error t value             Pr(>|t|)
(Intercept)                         9.3926597469  0.1308791018  71.766 < 0.0000000000000002 ***
loan_amnt                           0.0000355932  0.0000017644  20.173 < 0.0000000000000002 ***
term 60 months                      3.9731059276  0.0305336999 130.122 < 0.0000000000000002 ***
annual_inc                         -0.0000021151  0.0000001514 -13.973 < 0.0000000000000002 ***
verification_statusSource Verified  0.8969973247  0.0298717705  30.028 < 0.0000000000000002 ***
verification_statusVerified         1.9443032678  0.0331449015  58.661 < 0.0000000000000002 ***
purposecredit_card                 -1.6726787730  0.1177270798 -14.208 < 0.0000000000000002 ***
purposedebt_consolidation          -0.1385391590  0.1154019156  -1.200             0.229951
purposeeducational                 -1.4549315000  0.6092176617  -2.388             0.016933 *
purposehome_improvement            -0.0660993870  0.1211170074  -0.546             0.585240
purposehouse                        2.1165992080  0.1825686151  11.593 < 0.0000000000000002 ***
purposemajor_purchase               0.2344977012  0.1399364257   1.676             0.093792 .
purposemedical                      0.9393098628  0.1533380834   6.126    0.000000000906101 ***
purposemoving                       1.6067329736  0.1882528239   8.535 < 0.0000000000000002 ***
purposeother                        1.3409782790  0.1235469419  10.854 < 0.0000000000000002 ***
purposerenewable_energy             1.4566910609  0.4344986705   3.353             0.000801 ***
purposesmall_business               2.1046407416  0.1526934362  13.783 < 0.0000000000000002 ***
purposevacation                     0.7557328339  0.1844653514   4.097    0.000041909520263 ***
purposewedding                      0.9507426413  0.5498346812   1.729             0.083787 .
dti                                 0.0853334091  0.0015639323  54.563 < 0.0000000000000002 ***
delinq_2yrs                         0.0978157893  0.0136021488   7.191    0.000000000000647 ***
inq_last_6mths                      0.8924179719  0.0113659778  78.517 < 0.0000000000000002 ***
mths_since_last_delinq             -0.0074387618  0.0006196685 -12.004 < 0.0000000000000002 ***
mths_since_last_record              0.0051398810  0.0004626351  11.110 < 0.0000000000000002 ***
pub_rec                             0.0257892688  0.0127372422   2.025             0.042900 *
revol_bal                          -0.0000177471  0.0000007636 -23.243 < 0.0000000000000002 ***
revol_util                          0.0233744066  0.0005784969  40.405 < 0.0000000000000002 ***
total_acc                          -0.0235820023  0.0010905773 -21.623 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.725 on 95904 degrees of freedom
  (1182909 observations deleted due to missingness)
Multiple R-squared:  0.3302,    Adjusted R-squared:   0.33
F-statistic:  1751 on 27 and 95904 DF,  p-value: < 0.00000000000000022
```

```
> Club2Step2 <- step(Club2, direction = "backward")
Start:  AIC=252354.6
int_rate ~ loan_amnt + term + annual_inc + verification_status +
    purpose + dti + delinq_2yrs + inq_last_6mths + mths_since_last_delinq +
    mths_since_last_record + pub_rec + revol_bal + revol_util +
    total_acc

                        Df Sum of Sq     RSS    AIC
<none>                                1330903 252355
- pub_rec                1        57 1330960 252357
- delinq_2yrs            1       718 1331621 252404
- mths_since_last_record 1      1713 1332616 252476
- mths_since_last_delinq 1      2000 1332903 252497
- annual_inc             1      2710 1333613 252548
- loan_amnt              1      5647 1336551 252759
- total_acc              1      6489 1337392 252819
- revol_bal              1      7497 1338400 252892
- revol_util             1     22656 1353560 253972
- dti                    1     41315 1372219 255285
- verification_status    2     48090 1378993 255756
- purpose               13     67011 1397914 257041
- inq_last_6mths         1     85553 1416456 258329
- term                   1    234969 1565873 267950
```

```
> summary(Club2Step2)

Call:
lm(formula = int_rate ~ loan_amnt + term + annual_inc + verification_status +
    purpose + dti + delinq_2yrs + inq_last_6mths + mths_since_last_delinq +
    mths_since_last_record + pub_rec + revol_bal + revol_util +
    total_acc, data = train3)

Residuals:
     Min       1Q   Median       3Q      Max
-27.7809  -2.5786  -0.4371   2.0427  24.0031

Coefficients:
                                    Estimate    Std. Error t value             Pr(>|t|)
(Intercept)                       9.3926597469  0.1308791018  71.766 < 0.0000000000000002 ***
loan_amnt                         0.0000355932  0.0000017644  20.173 < 0.0000000000000002 ***
term 60 months                    3.9731059276  0.0305336999 130.122 < 0.0000000000000002 ***
annual_inc                       -0.0000021151  0.0000001514 -13.973 < 0.0000000000000002 ***
verification_statusSource Verified  0.8969973247  0.0298717705  30.028 < 0.0000000000000002 ***
verification_statusVerified       1.9443032678  0.0331449015  58.661 < 0.0000000000000002 ***
purposecredit_card               -1.6726787730  0.1177270798 -14.208 < 0.0000000000000002 ***
purposedebt_consolidation        -0.1385391590  0.1154019156  -1.200             0.229951
purposeeducational               -1.4549315000  0.6092176617  -2.388             0.016933 *
purposehome_improvement          -0.0660993870  0.1211170074  -0.546             0.585240
purposehouse                      2.1165992080  0.1825686151  11.593 < 0.0000000000000002 ***
purposemajor_purchase             0.2344977012  0.1399364257   1.676             0.093792 .
purposemedical                    0.9393098628  0.1533380834   6.126    0.000000000906101 ***
purposemoving                     1.6067329736  0.1882528239   8.535 < 0.0000000000000002 ***
purposeother                      1.3409782790  0.1235469419  10.854 < 0.0000000000000002 ***
purposerenewable_energy           1.4566910609  0.4344986705   3.353             0.000801 ***
purposesmall_business             2.1046407416  0.1526934362  13.783 < 0.0000000000000002 ***
purposevacation                   0.7557328339  0.1844653514   4.097    0.000041909520263 ***
purposewedding                    0.9507426413  0.5498346812   1.729             0.083787 .
dti                               0.0853334091  0.0015639323  54.563 < 0.0000000000000002 ***
delinq_2yrs                       0.0978157893  0.0136021488   7.191    0.000000000000647 ***
inq_last_6mths                    0.8924179719  0.0113659778  78.517 < 0.0000000000000002 ***
mths_since_last_delinq           -0.0074387618  0.0006196685 -12.004 < 0.0000000000000002 ***
mths_since_last_record            0.0051398810  0.0004626351  11.110 < 0.0000000000000002 ***
pub_rec                           0.0257892688  0.0127372422   2.025             0.042900 *
revol_bal                        -0.0000177471  0.0000007636 -23.243 < 0.0000000000000002 ***
revol_util                        0.0233744066  0.0005784969  40.405 < 0.0000000000000002 ***
total_acc                        -0.0235820023  0.0010905773 -21.623 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.725 on 95904 degrees of freedom
  (1182909 observations deleted due to missingness)
```

```
  (1182909 observations deleted due to missingness)
Multiple R-squared:  0.3302,    Adjusted R-squared:   0.33
F-statistic:  1751 on 27 and 95904 DF,  p-value: < 0.00000000000000022
```

Graph 8

6.

```
train4<-na.omit(train3)

ratemove<-train4$int_rate-mean(train4$int_rate)

ratemovesq<-ratemove^2

SST<-sum(ratemovesq)
```

```
SST

modelexp<-Club2Step2$fitted.values-mean(train4$int_rate)

modelexpsq<-modelexp^2

SSR<-sum(modelexpsq)

SSR

SSR/SST
```

```
> #6
> train4<-na.omit(train3)
> ratemove<-train4$int_rate-mean(train4$int_rate)
> ratemovesq<-ratemove^2
> SST<-sum(ratemovesq)
> SST
[1] 1986941
>
> modelexp<-Club2Step2$fitted.values-mean(train4$int_rate)
> modelexpsq<-modelexp^2
> SSR<-sum(modelexpsq)
> SSR
[1] 656037.3
>
> SSR/SST
[1] 0.3301746
```

Graph 9

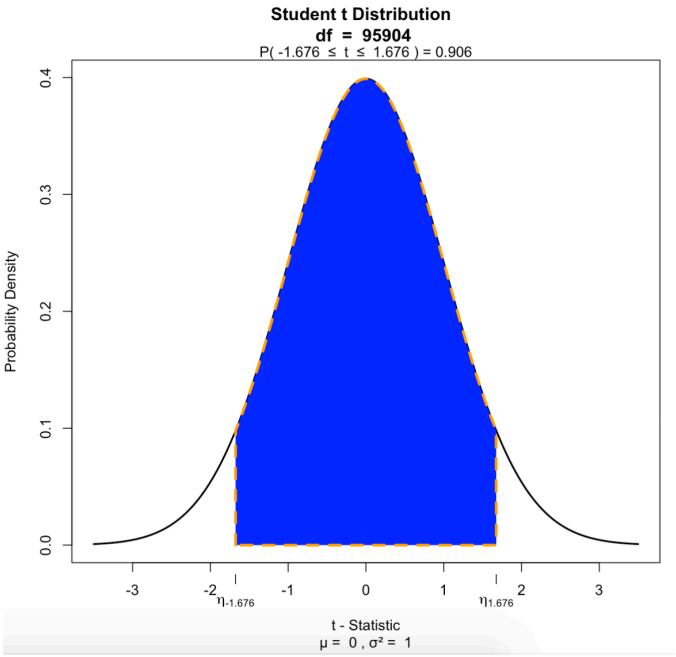
The total sum of squares for my model is 1986941, and the total sum of squares due to regression for my model is 656037.3. The result of SSR/SST is 0.3301746. The place where I found it is circled in the following pictures, which is in the summary of my model and is represented as "Multiple R-squared".

```
  (1182909 observations deleted due to missingness)
Multiple R-squared:  0.3302,    Adjusted R-squared:   0.33
F-statistic:  1751 on 27 and 95904 DF,  p-value: < 0.00000000000000022
```

Graph 10

7.

```
library(visualize)

visualize.t(stat=c(-1.676,1.676),df=95904,section = "bounded")

visualize.t(stat=c(-1.676,1.676),df=95904,section = "tails")
```



Graph 11

I chose the “purposemajor_purchase” factor. The p-value and the t-value of it are 0.093792 and -1.676 respectively.

90.6% of the curve is shaded.

The p-value equals to 1 – 90.6% = 9.4%, which also equals to the result shown in summary – 0.093792.

8.

```
temp<-train2[,c(1,6,9:17)]
sapply(temp,range,na.rm=TRUE)
table(train2$dti)
Tina <- data.frame(loan_amnt =30000,term=" 60 months",emp_length="7
years",home_ownership="RENT",annual_inc=777777,verification_status="Verified",
purpose="small_business",dti=2,delinq_2yrs=32,inq_last_6mths =15,mths_since_last_delinq
=160,mths_since_last_record=77,pub_rec =5,revol_bal=500000,revol_util=17,total_acc=60)
```

predTina <-predict(Club2Step2,Tina)

predTina

There is a person named Tina, her loan_amnt is 30000 with term 60 months, and her emp_length is 7 years with a rent home ownership. Her annual income is $777777 with verified status. Her lending purpose is a small business with dti=2, delinq_2yrs=32, inq_last_6mths =15, mths_since_last_delinq =160, mths_since_last_record=77, pub_rec =5, revol_bal=500000, revol_util=17, total_acc=60.

After running my model, Tina's interest rate is 22.96783%.

```
> sapply(temp,range,na.rm=TRUE)
     loan_amnt annual_inc   dti delinq_2yrs inq_last_6mths mths_since_last_delinq mths_since_last_record pub_rec revol_bal revol_util total_acc
[1,]       500       2000  0.00          0             0                      0                      0       0         0          0         1
[2,]     40000  110000000 49.96         42            33                    202                    129      63   2904836        100       176
>
```

```
> predTina
       1
22.96783
```

Graph 12

9.

predTrain2 <- predict(Club2Step2, train2)

accuracy(predTrain2, train2$int_rate)

predValid2 <- predict(Club2Step2, valid)

accuracy(predValid2, valid$int_rate)

As the results shown that the mean error of my model against the training set is extremely, however, the RMSE, MAE, MPE and MAPE is relatively high. Then we can see that values of ME, RMSE and MAE have become even higher in accuracy test against validation set. All of which suggest that the model work fine with training dataset, but it may become less accurate when it comes to validation dataset. In general, the model is able to give a relatively reliable prediction to the interest rate in training set.

As previous mentioned, adding more variables will increase the R square, because adding more variables might lead to overfit the model. Overfitting a model is a condition where a statistical model begins to describe the random error in the data rather than the relationships between variables. Although we already removed some variables in previous steps, we still face the overfitting risk. But since the results of training set and the validation set are very similar, our model is not overfitting at this time.

Comparing to my SLR model, my MLR model is more accurate because the mean error, RMSE, MAE, MPE and MAPE are all decreased. It makes sense to me because multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. Thus, the multiple linear regression is more accurate than the simple linear regression.

```
> predTrain2 <- predict(Club2Step2, train2)
> accuracy(predTrain2, train2$int_rate)
                            ME     RMSE      MAE       MPE     MAPE
Test set -0.000000000000009359953 3.724702 2.872265 -7.227733 22.32875
>
>
> predValid2 <- predict(Club2Step2, valid)
> accuracy(predValid2, valid$int_rate)
                 ME    RMSE      MAE       MPE     MAPE
Test set 0.03937234 3.74857 2.885954 -6.929995 22.24082
```

Graph 13