Report_LendingClub

Qi Huang

12/7/2019

Outline - Abstract This project is about investigating if the interest rate and loan amount have relationship. When the company decides to issue a loan to borrowers, would the larger amount of loan has higher interest rate or not. Data used is the transaction data between 2012 to 2015, from Lending Club.Based on brief findings on EDA, credit rating is another important influencer on interest rate, in other words, borrowers with different credit rating tend to have different interest rate, and borrowers within the group of same credit rating tend to have similar interest rate. Therefore, mixed effect model is applied and selected in this project. The results of the mixed effect model show that loan amount is negatively correlated with interest rate. Another important finding is about transactions that have 6% interest rate. A special group of transactions, with borrowers' credit rating varies from A to F, has the same 6% interest rate, which is much lower than average. This finding deserves future investigation to avoid default risk.

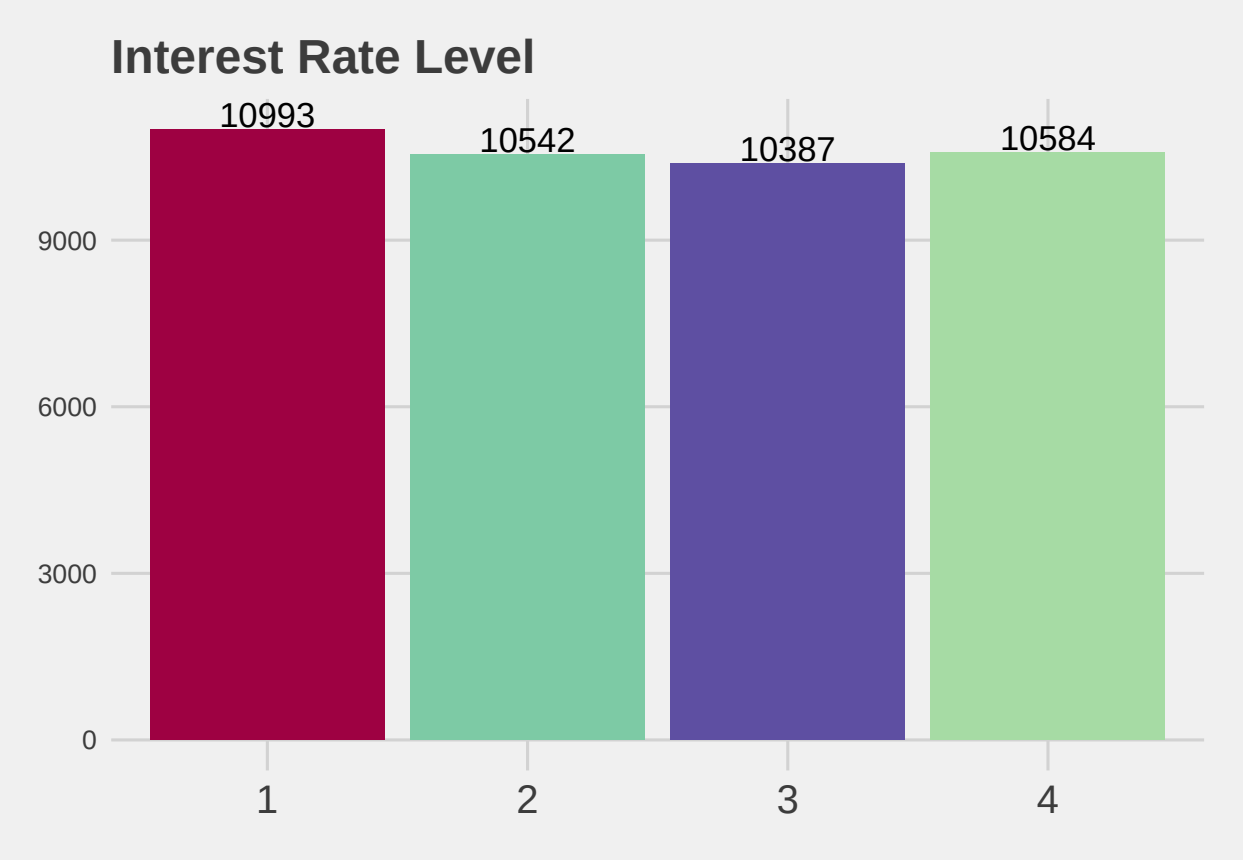
- Introduction Introduction Deciding a reasonable interest rate is significant for lending companies. As the major business, personal financing contributes to a large portion of revenues of lending companies, and revenues directly related to interest rates of every transaction. Usually, interest rates depend both on intrinsic features of loans issued and on client's credit information. For example, long-term loan tends to have higher interest rates compared to short-term loans because of the time risk, and clients who have lower credit scores tend to borrow loans with higher expenses. However, the relationship between loan amount and interest rate is not as clear as other factors. If different loan amount does contribute to different level of interest rates based on this dataset, then loan amount should be incorporated in loan pricing models.
- Data Description

```
##        id          loan_amnt             term          int_rate
##  1000007:    1   Min.   :  500    36 months:31505    10.99%:  970
##  1000030:    1   1st Qu.: 5200    60 months:11001    11.49%:  837
##  1000033:    1   Median : 9750                       13.49%:  832
##  1000045:    1   Mean   :11095                        7.51%:  787
##  1000067:    1   3rd Qu.:15000                        7.88%:  742
##  1000095:    1   Max.   :35000                        7.49%:  656
##  (Other):42500                                     (Other):37682
##   installment      grade       sub_grade         emp_length
##  Min.   :  15.67   A:10172   B3     : 2994   10+ years: 9366
##  1st Qu.: 165.67   B:12379   A4     : 2904   < 1 year : 5044
##  Median : 277.86   C: 8734   B5     : 2805   2 years  : 4742
##  Mean   : 322.76   D: 6015   A5     : 2791   3 years  : 4362
##  3rd Qu.: 428.50   E: 3393   B4     : 2587   4 years  : 3649
##  Max.   :1305.19   F: 1301   C1     : 2262   1 year   : 3592
##                    G:  512   (Other):26163   (Other)  :11751
##   home_ownership    annual_inc          verification_status
##  MORTGAGE:18952   Min.   :   1896   Not Verified   :18729
##  NONE    :    4   1st Qu.:  40000   Source Verified:10306
##  OTHER   :  136   Median :  59000   Verified       :13471
##  OWN     : 3250   Mean   :  69135
##  RENT    :20164   3rd Qu.:  82485
##                   Max.   :6000000
##
##      issue_d       delinq_2yrs      fico_range_low    open_acc
```

```
##  Dec-2011: 2267   Min.   : 0.0000   Min.   :610    Min.   : 1.000
##  Nov-2011: 2232   1st Qu.: 0.0000   1st Qu.:685    1st Qu.: 6.000
##  Oct-2011: 2118   Median : 0.0000   Median :710    Median : 9.000
##  Sep-2011: 2067   Mean   : 0.1524   Mean   :713    Mean   : 9.344
##  Aug-2011: 1934   3rd Qu.: 0.0000   3rd Qu.:740    3rd Qu.:12.000
##  Jul-2011: 1875   Max.   :13.0000   Max.   :825    Max.   :47.000
##  (Other) :30013
##    recoveries
##  Min.   :    0.0
##  1st Qu.:    0.0
##  Median :    0.0
##  Mean   :  103.2
##  3rd Qu.:    0.0
##  Max.   :29623.3
##
```

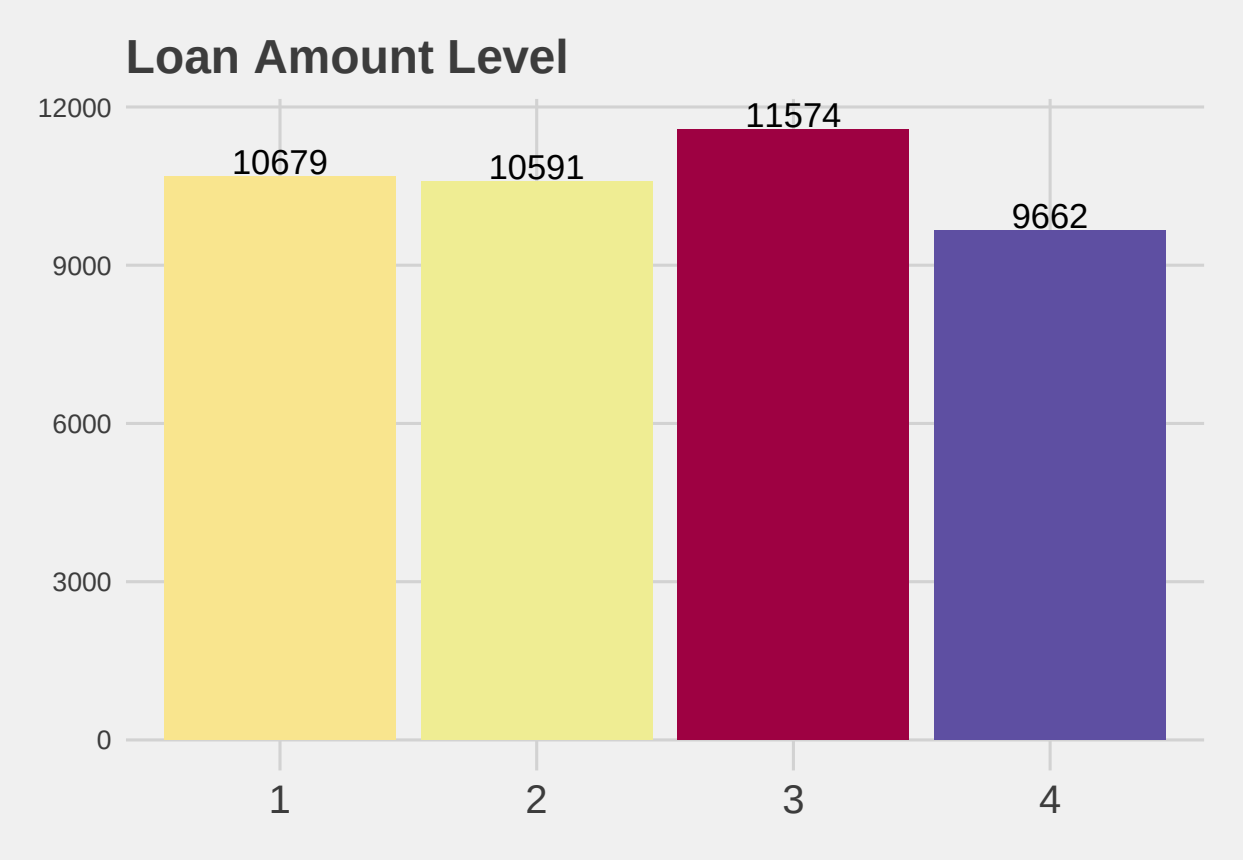
Data used in this project is the lending data between 2012 and 2015, from Lending club website. The original dataset has 150 variables, including both the borrower's information as well as the transaction information. Each transaction has a unique transaction id that serves as the primary key. There steps are taken when dealing with the data cleaning. First, rather than using all variables in the original dataset, certain variables that are related with the research question are selected to a new dataset. Selected variables are loan amount, term, installment, purpose, grade, subgrade, employment length, home ownership, annual income, verification status, delinquency within last two years, fico range, open account numbers, recoveries and application types. The new dataset has 18 total columns and 42506 transactions in total. Second, N/As are removed in the dataset. The new dataset includes 6 N/A, which accounts for 0.0001% of the whole data population. Therefore, removing those N/A will not strongly influence the pattern of the dataset. Third, the interest rate is a continuous variable that has different values for 42506 transactions, which is not convenient for EDA. Therefore, interest rate is classified to four different levels based on quantile and a new column of interest level is added into the dataset. Same for loan amount.

- EDA Two general goals are incorporated at the 8 plots in EDA part. First, to visualize if loan amount and interest rate has relationship, which is a brief view of the research question. Second, to visualize if other variables selected have relationship with the interest rate.
  Plot1: interest rate distribution



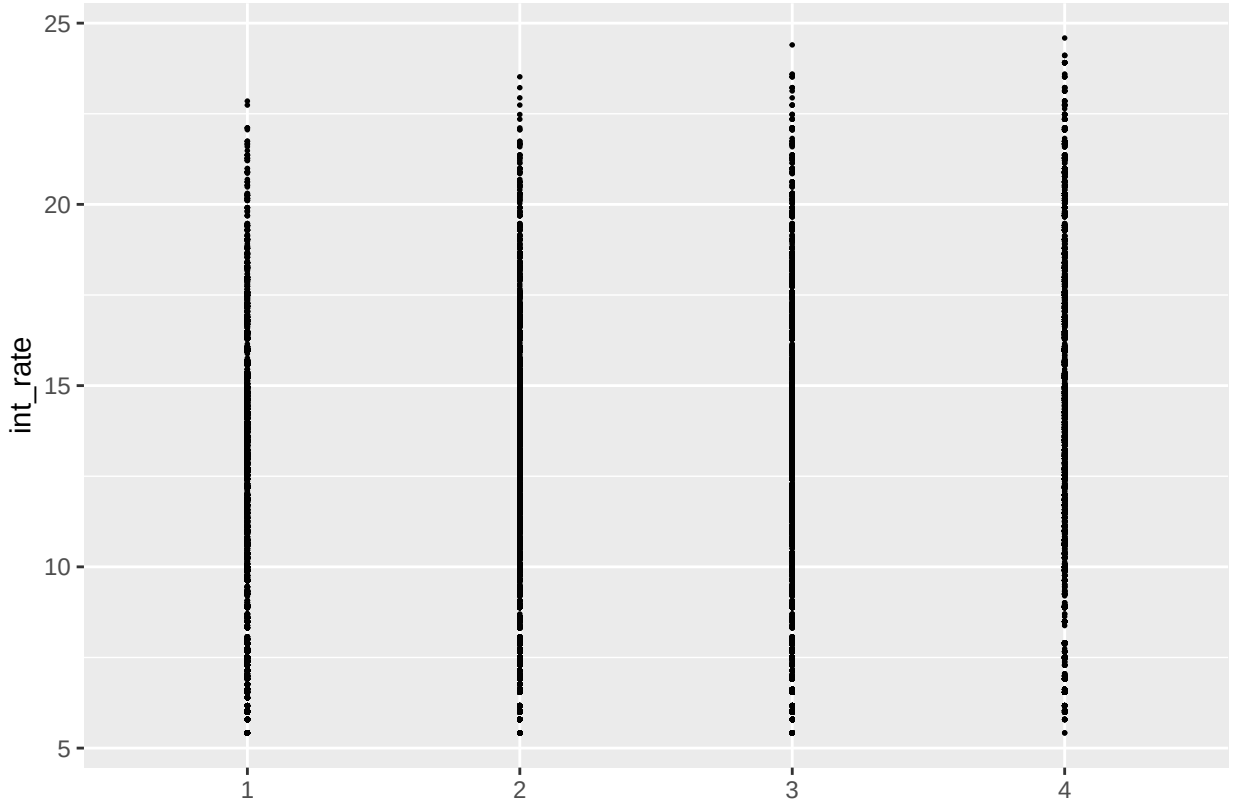
Plot1 shows the number of people in 4 different interest rate levels. Based on the plot, the number of people on each level of interest rate is similar.

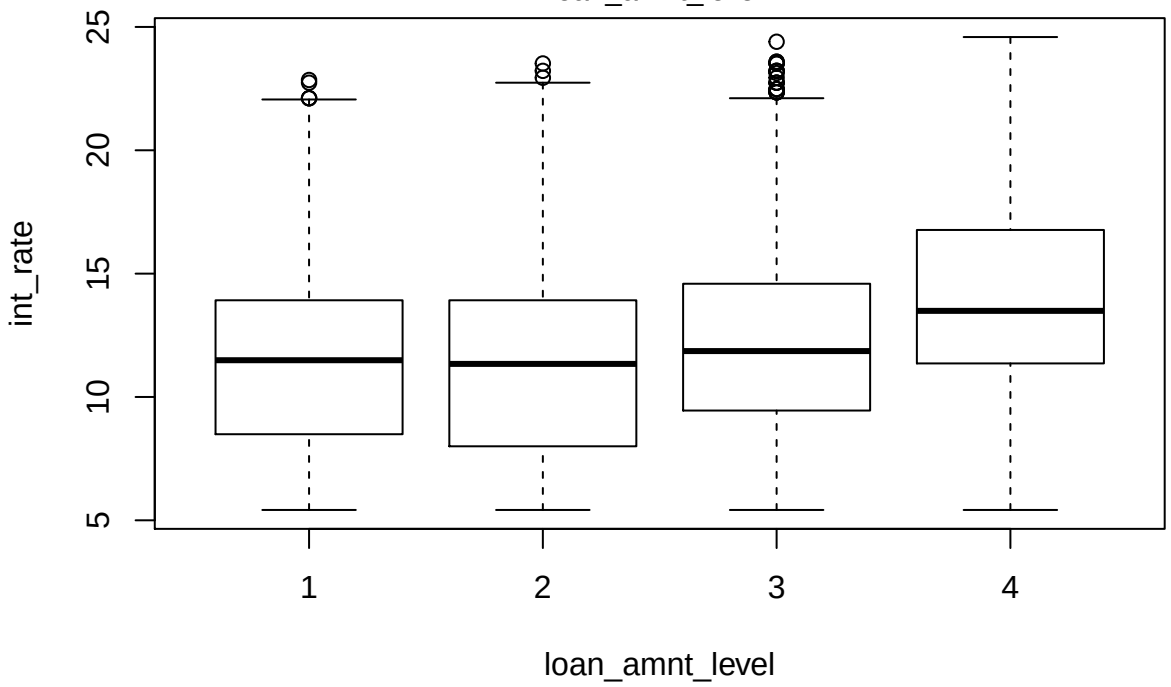
Plot2: loan amount distribution



Plot2 shows the number of people borrowing different amount of loan. Based on the plot, the loan amount that most people borrow is level 3, which is between 9750 to 15000. And least of people borrow loan that is larger than 15000.

Plot3:potential relationship between interest rate and loan amount



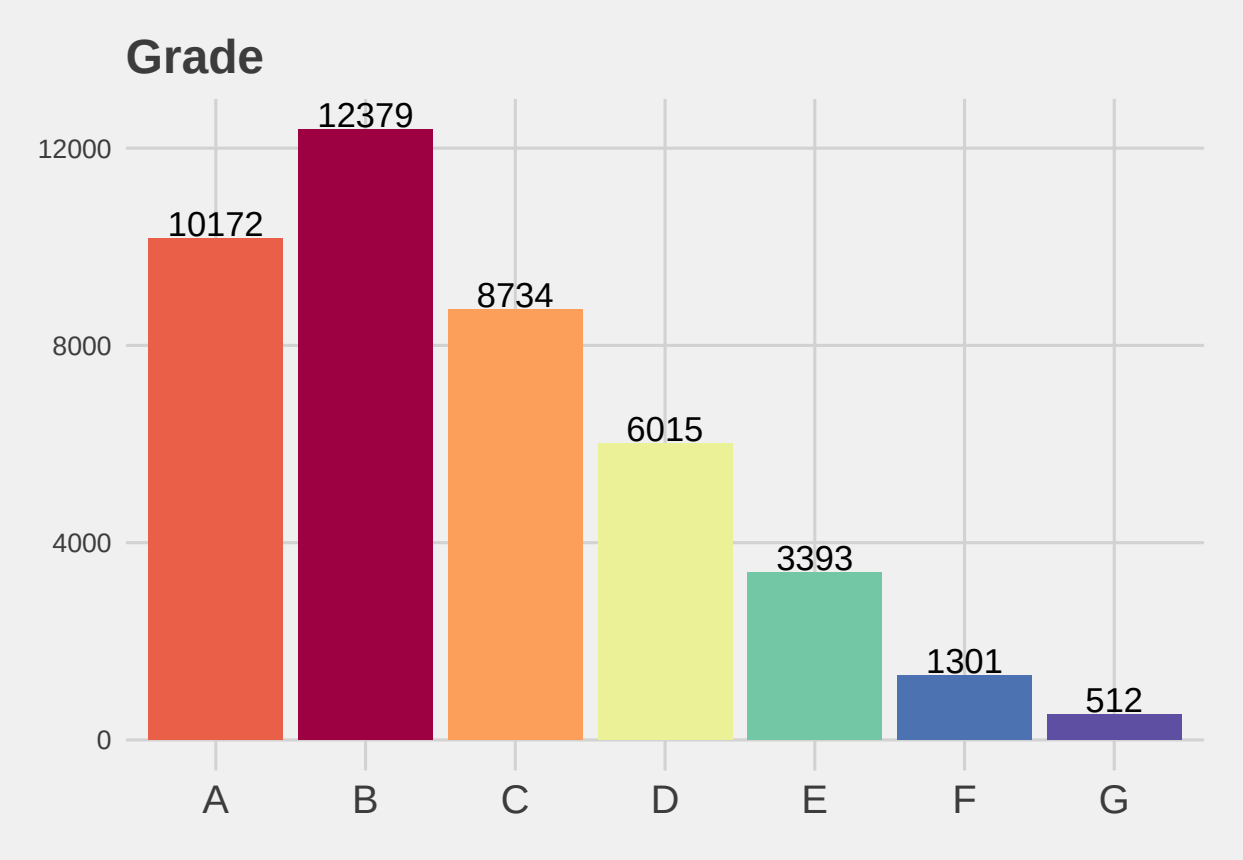



```
## $stats
##       [,1]  [,2]  [,3]  [,4]
## [1,]  5.42  5.42  5.42  5.42
## [2,]  8.49  8.00  9.45 11.36
## [3,] 11.49 11.34 11.86 13.49
## [4,] 13.92 13.92 14.59 16.77
```

```
## [5,] 22.06 22.74 22.11 24.59
##
## $n
## [1] 10679 10591 11574  9662
##
## $conf
##          [,1]     [,2]     [,3]     [,4]
## [1,] 11.40698 11.24911 11.78451 13.40304
## [2,] 11.57302 11.43089 11.93549 13.57696
##
## $out
##  [1] 22.74 22.11 22.11 22.85 22.11 23.52 23.22 22.94 23.52 23.13 22.35
## [12] 22.35 22.35 22.74 23.52 22.74 23.52 22.35 22.74 23.59 23.22 23.59
## [23] 23.22 22.48 22.48 23.22 22.94 24.40 22.48
##
## $group
##  [1] 1 1 1 1 1 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##
## $names
## [1] "1" "2" "3" "4"
```

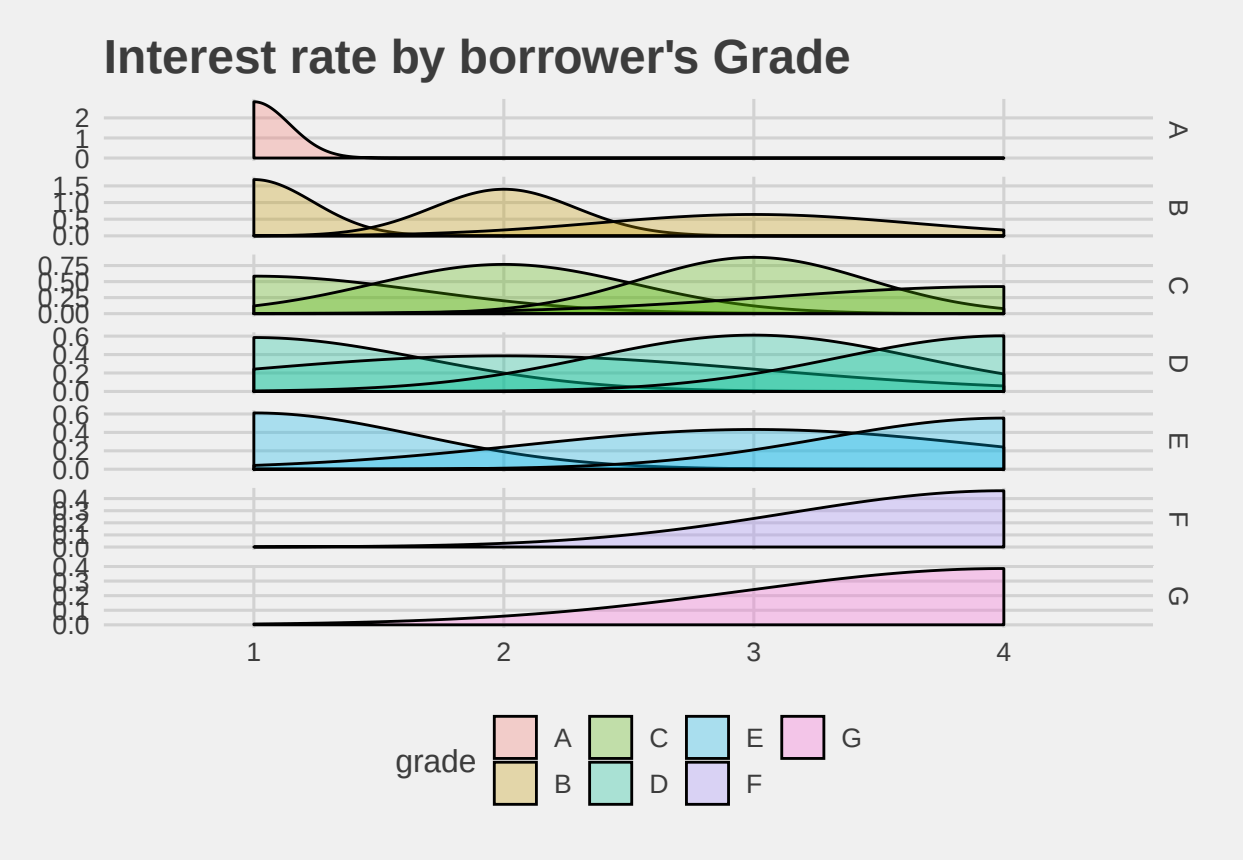
Plot3 gives a visualization of potential relationship between different level of loan amount and different level of interest rate. Based on the results, higher level of loan amount, which is the larger loan amount, tends to have higher interest rate. The relationship is clearer with the boxplot, where the mean of interest rate increases with the higher loan amount level. One finding related with the boxplot is the extremely high interest rate with loan amount level of 1,2,3. Further investigation on those extreme values are implemented.

Plot4: grade distributon



Plot 4 gives a view on the credit rating distribution on the dataset. Based on the plot result, most of borrowers are in graded at A to D, with most borrowers graded at level B and least of borrowers graded at level G. This distribution indicates that the borrowers that have transaction records are in good credit condition, which could potentially result from credit requirements and certification when applying for a loan.

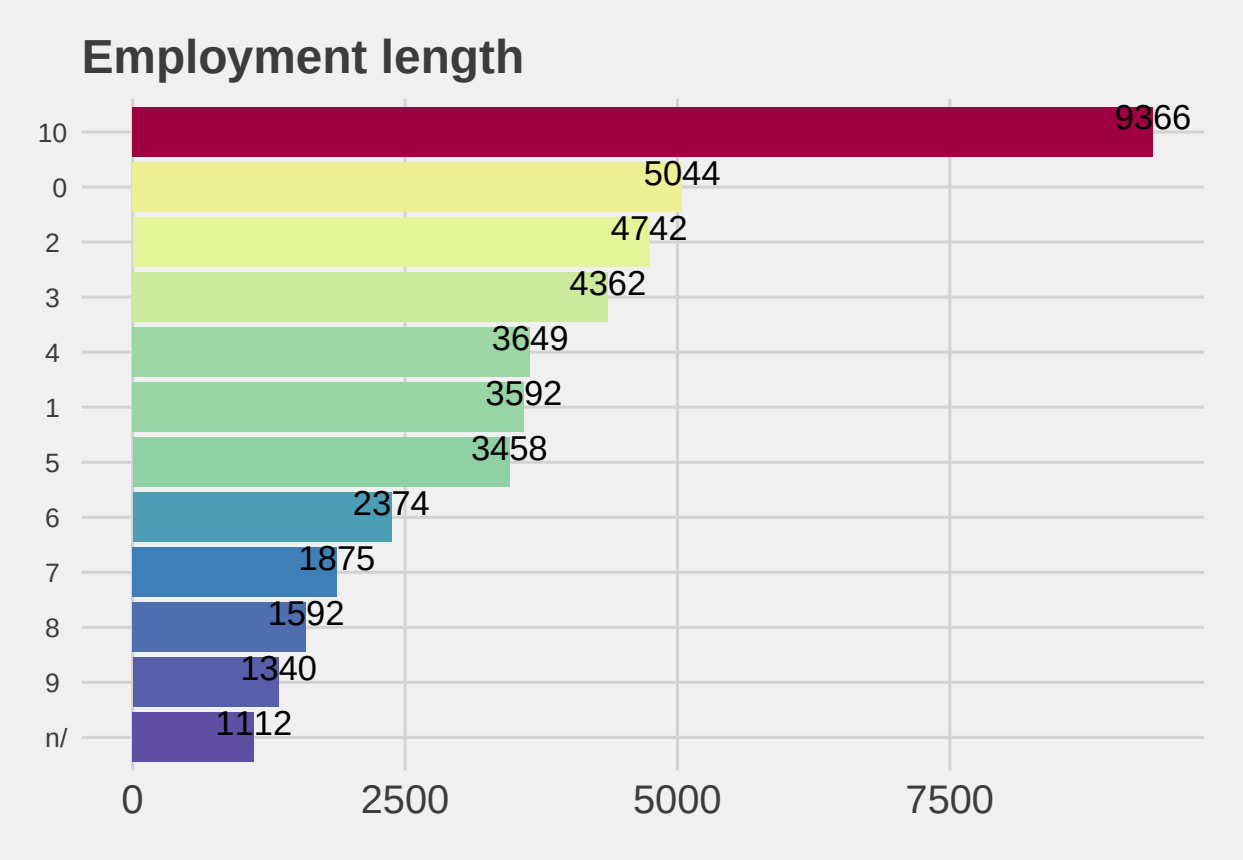
Plot5: grade related with interest rates



Plot 5 is the distribution of interest rate grouped by borrower's grade. Based on the plot result, people with worse credit grading have higher interest rate level. For example, most people with credit rating at A have interest rate at level 1, which is the lowest level of interest rate; while most people with credit rating at F and G have interest rate at level 4, which is the highest level of interest rate.

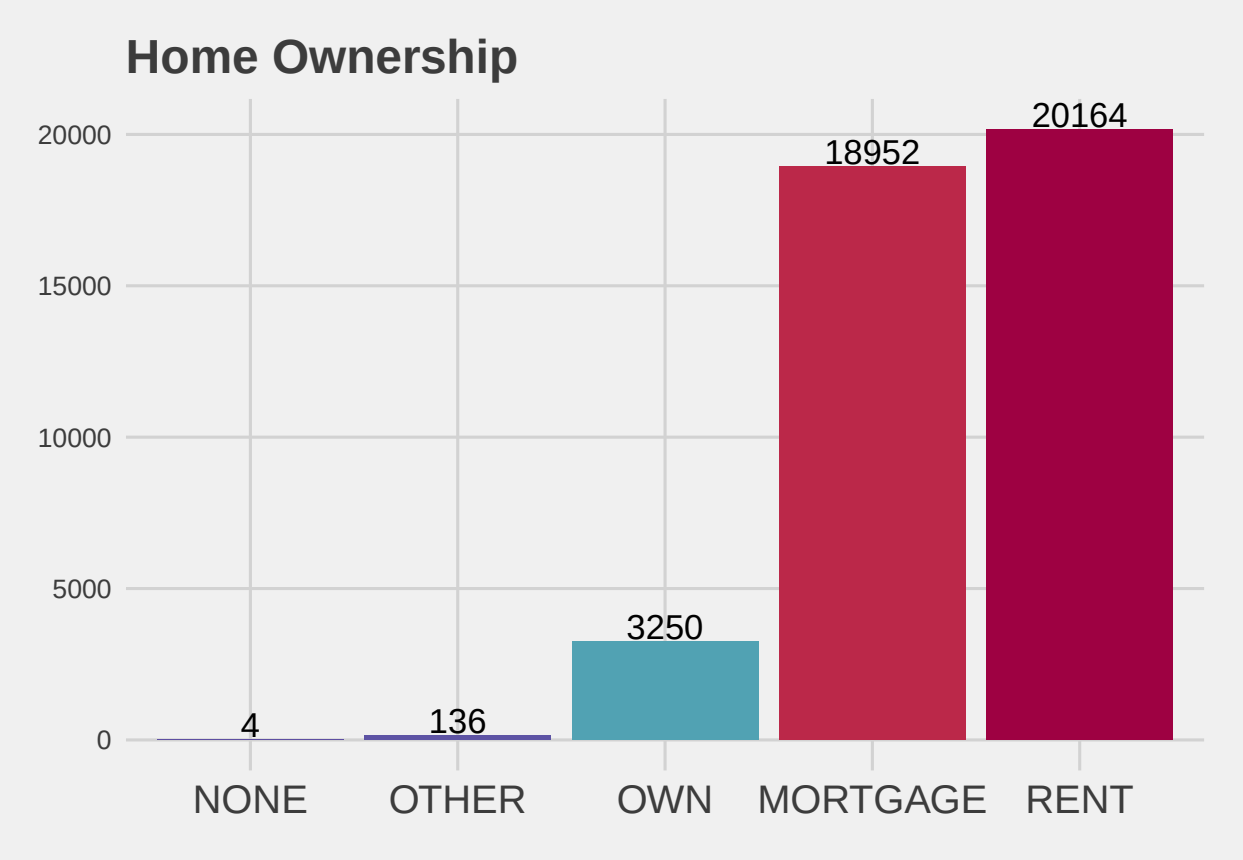
Plot 6:employment length

```
## Selecting by freq
```



Plot 6 is the distribution of employment length of borrowers. Based on the plot result, most of borrowers have been working for more than 10 years. However, there also comes a large number of borrowers that have no working experience. This finding might be caused by different reasons of borrowing. For example, people with long employment length might borrow money to purchase real estate, and people with no working experience could be student who are applying for student loan.

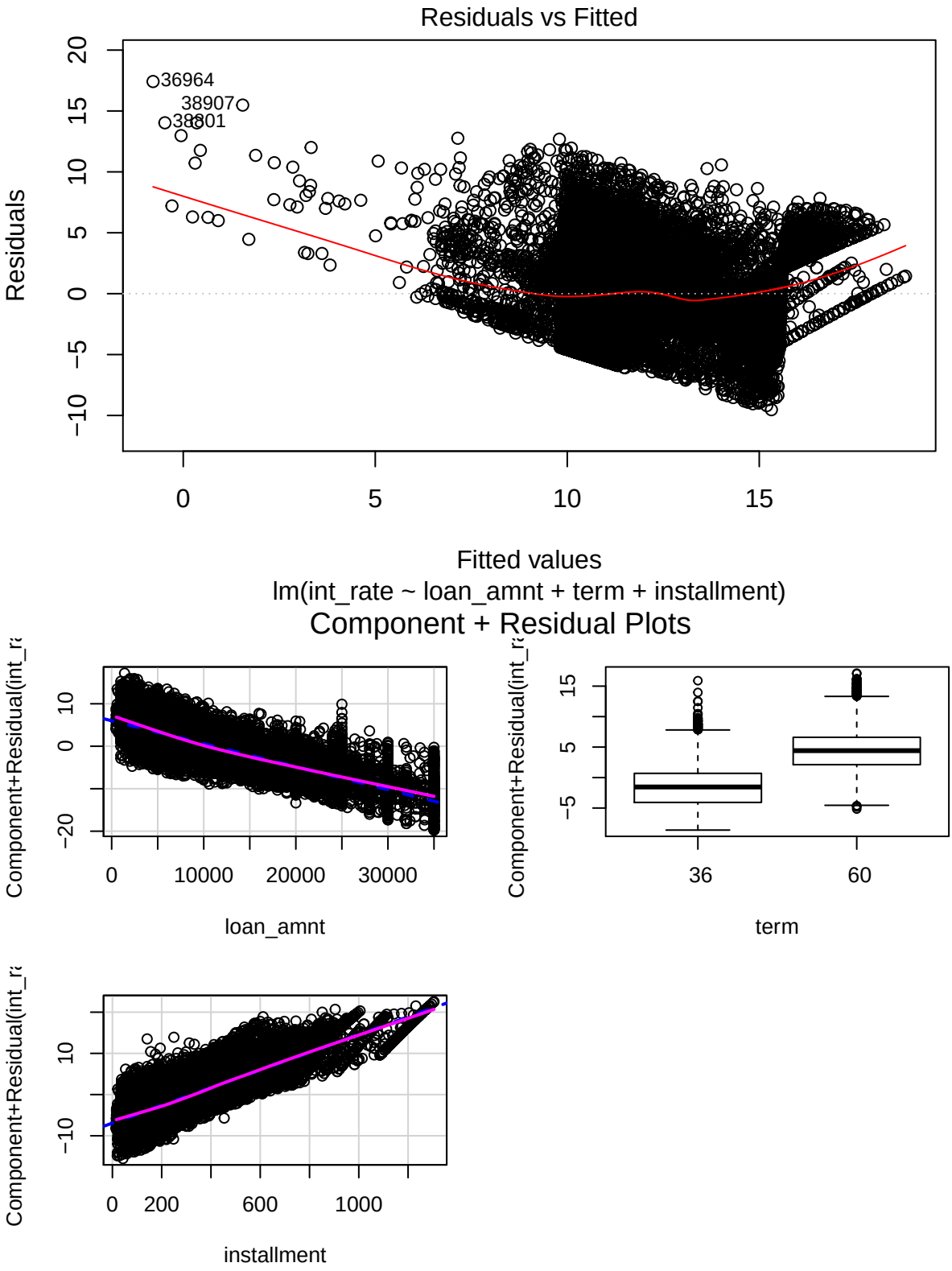
Plot 7: borrowers home ownership type



Plot 7 is the distribution of home ownership of borrowers. Home ownership is related with credit risk assessment in the sense of asset-backed loan. Also, the home ownership might indicate the reason for borrowing. Based on the plot result, most of borrowers rent a house or has a mortgage.

- Model fitting and results Model 1: simple linear model without confounding variables (personal information realted with interest rate)

```
##
## Call:
## lm(formula = int_rate ~ loan_amnt + term + installment, data = subdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5260 -2.4857 -0.0055  2.2210 17.4167
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.715e+00  2.849e-02  340.95   <2e-16 ***
## loan_amnt   -5.412e-04  8.048e-06  -67.25   <2e-16 ***
## term 60      5.886e+00  5.001e-02  117.70   <2e-16 ***
## installment  2.148e-02  2.681e-04   80.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.08 on 42502 degrees of freedom
## Multiple R-squared:  0.3102, Adjusted R-squared:  0.3101
## F-statistic:  6370 on 3 and 42502 DF,  p-value: < 2.2e-16
```

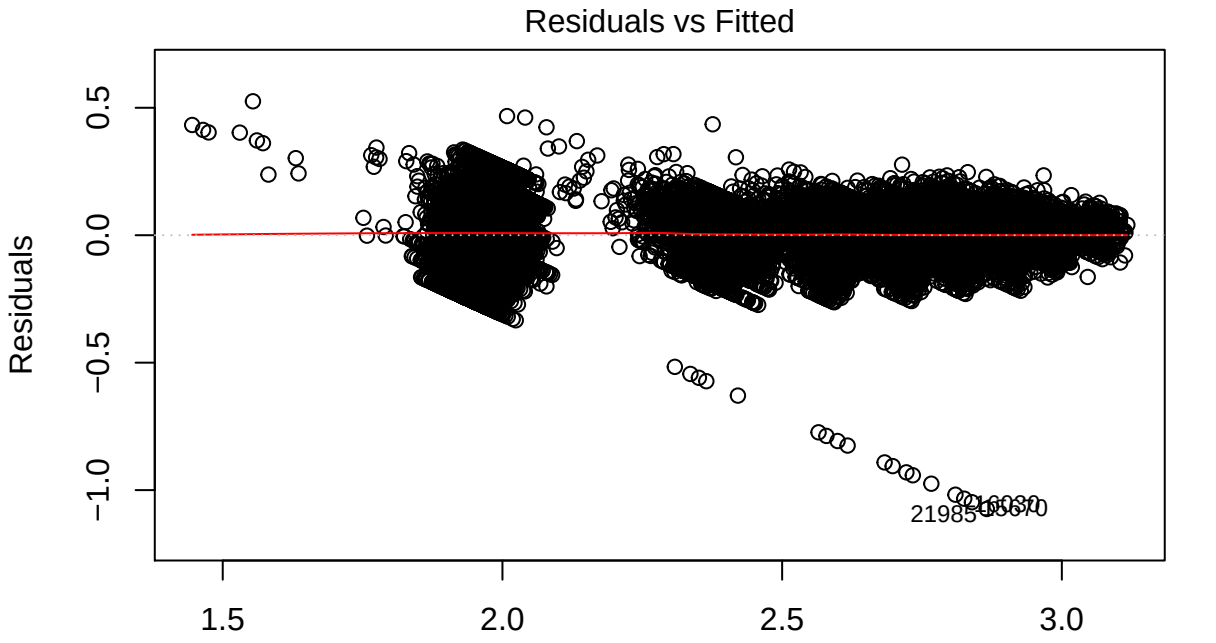


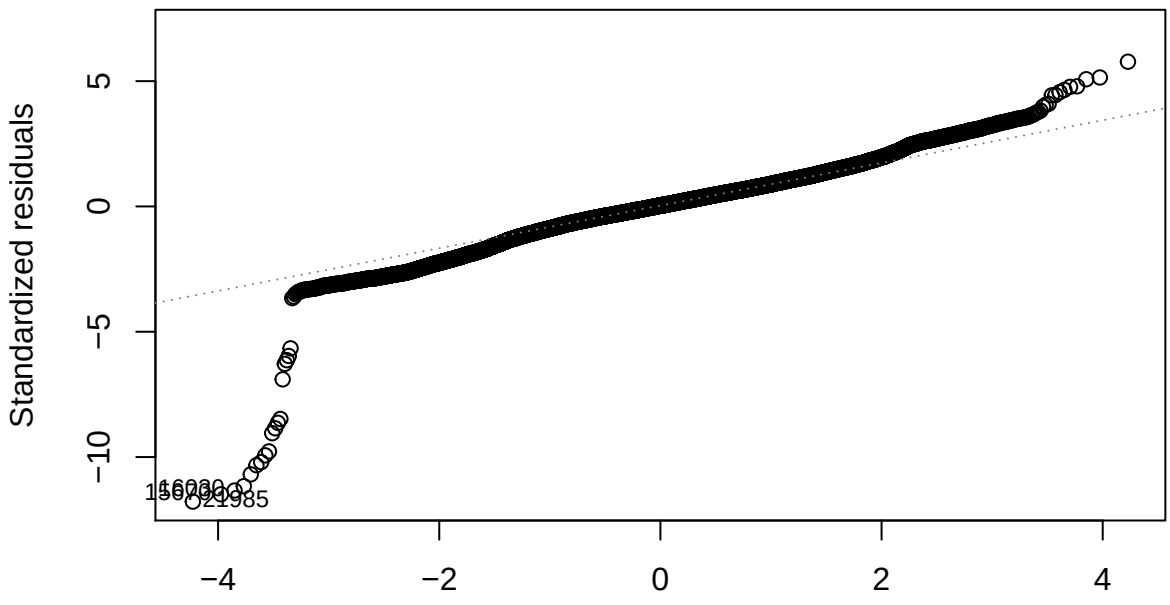
The first model (fit1) is trying to assess if the interest rate is strongly related with loan structure under linear regression, including loan amount, term and installment. Based on the model fitting results, loan amount, term and installment are all significantly correlated with interest rates. For every unit increase of log loan amount, the interest rate will decrease 20%. To check the assumption of the model, residual plots are provided. Component residual plots indicates that the variables of loan amount and installment align

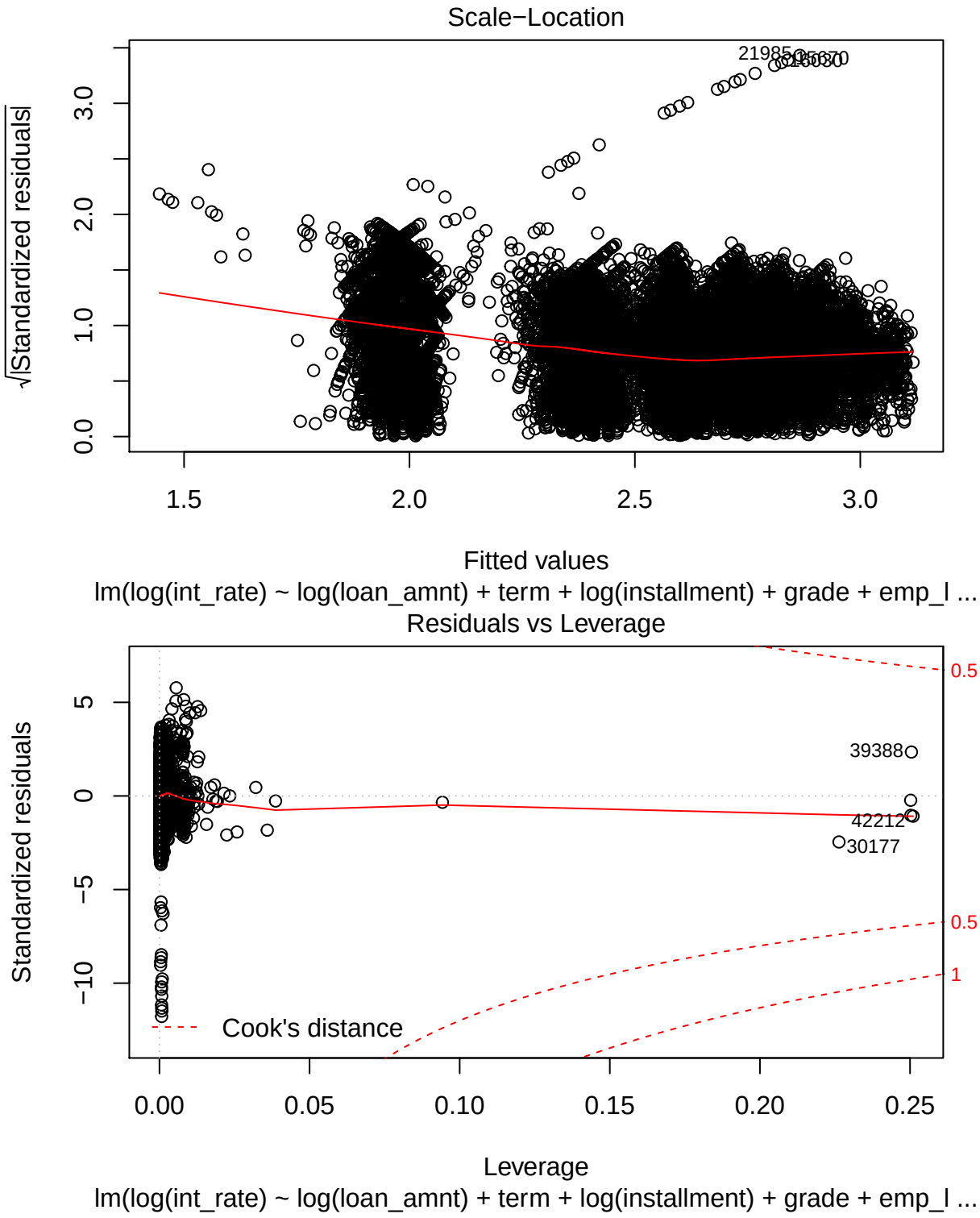
with the linear assumption. However, adjusted R square of fit 1 is only 0.25, indicating that the linear model does not capture enough information in the dataset.

Model 2: simple liner model with confounding variables (personal information related with interest rate)

```
##
## Call:
## lm(formula = log(int_rate) ~ log(loan_amnt) + term + log(installment) +
##     grade + emp_length + home_ownership + annual_inc + verification_status +
##     delinq_2yrs + fico_range_low + open_acc + recoveries, data = subdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07408 -0.04918  0.00304  0.05553  0.52544
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.611e+00  2.125e-02 169.901  < 2e-16 ***
## log(loan_amnt)         -2.234e-01  4.495e-03 -49.695  < 2e-16 ***
## term 60                 1.646e-01  2.174e-03  75.710  < 2e-16 ***
## log(installment)        2.505e-01  4.502e-03  55.654  < 2e-16 ***
## gradeB                  3.381e-01  1.494e-03 226.225  < 2e-16 ***
## gradeC                  4.991e-01  1.915e-03 260.614  < 2e-16 ***
## gradeD                  6.053e-01  2.309e-03 262.169  < 2e-16 ***
## gradeE                  6.763e-01  2.798e-03 241.717  < 2e-16 ***
## gradeF                  7.522e-01  3.613e-03 208.230  < 2e-16 ***
## gradeG                  8.041e-01  4.850e-03 165.799  < 2e-16 ***
## emp_length1            -1.640e-03  1.994e-03  -0.822 0.410863
## emp_length10           -3.902e-03  1.644e-03  -2.373 0.017637 *
## emp_length2            -1.731e-03  1.849e-03  -0.936 0.349201
## emp_length3            -3.715e-03  1.893e-03  -1.963 0.049636 *
## emp_length4            -2.058e-03  1.991e-03  -1.033 0.301404
## emp_length5            -3.618e-03  2.027e-03  -1.785 0.074319 .
## emp_length6            -5.660e-03  2.285e-03  -2.477 0.013241 *
## emp_length7            -1.922e-03  2.482e-03  -0.774 0.438725
## emp_length8            -5.327e-03  2.640e-03  -2.018 0.043641 *
## emp_length9            -1.711e-03  2.822e-03  -0.606 0.544343
## emp_lengthn/           -1.636e-02  3.054e-03  -5.355 8.59e-08 ***
## home_ownershipNONE     -7.574e-02  4.564e-02  -1.660 0.096988 .
## home_ownershipOTHER     3.785e-02  7.873e-03   4.807 1.53e-06 ***
## home_ownershipOWN       1.061e-02  1.754e-03   6.051 1.45e-09 ***
## home_ownershipRENT      5.041e-03  1.002e-03   5.030 4.92e-07 ***
## annual_inc              3.159e-08  7.317e-09   4.317 1.58e-05 ***
## verification_status1    3.285e-03  1.141e-03   2.879 0.003994 **
## verification_status2    1.523e-02  1.130e-03  13.473  < 2e-16 ***
## delinq_2yrs            -3.047e-03  8.883e-04  -3.429 0.000605 ***
## fico_range_low         -1.319e-03  2.151e-05 -61.324  < 2e-16 ***
## open_acc               -1.044e-03  1.037e-04 -10.063  < 2e-16 ***
## recoveries              2.265e-06  6.082e-07   3.725 0.000196 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09123 on 42474 degrees of freedom
## Multiple R-squared:  0.9202, Adjusted R-squared:  0.9202
## F-statistic: 1.581e+04 on 31 and 42474 DF,  p-value: < 2.2e-16
```


Residuals vs Fitted
Residuals
0.5
0.0
−0.5
−1.0
1.5
2.0
2.5
3.0
Fitted values
lm(log(int_rate) ~ log(loan_amnt) + term + log(installment) + grade + emp_l ...


Normal Q−Q
Standardized residuals
5
0
−5
−10
−4
−2
0
2
4
Theoretical Quantiles
lm(log(int_rate) ~ log(loan_amnt) + term + log(installment) + grade + emp_l ...



The second model (fit2) also uses the linear regression but incorporates borrowers' personal information in confounding variables. Based on the model fitting results, all the variables selected are significantly correlated with interest rate. For every unit increase of log loan amount, the interest rate will decrease 20%. The R square for fit 2 is 0.92, indicating that this model captures much more information than model 1. Furthermore, the improvement of R square indicates that personal information is important when determining the interest rate.

ANOVA test

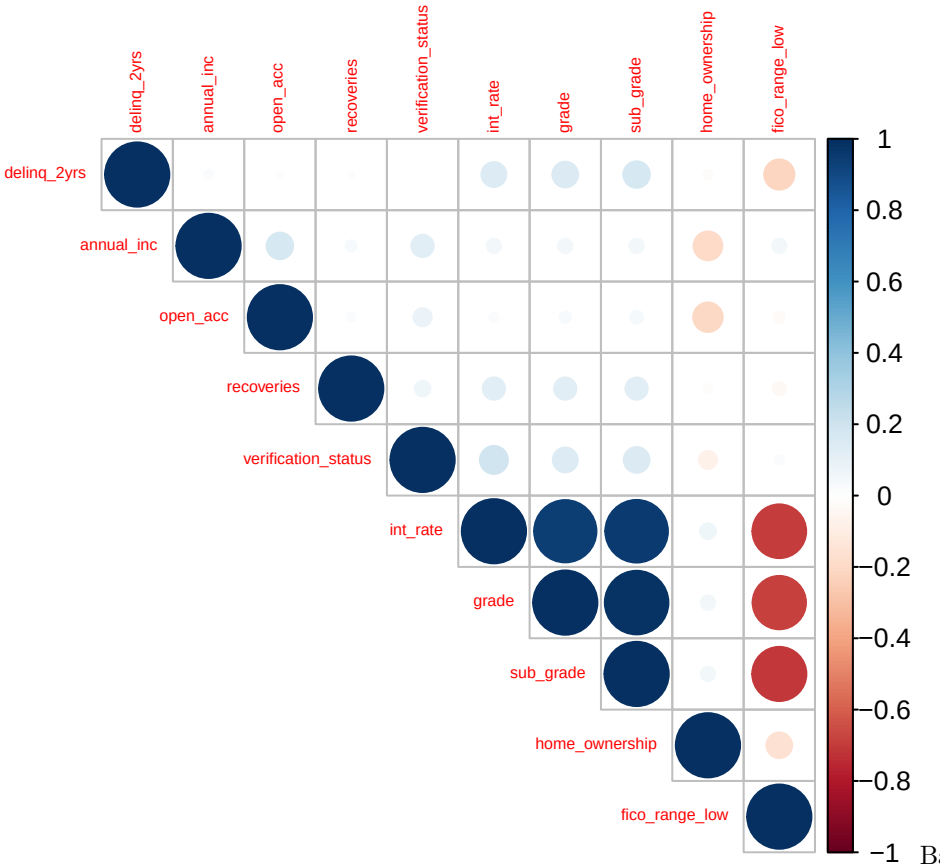
```
## Warning in anova.lmlist(object, ...): models with response
## '"log(int_rate)"' removed because response differs from model 1
```

```
## Analysis of Variance Table
##
## Response: int_rate
##                Df Sum Sq Mean Sq F value    Pr(>F)
## loan_amnt       1  49848   49848  5254.5 < 2.2e-16 ***
## term            1  70537   70537  7435.4 < 2.2e-16 ***
## installment     1  60898   60898  6419.4 < 2.2e-16 ***
## Residuals   42502 403203       9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

An ANOVA test is implemented for comparing fit 1 and fit2, RSS (residual sum of square) decreases from 3309 to 353, indicates that the model with borrowers' personal information is much better than the first model.

Correlation map to check if the confounding variables are correlated with the outcome (interst rate).

```
## 'data.frame':    42506 obs. of  10 variables:
##  $ int_rate           : num  10.7 15.3 16 13.5 12.7 ...
##  $ grade              : num  2 3 3 3 2 1 3 5 6 2 ...
##  $ sub_grade          : num  7 14 15 11 10 4 15 21 27 10 ...
##  $ home_ownership     : num  5 5 5 5 5 5 5 5 4 5 ...
##  $ annual_inc         : num  24000 30000 12252 49200 80000 ...
##  $ verification_status: num  3 2 1 2 2 2 1 2 2 3 ...
##  $ delinq_2yrs        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ fico_range_low     : int  735 740 735 690 695 730 690 660 675 725 ...
##  $ open_acc           : int  3 3 2 10 15 9 7 4 11 2 ...
##  $ recoveries         : num  0 123 0 0 0 ...
```



Based on the correlation plot, int_rate and grade, int_rate and fico_range have strong relationship.

Model 3: Lasso regression to choose the confounding varibales

```
## [1] 28479
```

```
## [1] 14027
```

```
## Warning in plot.window(...): "label" is not a graphical parameter
```

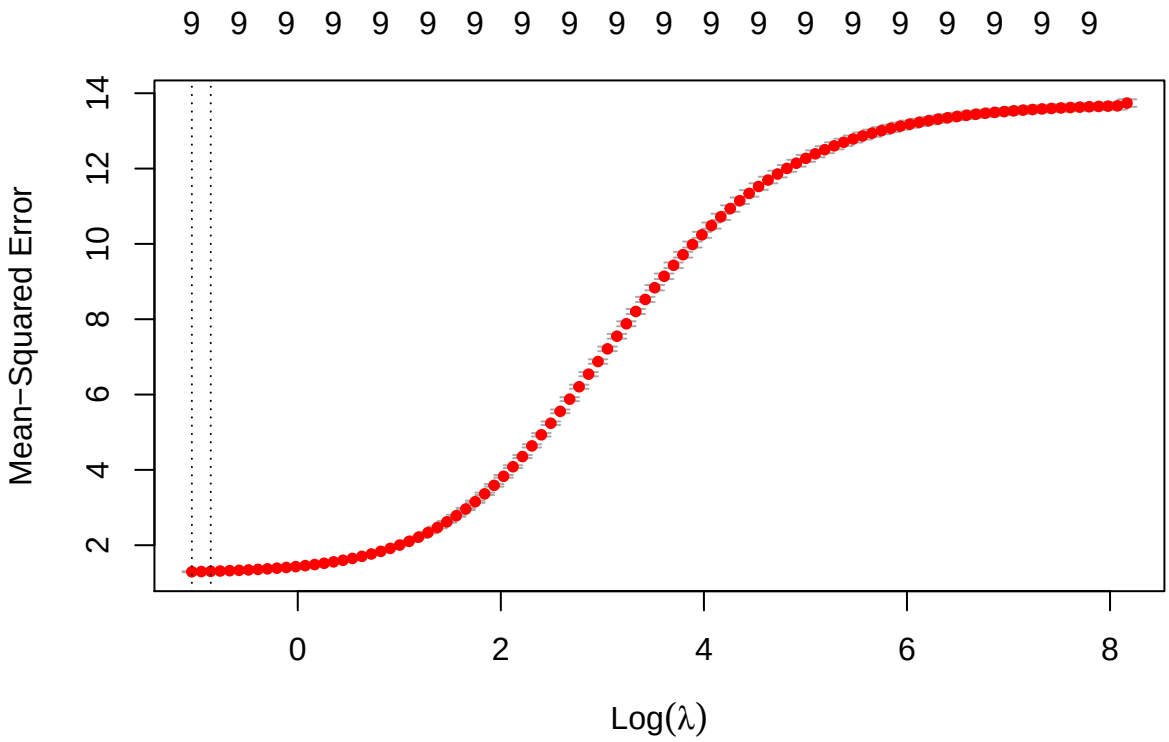
```
## Warning in plot.xy(xy, type, ...): "label" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "label" is not
## a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "label" is not
## a graphical parameter
```

```
## Warning in box(...): "label" is not a graphical parameter
```

```
## Warning in title(...): "label" is not a graphical parameter
```

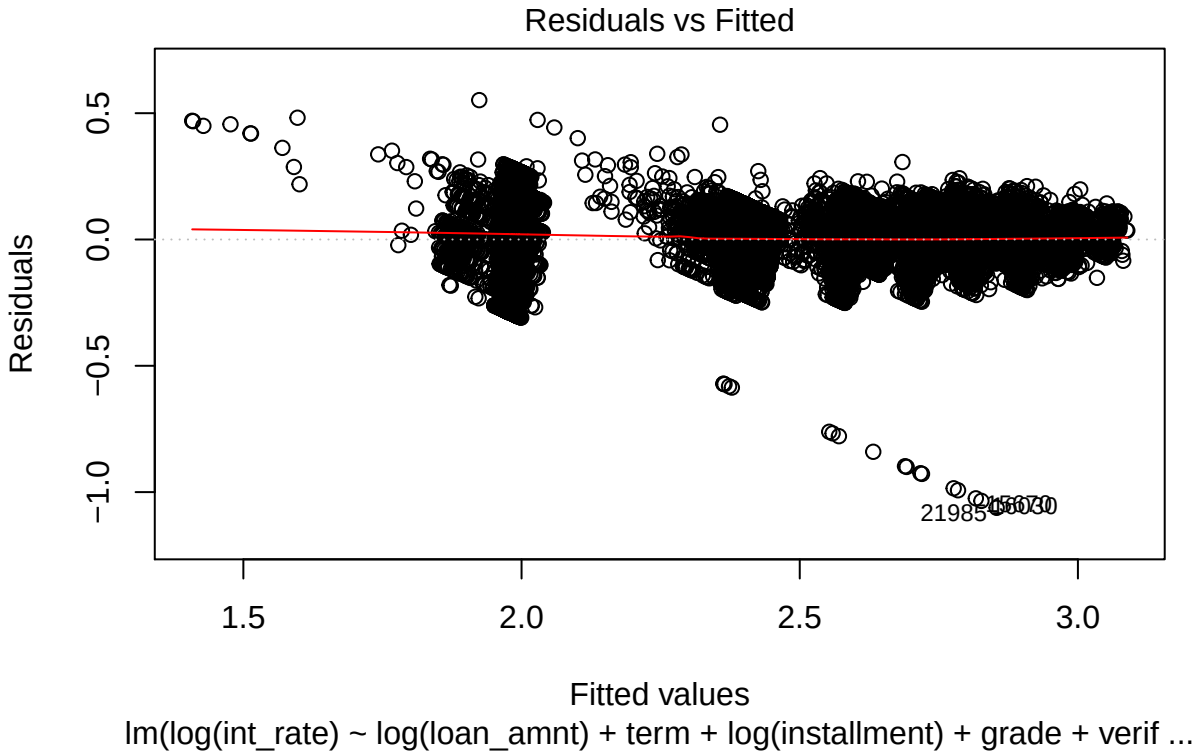



```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                                    1
## (Intercept)           1.194264e+01
## (Intercept)           .
## grade                 5.029704e-01
## sub_grade             3.647535e-01
## home_ownership        3.110576e-03
## annual_inc            1.823630e-07
## verification_status   2.156435e-01
## delinq_2yrs          -4.801485e-02
## fico_range_low       -7.898908e-03
## open_acc             -7.934209e-03
## recoveries            2.539326e-05
```

```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                                   s0
## (Intercept)           1.194264e+01
## (Intercept)           .
## grade                 5.029704e-01
## sub_grade             3.647535e-01
## home_ownership        3.110576e-03
## annual_inc            1.823630e-07
## verification_status   2.156435e-01
## delinq_2yrs          -4.801485e-02
## fico_range_low       -7.898908e-03
## open_acc             -7.934209e-03
## recoveries            2.539326e-05
```

```
##
## Call:
## lm(formula = log(int_rate) ~ log(loan_amnt) + term + log(installment) +
```

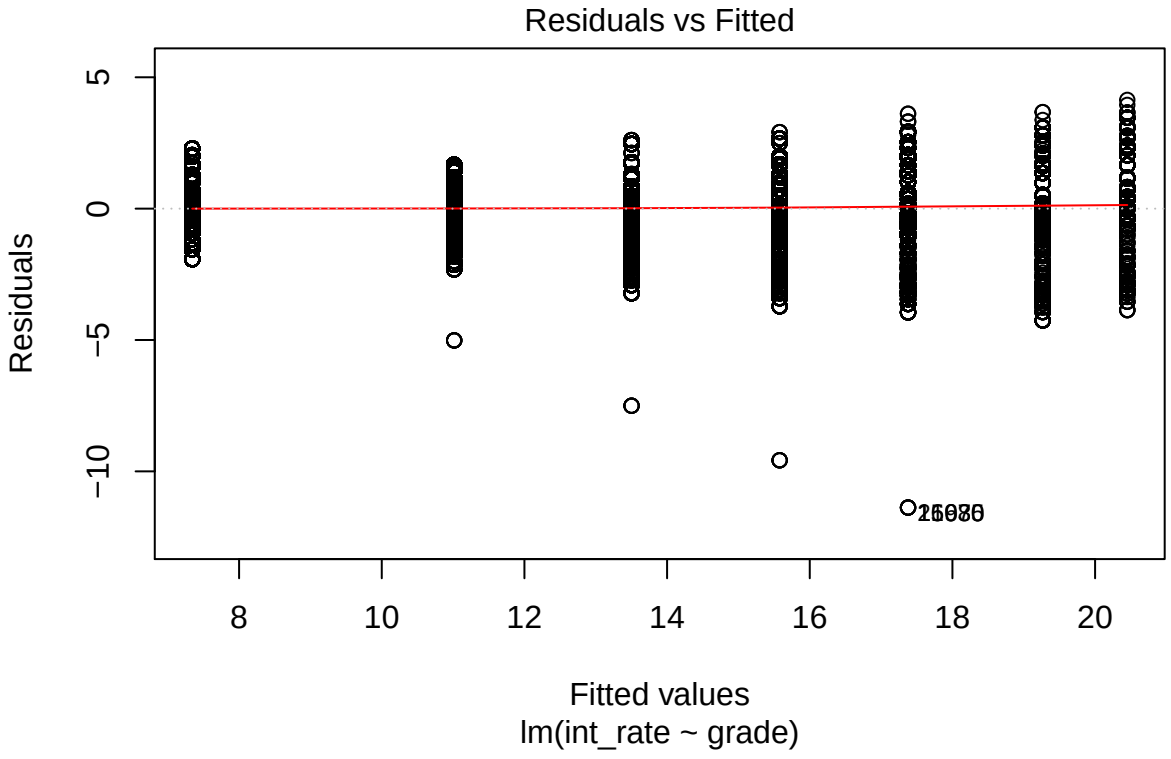
```
##     grade + verification_status + delinq_2yrs + recoveries, data = subdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06218 -0.05140  0.00428  0.05947  0.55169
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.787e+00  1.689e-02 164.983  < 2e-16 ***
## log(loan_amnt)         -2.505e-01  4.668e-03 -53.650  < 2e-16 ***
## term 60                 1.410e-01  2.232e-03  63.184  < 2e-16 ***
## log(installment)        2.625e-01  4.698e-03  55.871  < 2e-16 ***
## gradeB                  3.874e-01  1.328e-03 291.677  < 2e-16 ***
## gradeC                  5.781e-01  1.506e-03 383.931  < 2e-16 ***
## gradeD                  7.049e-01  1.747e-03 403.587  < 2e-16 ***
## gradeE                  7.927e-01  2.182e-03 363.228  < 2e-16 ***
## gradeF                  8.804e-01  3.104e-03 283.675  < 2e-16 ***
## gradeG                  9.388e-01  4.533e-03 207.113  < 2e-16 ***
## verification_status1    7.800e-03  1.188e-03   6.564 5.31e-11 ***
## verification_status2    1.436e-02  1.176e-03  12.203  < 2e-16 ***
## delinq_2yrs             3.393e-03  9.192e-04   3.692 0.000223 ***
## recoveries              2.109e-06  6.357e-07   3.318 0.000907 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09539 on 42492 degrees of freedom
## Multiple R-squared:  0.9128, Adjusted R-squared:  0.9127
## F-statistic: 3.42e+04 on 13 and 42492 DF,  p-value: < 2.2e-16
```




```
##  [1] id                  loan_amnt           term
```

```
##  [4] int_rate            installment         grade
##  [7] sub_grade           emp_length          home_ownership
## [10] annual_inc          verification_status issue_d
## [13] delinq_2yrs         fico_range_low      open_acc
## [16] recoveries          int_level           loan_amnt_level
## <0 rows> (or 0-length row.names)
```

```
## [1] 0.9391812
```



The third model is to choose confounding variables by Lasso regression. Based on the model results, variables of annual income, recoveries, fico range have been shrinking to around zero. Thus, confounding variables are now left with grade, verification status and delinquency to be fitted to linear regression model. The model named "fittest" is the linear model with variables selected by lasso regression. On the one hand, with R square at 0.91, the linear model still captures most of the information in the dataset. On the other hand, two important findings are shown in the residual plot. First, the residuals are grouped. Second, outliners at the button of the residual plot tends to have some pattern. To investigate more about the outliners, those outliners are filtered. One important finding is that all those outliners have interest rate at 6%, no matter how the other factors are varied. Further concerns about the same 6% are recommended. The grouped pattern of residual is caused by different credit grades, with a strong correlation between grade and interest rate. Thus, the mixed effect model is implemented rather than simple linear regression model.

Model 4:multilevel linear regression model with varying intercept

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : unable to evaluate scaled gradient
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge: degenerate Hessian with 1
## negative eigenvalues
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## log(int_rate) ~ log(loan_amnt) + term + installment + (1 | grade) +
```

```
##     (1 | verification_status) + (1 | delinq_2yrs)
##    Data: subdata
##
## REML criterion at convergence: -76712.6
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -10.9437  -0.5408   0.0437   0.6292   2.9942
##
## Random effects:
##  Groups              Name        Variance  Std.Dev.
##  delinq_2yrs         (Intercept) 1.709e-05 0.004134
##  grade               (Intercept) 1.177e-01 0.343007
##  verification_status (Intercept) 3.650e-05 0.006041
##  Residual                        9.602e-03 0.097992
## Number of obs: 42506, groups:
## delinq_2yrs, 12; grade, 7; verification_status, 3
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)     2.853e+00  1.303e-01   21.89
## log(loan_amnt) -3.009e-02  1.564e-03  -19.24
## term 60         5.529e-02  1.420e-03   38.92
## installment     1.485e-04  5.462e-06   27.20
##
## Correlation of Fixed Effects:
##             (Intr) lg(l_) term60
## log(ln_mnt) -0.095
## term 60      0.045 -0.515
## installment  0.081 -0.887  0.486
## convergence code: 0
## unable to evaluate scaled gradient
## Model failed to converge: degenerate  Hessian with 1 negative eigenvalues
```

Model 5: multilevel linear regression model with varying slopes

```
## boundary (singular) fit: see ?isSingular
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## log(int_rate) ~ log(loan_amnt) + term + installment + (1 + log(loan_amnt) |
##     grade) + (1 + log(loan_amnt) | verification_status) + (1 +
##     log(loan_amnt) | delinq_2yrs)
##    Data: subdata
##
## REML criterion at convergence: -76878.8
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -11.0944  -0.5380   0.0446   0.6224   3.0189
##
## Random effects:
##  Groups              Name           Variance  Std.Dev. Corr
##  delinq_2yrs         (Intercept)    3.001e-04 0.017325
##                      log(loan_amnt) 2.163e-06 0.001471 -1.00
```

```
##  grade               (Intercept)    7.442e-02 0.272808
##                      log(loan_amnt) 1.199e-04 0.010952 0.16
##  verification_status (Intercept)    1.145e-03 0.033833
##                      log(loan_amnt) 1.079e-05 0.003285 -0.99
##  Residual                           9.560e-03 0.097777
## Number of obs: 42506, groups:
## delinq_2yrs, 12; grade, 7; verification_status, 3
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)     2.808e+00  1.067e-01   26.32
## log(loan_amnt) -2.549e-02  5.039e-03   -5.06
## term 60         5.609e-02  1.492e-03   37.60
## installment     1.481e-04  5.903e-06   25.09
##
## Correlation of Fixed Effects:
##             (Intr) lg(l_) term60
## log(ln_mnt) -0.021
## term 60      0.076 -0.211
## installment  0.120 -0.328  0.543
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

On model 4, for each increase in the level of grade, the intercept will increase 0.1177, causing 12% increase of interest rate keeping other variables constant. For the fixed effect, each unit increase in the log loan amount corresponds to 3% decrease of the interest rate. On model 5, for each increase in the grade level, the intercept will increase 0.003, and the slope of log loan amount will increase 0.00002. For the fixed effect, each unit increase in the log loan amount corresponds to 3% decrease of the interest rate. There is not big difference between adding the random effect to varying slope or not.

Model 6:Beyasian multilevel linear regression model (varying intercept)

#fit6 <- stan_lmer(log(int_rate)~log(loan_amnt)+term+installment+(1|grade)+(1|verification_status)+(1|delinq_2yrs),data
#summary(fit6)

Model7: Beyasian multilevel linear regression model (varying intercept and slope)

#fit7 <- stan_lmer(log(int_rate)~log(loan_amnt)+term+installment+(1+log(loan_amnt)|grade)+(1+log(loan_amnt)|verifica
#summary(fit7)

Model8: Categorical regression model

```
## Call:
## polr(formula = as.factor(int_level) ~ loan_amnt_level + as.factor(term) +
##     installment, data = subdata, Hess = TRUE)
##
## Coefficients:
##                         Value Std. Error t value
## loan_amnt_level2   -0.802156  0.0287179  -27.93
## loan_amnt_level3   -1.551257  0.0393443  -39.43
## loan_amnt_level4   -2.540464  0.0636847  -39.89
## as.factor(term) 60  2.187453  0.0265864   82.28
## installment         0.005826  0.0001122   51.90
##
## Intercepts:
##     Value    Std. Error t value
## 1|2  -0.0085   0.0229    -0.3693
```

```
## 2|3   1.2272   0.0234    52.3840
## 3|4   2.5213   0.0259    97.1654
##
## Residual Deviance: 107672.44
## AIC: 107688.44
```

```
## Waiting for profiling to be done...
```

```
##   loan_amnt_level2   loan_amnt_level3   loan_amnt_level4
##         0.44836111         0.21198138         0.07882983
## as.factor(term) 60        installment
##         8.91248257         1.00584277
```

```
##                            OR      2.5 %    97.5 %
## loan_amnt_level2   0.44836111 0.42379711 0.4743177
## loan_amnt_level3   0.21198138 0.19618450 0.2289977
## loan_amnt_level4   0.07882983 0.06951141 0.0893653
## as.factor(term) 60 8.91248257 8.45878011 9.3922038
## installment        1.00584277 1.00562233 1.0060644
```

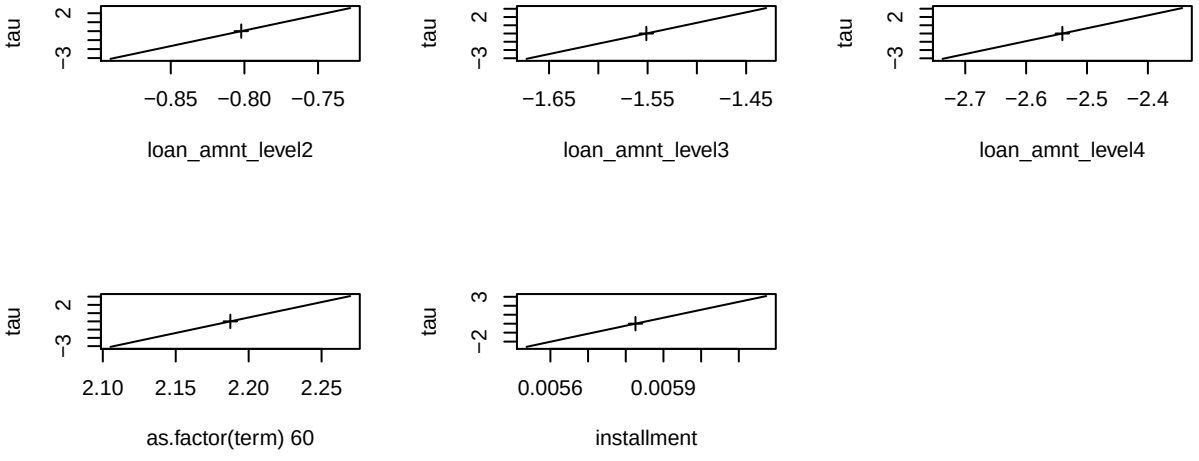
```
## Call:
## polr(formula = as.factor(int_level) ~ loan_amnt_level + as.factor(term) +
##     installment, data = subdata, Hess = TRUE, method = "probit")
##
## Coefficients:
##                       Value Std. Error t value
## loan_amnt_level2   -0.4548   0.016796   -27.1
## loan_amnt_level3   -0.8709   0.022454   -38.8
## loan_amnt_level4   -1.4172   0.036651   -38.7
## as.factor(term) 60  1.3017   0.015241    85.4
## installment         0.0033   0.000062    53.2
##
## Intercepts:
##     Value   Std. Error t value
## 1|2  -0.029   0.013     -2.199
## 2|3   0.723   0.014     53.127
## 3|4   1.491   0.015    102.626
##
## Residual Deviance: 107474.95
## AIC: 107490.95
```

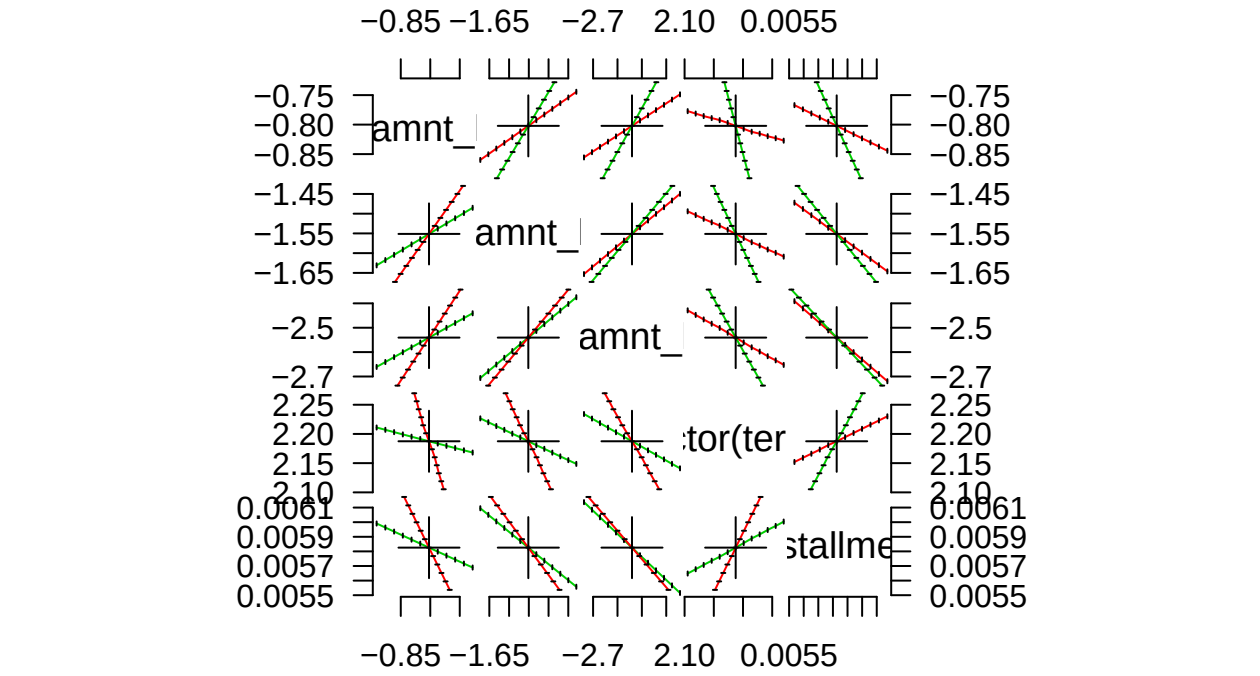
```
## Call:
## polr(formula = as.factor(int_level) ~ loan_amnt_level + as.factor(term) +
##     installment, data = subdata, Hess = TRUE, method = "logistic")
##
## Coefficients:
##                       Value Std. Error t value
## loan_amnt_level2   -0.80216   0.028718   -27.9
## loan_amnt_level3   -1.55126   0.039344   -39.4
## loan_amnt_level4   -2.54046   0.063685   -39.9
## as.factor(term) 60  2.18745   0.026586    82.3
## installment         0.00583   0.000112    51.9
##
## Intercepts:
##     Value   Std. Error t value
## 1|2  -0.008   0.023     -0.369
```

```
## 2|3   1.227   0.023     52.384
## 3|4   2.521   0.026     97.165
##
## Residual Deviance: 107672.44
## AIC: 107688.44
```

```
## Call:
## polr(formula = as.factor(int_level) ~ loan_amnt_level + as.factor(term) +
##     installment, data = subdata, Hess = TRUE, method = "cloglog")
##
## Coefficients:
##                         Value Std. Error t value
## loan_amnt_level2   -0.51933   1.79e-02   -29.0
## loan_amnt_level3   -1.03363   2.49e-02   -41.5
## loan_amnt_level4   -1.66001   4.01e-02   -41.4
## as.factor(term) 60  1.40266   1.80e-02    77.9
## installment         0.00373   6.86e-05    54.4
##
## Intercepts:
##     Value   Std. Error t value
## 1|2  -0.533   0.015    -35.120
## 2|3   0.382   0.014     27.443
## 3|4   1.154   0.014     82.103
##
## Residual Deviance: 108226.24
## AIC: 108242.24
```

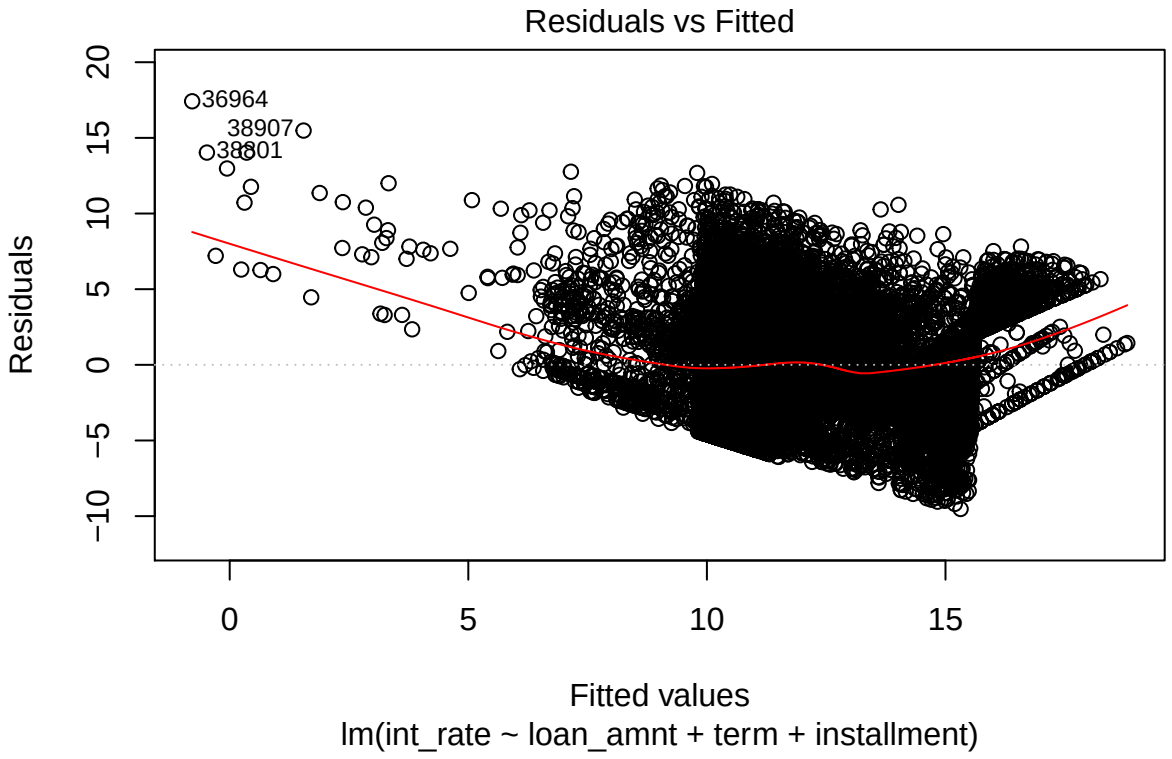
```
##                            2.5 %       97.5 %
## loan_amnt_level2   -0.858500410 -0.745877880
## loan_amnt_level3   -1.628697437 -1.474043644
## loan_amnt_level4   -2.666264278 -2.415022831
## as.factor(term) 60  2.135204496  2.239880636
## installment         0.005606579  0.006046076
```



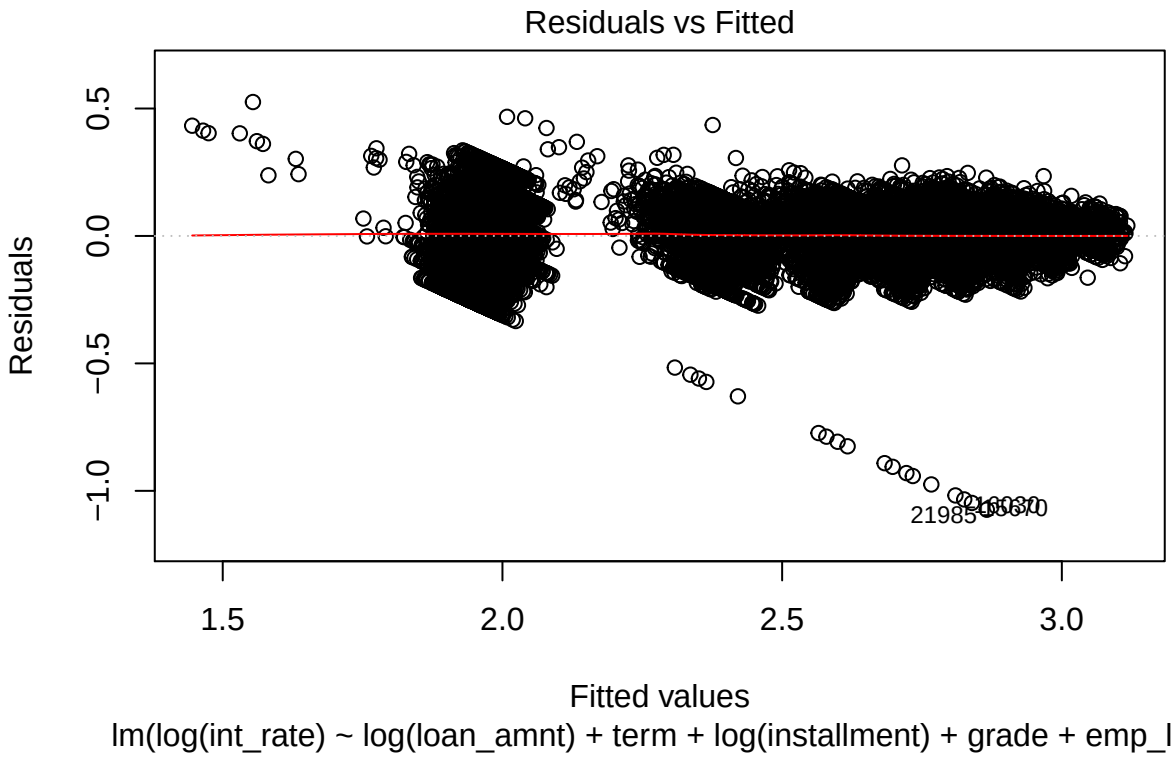


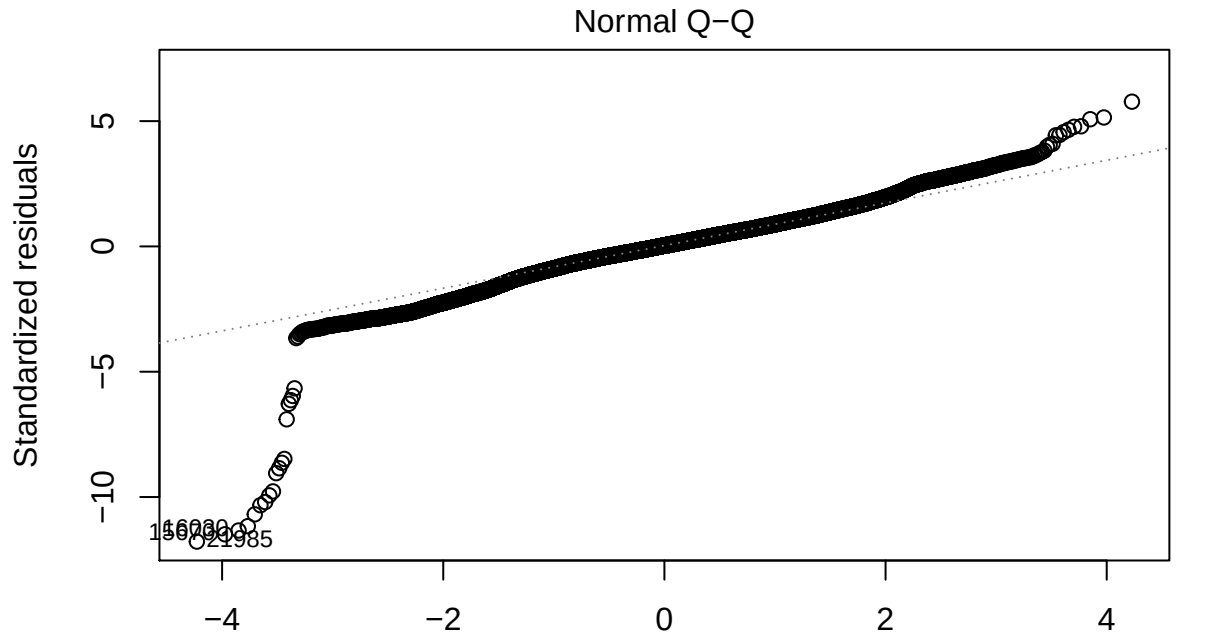
Model 8 is the categorical regression model to investigate the difference between each level of interest rate. At level 2 loan amount, the odds of interest rate of level 1 compared to level 2 are 0.45 greater. At level 3 loan amount, the odds of interest rate of level 1 compared to level 2 are 0.21 greater. At level 4 loan amount, the odds of interest rate of level 1 compared to level 2 are 0.08 greater. For installment, when the installment amount moves 1 unit, the odds of interest rate level moving from "level 1" to other levels are nearly the same.

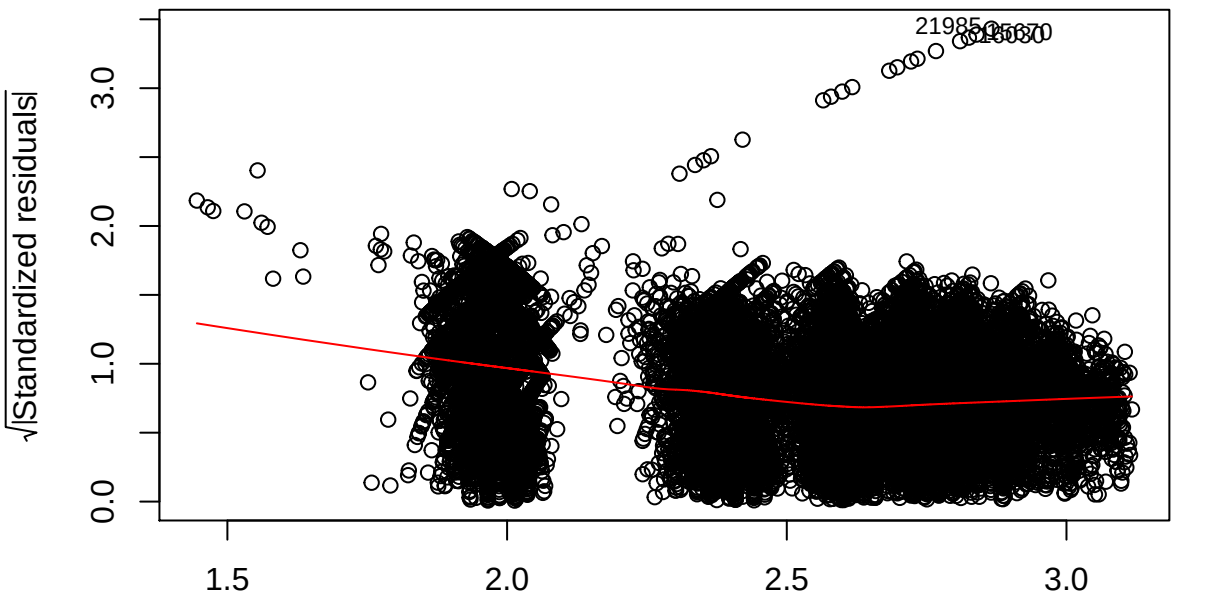
- Model validation In this project, model validation is implemented on confirming the alignment of assumptions and accuracy of predicted results. Model 1 and model 2: qqplots, residual plots and R square.

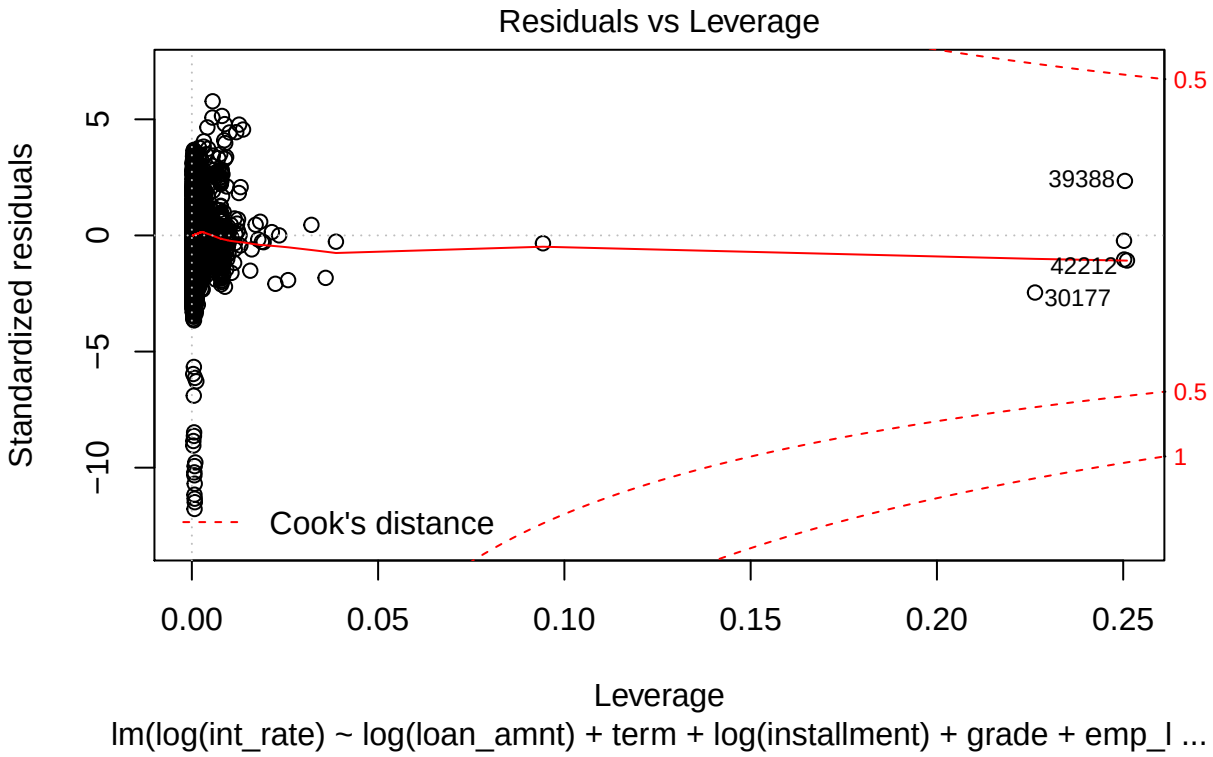


Based on the residual plot, the assumption of normality of residuals is not satisfied. Outliers in the left side is another problem. After investigation the residuals, those outliers are a set of transactions with 6% interest rate. With R square at 0.25, the model is not accurate enough.



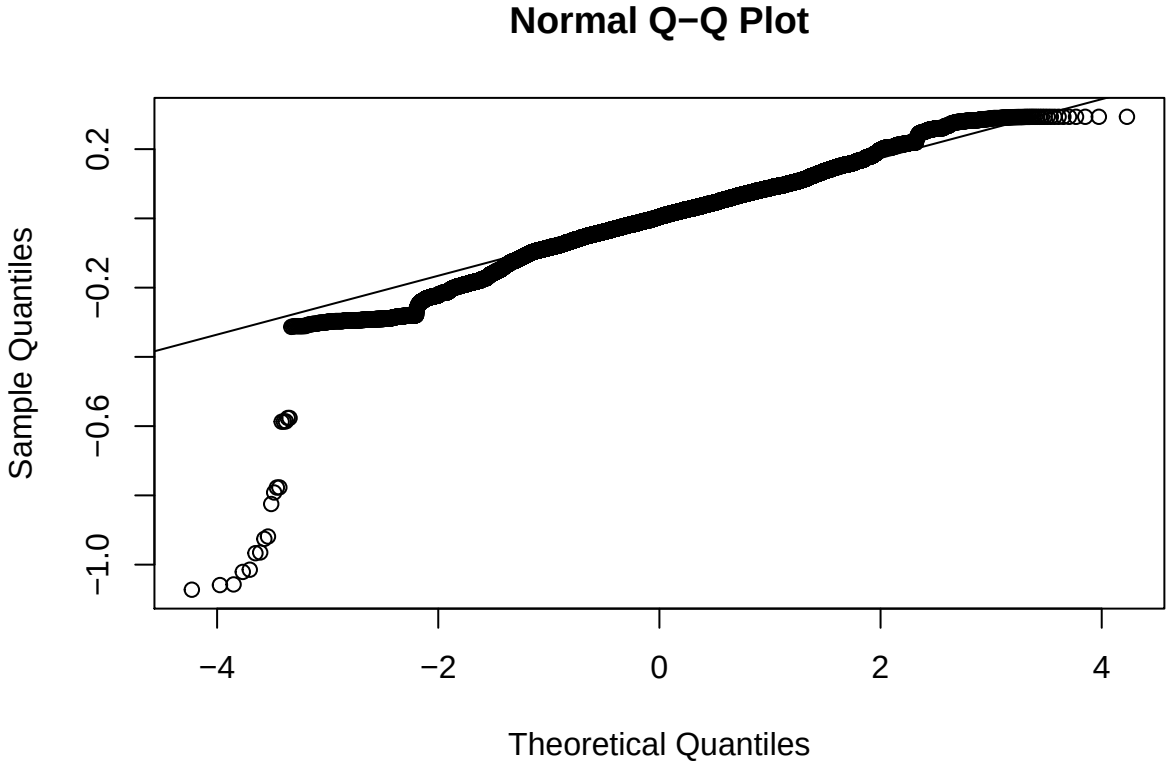

Normal Q−Q
Standardized residuals
5
0
−5
−10
−4
−2
0
2
4
Theoretical Quantiles
lm(log(int_rate) ~ log(loan_amnt) + term + log(installment) + grade + emp_l ...

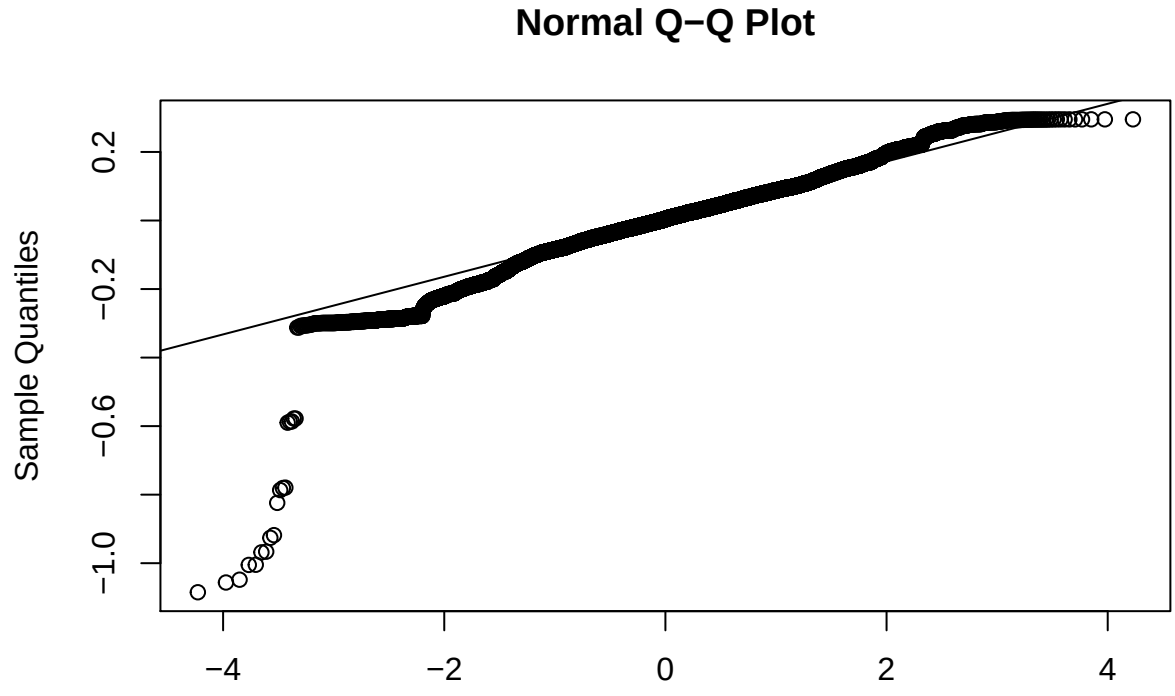

Scale−Location
√|Standardized residuals|
3.0
2.0
1.0
0.0
1.5
2.0
2.5
3.0
Fitted values
lm(log(int_rate) ~ log(loan_amnt) + term + log(installment) + grade + emp_l ...

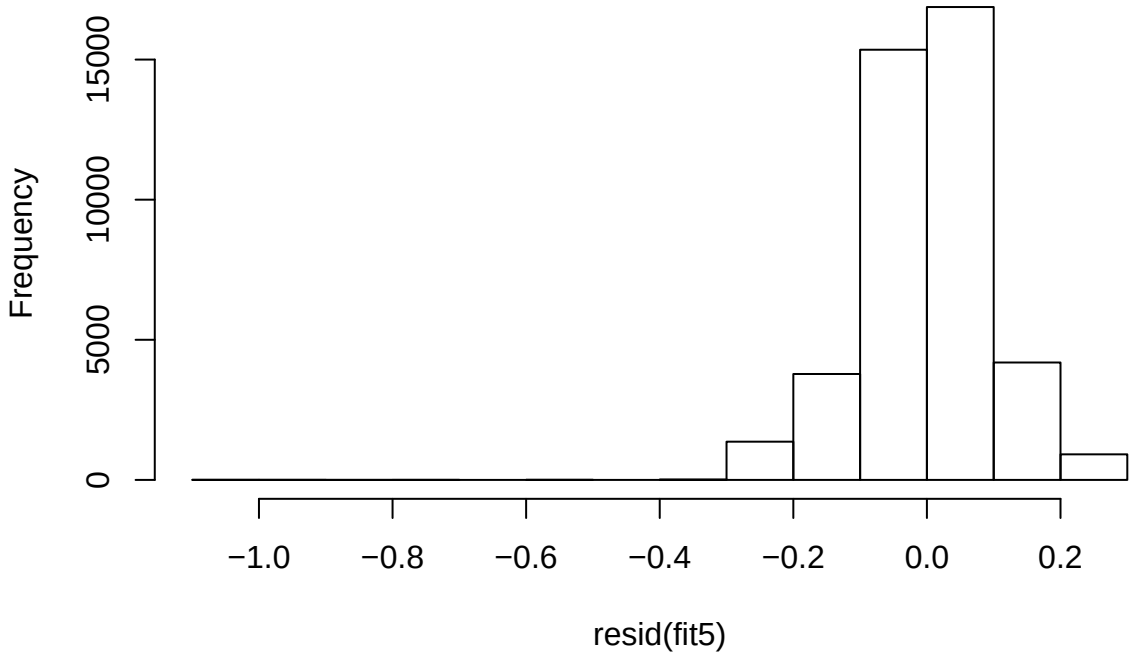


For the second model, the assumption of normality of residuals is satisfied with most of data equally distributed within the zero-horizontal line. The assumption of normality of residuals is satisfied with most of dots plotted as a straight line in normal Q-Q plot. With R square at 0.92, the model is expected to make good determination.

Model 4 and Model 5: qq plots, residual plots, AIC

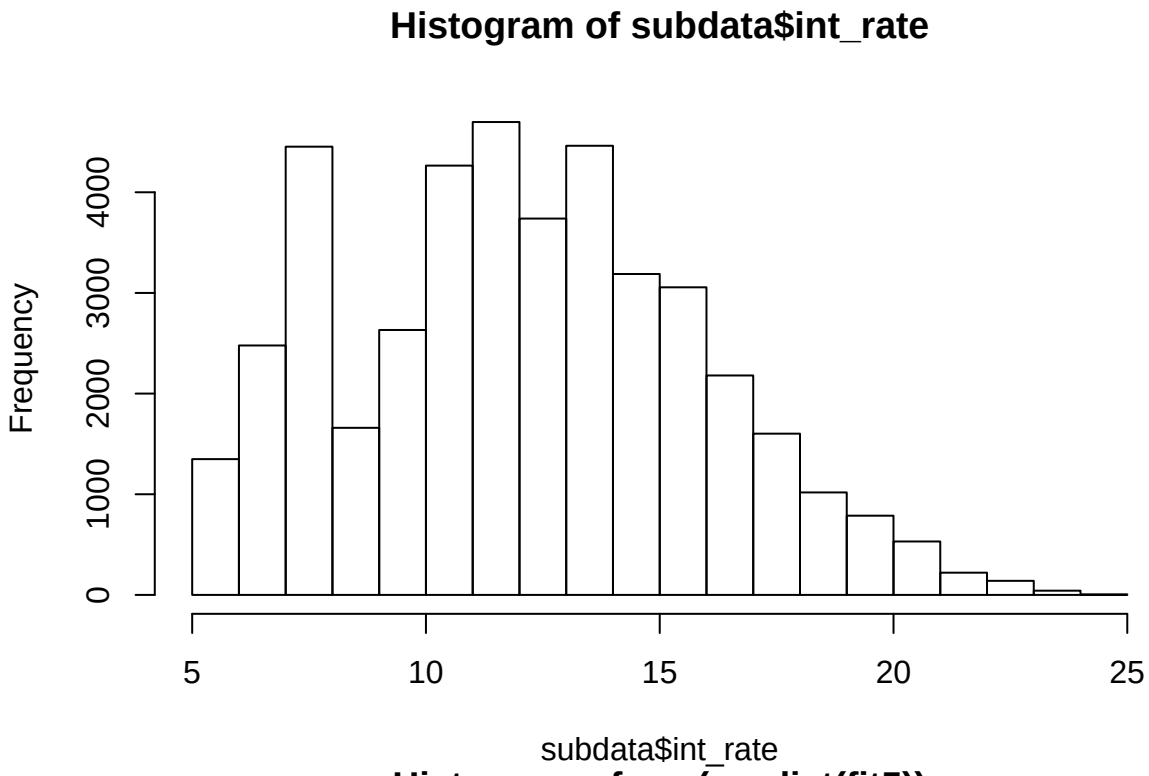


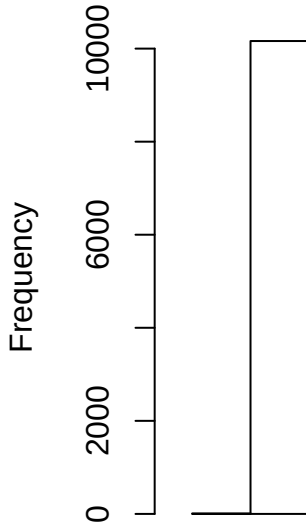


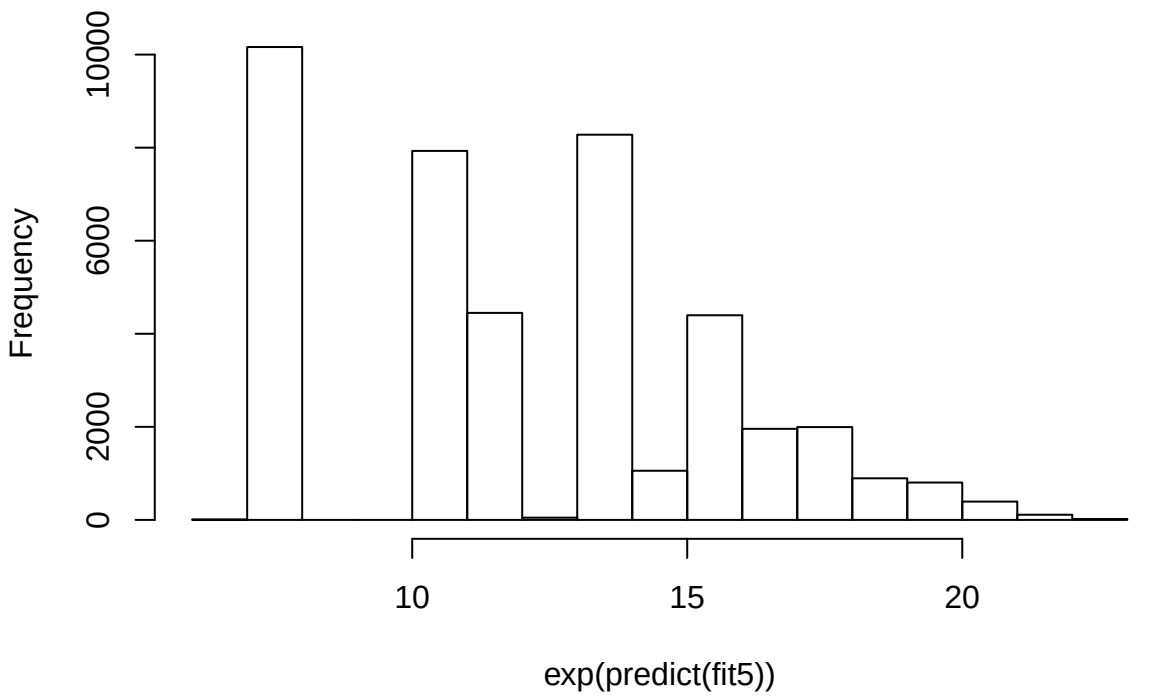


The normal Q-Q plot for model 4 and model 5 indicate that these two models meet the assumption of normality of residuals.

```
##           aic4      aic5
## [1,] -76696.64 -76850.82
```







With lower level of AIC, model 4 tends to perform better than model 5. However, comparing the histogram of predicted value to raw data, the accuracy of the model still needs improvement.

- Summary of key findings

1. Credit grades is the most important factor to determine the interest rate. A certain level of credit grade would have a certain range of interest rate. People with lower credit grading have higher interest rate.

2. Loan amount has negative correlation with interest rate. For every unit increase of log loan amount, the interest rate will decrease 20%.
3. There is a group of people, with different credit rating grades, have the same interest rate at 6%, which is relatively low compared to the interest rate for whole population. Even for people with grade F, enjoys the interest rate of 6%. This finding deserves further consideration since it corporates huge default risks.

- Limitations

1. The accuracy of all the model is not high enough to make predictions.
2. Those outliers with 6% interest rate are not removed, which might impact the model validation.

- Future direction

1. Investigate transactions with 6% interest rater. Due dilligence is suggested to make sure the the interest rate is reasonable and the default risk is not too high. Also, investigate why the interest rate is the same for all those transactions, in a relatively low level.
2. Determine the interest rate is the pricing stage for lending. How to control the risk afterwards is considered to be the next step.