Introduction to Statistics Project

Sophie Zheng

July 27, 2016

Model 1

```
model <- lm(formula= int_rate ~  loan_amnt + revol_util + dti + tot_cur_bal +
              tot_hi_cred_lim + home_ownership + total_bc_limit + term +
              percent_bc_gt_75 + purpose + installment, data = Loan)

summary(model)
```

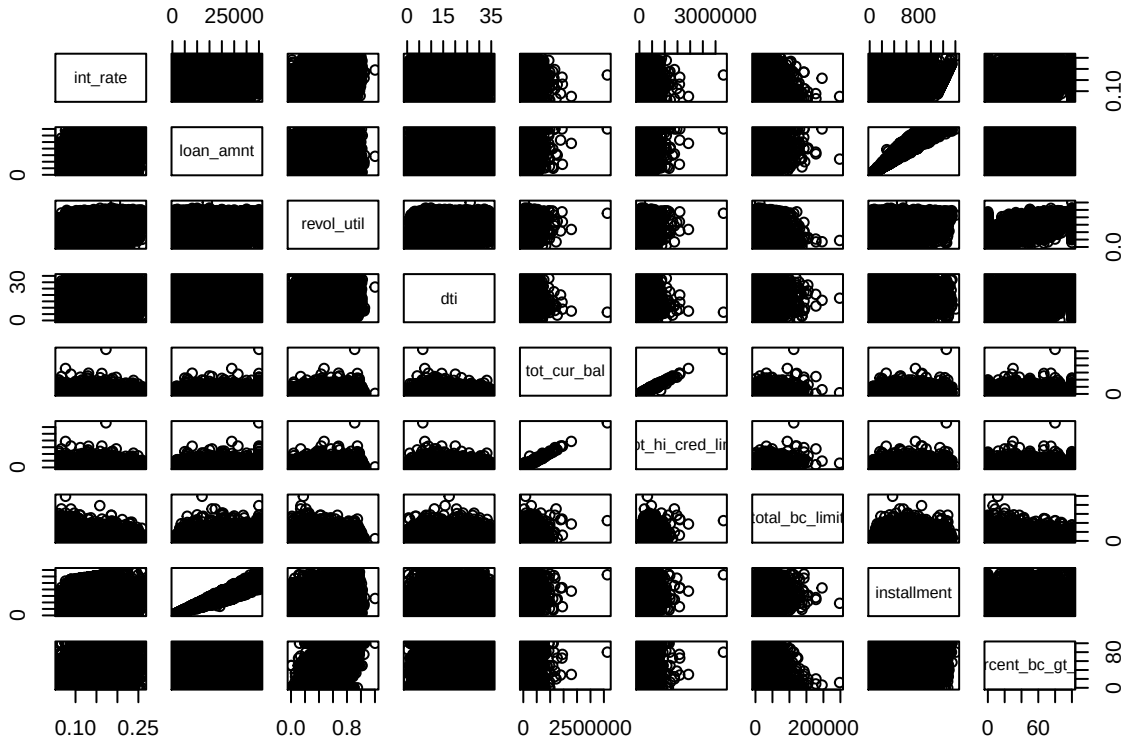
```
##
## Call:
## lm(formula = int_rate ~ loan_amnt + revol_util + dti + tot_cur_bal +
##     tot_hi_cred_lim + home_ownership + total_bc_limit + term +
##     percent_bc_gt_75 + purpose + installment, data = Loan)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.106882 -0.017914 -0.002578  0.016402  0.115353
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.031e-01  3.093e-03  33.326  < 2e-16 ***
## loan_amnt                   -1.517e-05  2.331e-07 -65.060  < 2e-16 ***
## revol_util                   1.893e-02  1.732e-03  10.924  < 2e-16 ***
## dti                          2.533e-04  3.536e-05   7.165 8.35e-13 ***
## tot_cur_bal                  9.118e-08  1.419e-08   6.425 1.38e-10 ***
## tot_hi_cred_lim             -1.053e-07  1.340e-08  -7.859 4.28e-15 ***
## home_ownershipOTHER          5.659e-03  2.574e-02   0.220  0.82598
## home_ownershipOWN            4.853e-03  1.036e-03   4.684 2.85e-06 ***
## home_ownershipRENT           5.786e-03  6.544e-04   8.842  < 2e-16 ***
## total_bc_limit              -2.216e-07  1.840e-08 -12.045  < 2e-16 ***
## term 60 months               1.211e-01  1.311e-03  92.385  < 2e-16 ***
## percent_bc_gt_75             1.503e-04  1.094e-05  13.745  < 2e-16 ***
## purposecredit_card          -5.988e-03  3.015e-03  -1.986  0.04705 *
## purposedebt_consolidation -8.804e-04  2.988e-03  -0.295  0.76826
## purposehome_improvement      2.529e-03  3.174e-03   0.797  0.42553
## purposehouse                 1.345e-02  4.724e-03   2.846  0.00444 **
## purposemajor_purchase        2.181e-03  3.505e-03   0.622  0.53370
## purposemedical               3.042e-02  4.388e-03   6.932 4.39e-12 ***
## purposemoving                4.593e-02  4.693e-03   9.788  < 2e-16 ***
## purposeother                 3.039e-02  3.205e-03   9.481  < 2e-16 ***
## purposerenewable_energy      2.228e-02  1.515e-02   1.470  0.14149
## purposesmall_business        2.578e-02  4.092e-03   6.299 3.12e-10 ***
## purposevacation              3.408e-02  5.072e-03   6.720 1.92e-11 ***
## purposewedding               3.362e-02  4.664e-03   7.208 6.09e-13 ***
## installment                  4.821e-04  6.981e-06  69.061  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02573 on 9975 degrees of freedom
## Multiple R-squared:  0.6603, Adjusted R-squared:  0.6595
## F-statistic:   808 on 24 and 9975 DF,  p-value: < 2.2e-16
```

Testing the assumptions

home_ownership, term and purpose are categorical string variables Hence removed those in checks for collinearity

```
pairs(~int_rate + loan_amnt + revol_util + dti + tot_cur_bal + tot_hi_cred_lim +
        total_bc_limit + installment + percent_bc_gt_75, data = Loan)
```




```
cor(Loan[,c("loan_amnt","revol_util","dti","tot_cur_bal","tot_hi_cred_lim","total_bc_limit","installmen
```

```
##                      loan_amnt   revol_util         dti   tot_cur_bal
## loan_amnt        1.000000000  0.102953568  0.04002413  0.3204235085
## revol_util       0.102953568  1.000000000  0.19826912  0.0608295777
## dti              0.040024130  0.198269123  1.00000000 -0.0365929272
## tot_cur_bal      0.320423509  0.060829578 -0.03659293  1.0000000000
## tot_hi_cred_lim  0.343617668  0.007232872 -0.02255970  0.9886883565
## total_bc_limit   0.373457902 -0.200506273  0.03414960  0.3141158787
## installment      0.953322951  0.131482531  0.03228388  0.2794928552
## percent_bc_gt_75 0.009192116  0.700251750  0.18631664 -0.0001950083
##                  tot_hi_cred_lim total_bc_limit installment
## loan_amnt            0.343617668      0.3734579  0.95332295
## revol_util           0.007232872     -0.2005063  0.13148253
## dti                 -0.022559698      0.0341496  0.03228388
## tot_cur_bal          0.988688357      0.3141159  0.27949286
## tot_hi_cred_lim      1.000000000      0.3888274  0.30006439
## total_bc_limit       0.388827358      1.0000000  0.33489123
```

```
## installment          0.300064394      0.3348912  1.00000000
## percent_bc_gt_75    -0.040306943     -0.2344359  0.03909079
##                  percent_bc_gt_75
## loan_amnt            0.0091921156
## revol_util           0.7002517499
## dti                  0.1863166365
## tot_cur_bal         -0.0001950083
## tot_hi_cred_lim     -0.0403069433
## total_bc_limit      -0.2344358626
## installment          0.0390907927
## percent_bc_gt_75     1.0000000000
```

'tot_cur_bal' is highly collinear with 'tot_hi_cred_lim'

Hence removing 'tot_cur_bal' (Since it has higher p-value among both).

'Installment' is highly collinear with 'loan_amnt' so removing installment as well. Plus installment does not make logical sense since it is derived in part from interest rate.

Model 2

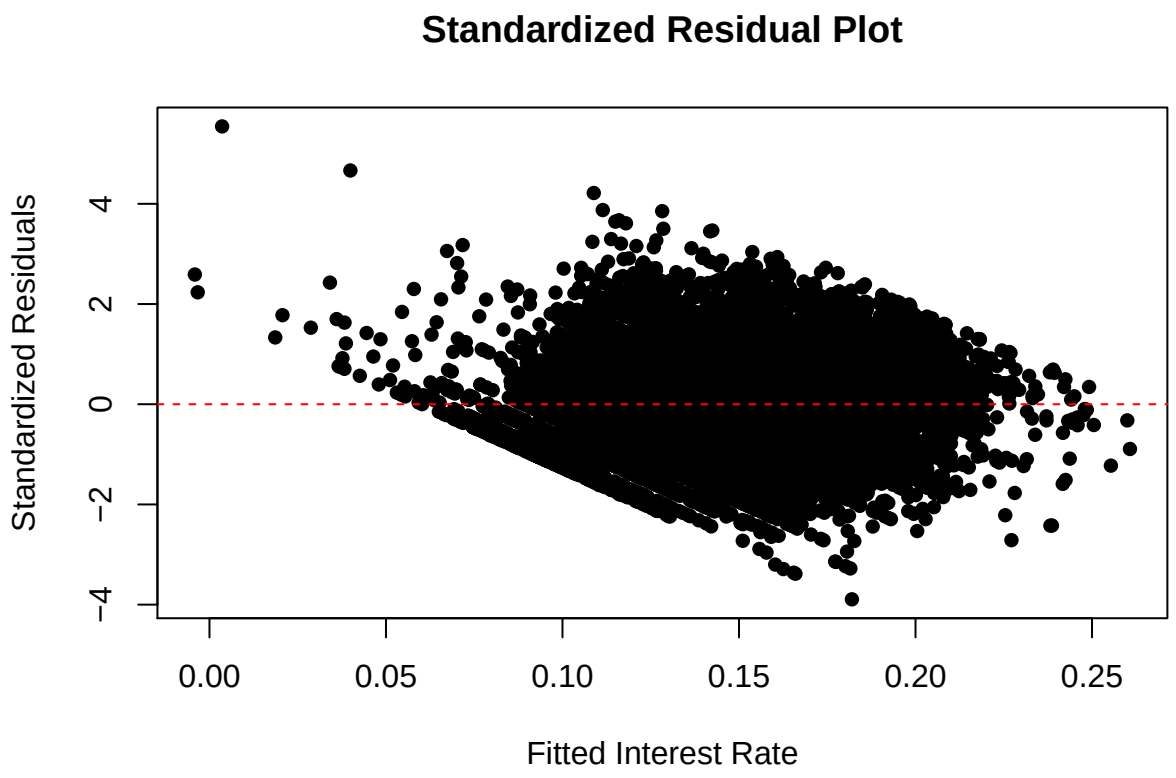
```
Loan$homeF <- factor(Loan$home_ownership)
Loan <- within(Loan,homeF <- relevel(homeF, ref="RENT"))

model <- lm(formula= int_rate ~  loan_amnt + revol_util + dti + tot_hi_cred_lim +
            homeF + total_bc_limit + term + percent_bc_gt_75 + purpose, data = Loan)

summary(model)
```

```
##
## Call:
## lm(formula = int_rate ~ loan_amnt + revol_util + dti + tot_hi_cred_lim +
##     homeF + total_bc_limit + term + percent_bc_gt_75 + purpose,
##     data = Loan)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.121981 -0.022313 -0.001856  0.020884  0.169125
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.097e-01  3.722e-03  29.464  < 2e-16 ***
## loan_amnt                   6.724e-07  4.861e-08  13.833  < 2e-16 ***
## revol_util                  3.155e-02  2.031e-03  15.533  < 2e-16 ***
## dti                         3.236e-04  4.249e-05   7.616 2.86e-14 ***
## tot_hi_cred_lim            -3.515e-08  2.500e-09 -14.063  < 2e-16 ***
## homeFMORTGAGE              -8.573e-03  7.956e-04 -10.776  < 2e-16 ***
## homeFOTHER                 -2.586e-03  3.135e-02  -0.082  0.93427
## homeFOWN                   -1.780e-03  1.265e-03  -1.407  0.15940
## total_bc_limit             -4.653e-07  2.000e-08 -23.261  < 2e-16 ***
## term 60 months              4.331e-02  8.090e-04  53.533  < 2e-16 ***
## percent_bc_gt_75            2.255e-04  1.324e-05  17.030  < 2e-16 ***
## purposecredit_card         -9.374e-03  3.671e-03  -2.553  0.01069 *
## purposedebt_consolidation -9.786e-04  3.639e-03  -0.269  0.78801
```

```
## purposehome_improvement    3.587e-03  3.866e-03   0.928  0.35353
## purposehouse               1.611e-02  5.755e-03   2.799  0.00514 **
## purposemajor_purchase      1.102e-03  4.269e-03   0.258  0.79634
## purposemedical             3.529e-02  5.345e-03   6.603 4.24e-11 ***
## purposemoving              4.811e-02  5.716e-03   8.417  < 2e-16 ***
## purposeother               3.987e-02  3.901e-03  10.219  < 2e-16 ***
## purposerenewable_energy    3.419e-02  1.845e-02   1.853  0.06395 .
## purposesmall_business      4.249e-02  4.976e-03   8.538  < 2e-16 ***
## purposevacation            3.467e-02  6.178e-03   5.612 2.06e-08 ***
## purposewedding             4.116e-02  5.679e-03   7.247 4.56e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03134 on 9977 degrees of freedom
## Multiple R-squared:  0.4959, Adjusted R-squared:  0.4948
## F-statistic: 446.1 on 22 and 9977 DF,  p-value: < 2.2e-16
```

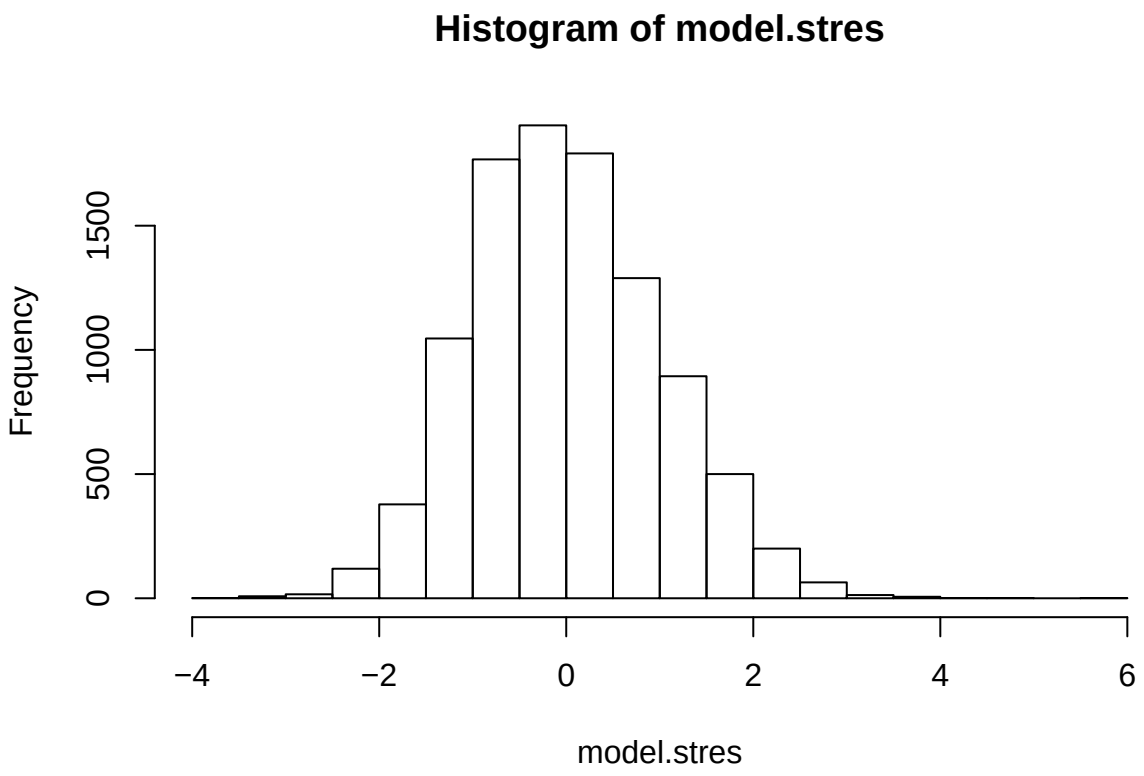
Check for Standardized residuals

```
model.stres <- rstandard(model)
plot(model$fitted.values, model.stres, pch = 16, main = "Standardized Residual Plot",
     xlab = "Fitted Interest Rate", ylab = "Standardized Residuals")
abline(0,0, lty=2, col="red")
```

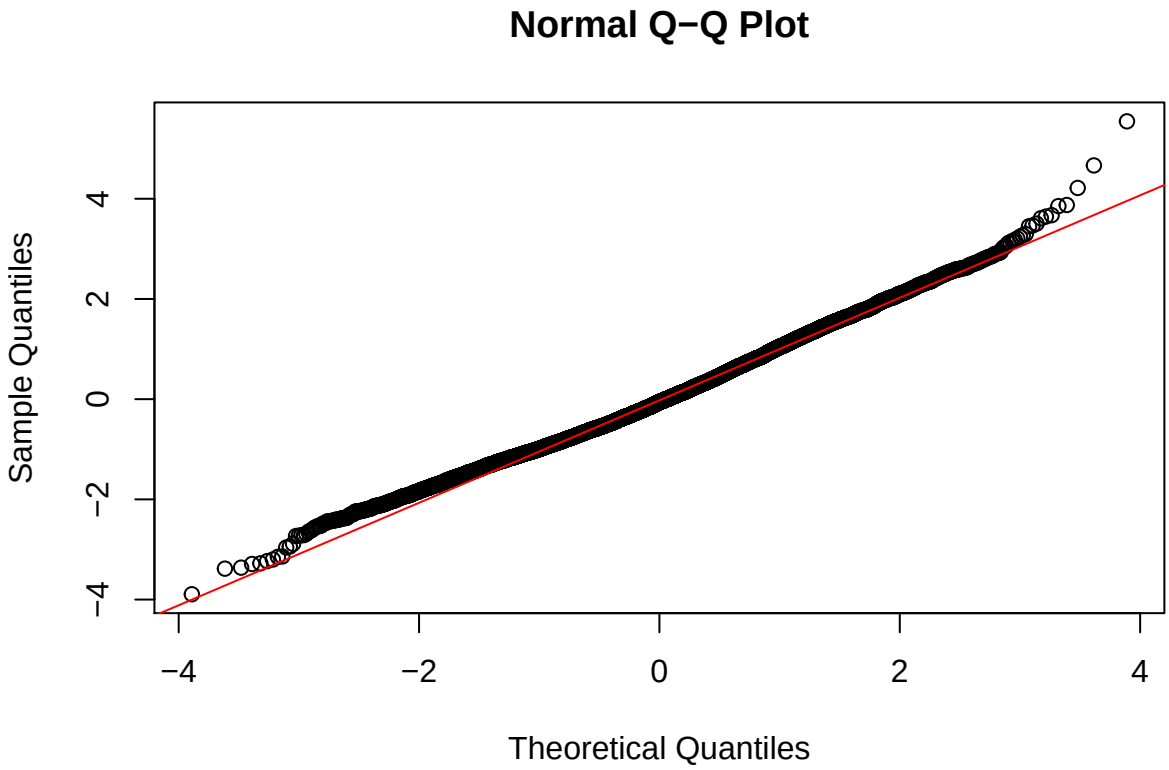


Check for normality

```
hist(model.stres)
```




```
# x <- model.stres
# xfit <- seq(min(x), max(x), length = 50)
# yfit <- dnorm(xfit, mean = mean(x), sd = sd(x))
# yfit <- yfit*diff(h$mids[1:2])*length(x)
# lines(xfit, yfit, col="blue")

qqnorm(model.stres)
qqline(model.stres,col="red")
```




```
#shapiro.test(model.stres)
```

Model 3

After changing home_ownership to factors the other variables become insignificant. Removing home_ownership.

```
Loan$homeF <- factor(Loan$home_ownership)
Loan <- within(Loan,homeF <- relevel(homeF, ref="RENT"))

model <- lm(formula= int_rate ~  loan_amnt + revol_util + dti + tot_hi_cred_lim +
              total_bc_limit + term + percent_bc_gt_75 + purpose, data = Loan)

summary(model)
```

```
##
## Call:
## lm(formula = int_rate ~ loan_amnt + revol_util + dti + tot_hi_cred_lim +
##     total_bc_limit + term + percent_bc_gt_75 + purpose, data = Loan)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.123594 -0.022622 -0.002051  0.020961  0.206869
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.074e-01  3.733e-03  28.765  < 2e-16 ***
## loan_amnt                  6.519e-07  4.885e-08  13.344  < 2e-16 ***
```

```
## revol_util                 3.162e-02  2.043e-03  15.480  < 2e-16 ***
## dti                        3.163e-04  4.272e-05   7.404 1.43e-13 ***
## tot_hi_cred_lim           -4.876e-08  2.183e-09 -22.329  < 2e-16 ***
## total_bc_limit            -4.498e-07  2.007e-08 -22.409  < 2e-16 ***
## term 60 months             4.290e-02  8.128e-04  52.777  < 2e-16 ***
## percent_bc_gt_75           2.269e-04  1.332e-05  17.040  < 2e-16 ***
## purposecredit_card        -9.418e-03  3.693e-03  -2.550  0.01078 *
## purposedebt_consolidation -1.135e-03  3.661e-03  -0.310  0.75644
## purposehome_improvement    1.702e-03  3.884e-03   0.438  0.66126
## purposehouse               1.699e-02  5.788e-03   2.936  0.00334 **
## purposemajor_purchase      1.725e-03  4.294e-03   0.402  0.68785
## purposemedical             3.560e-02  5.376e-03   6.622 3.73e-11 ***
## purposemoving              4.989e-02  5.746e-03   8.683  < 2e-16 ***
## purposeother               4.023e-02  3.924e-03  10.252  < 2e-16 ***
## purposerenewable_energy    3.597e-02  1.856e-02   1.938  0.05263 .
## purposesmall_business      4.302e-02  5.005e-03   8.594  < 2e-16 ***
## purposevacation            3.546e-02  6.214e-03   5.707 1.18e-08 ***
## purposewedding             4.188e-02  5.712e-03   7.332 2.44e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03152 on 9980 degrees of freedom
## Multiple R-squared:  0.4898, Adjusted R-squared:  0.4888
## F-statistic: 504.3 on 19 and 9980 DF,  p-value: < 2.2e-16
```

Model 4

Removing purpose as many of the variables are insignificant. loan_amnt becomes insignificant in the above model. Since the scale is too high. Taking log of loan_amnt and using it in the model.

```
model <- lm(formula= int_rate ~  log(loan_amnt) + revol_util + dti +
            tot_hi_cred_lim +  total_bc_limit + term + percent_bc_gt_75,
            data = Loan)

summary(model)
```

```
##
## Call:
## lm(formula = int_rate ~ log(loan_amnt) + revol_util + dti + tot_hi_cred_lim +
##     total_bc_limit + term + percent_bc_gt_75, data = Loan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12662 -0.02477 -0.00352  0.02312  0.19189
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.418e-01  5.663e-03  25.038  < 2e-16 ***
## log(loan_amnt)   -2.920e-03  6.276e-04  -4.653 3.32e-06 ***
## revol_util        3.311e-02  2.171e-03  15.249  < 2e-16 ***
## dti               2.490e-04  4.544e-05   5.480 4.35e-08 ***
## tot_hi_cred_lim  -4.035e-08  2.298e-09 -17.560  < 2e-16 ***
## total_bc_limit   -4.061e-07  2.115e-08 -19.200  < 2e-16 ***
```

```
## term 60 months    4.894e-02  8.552e-04  57.228  < 2e-16 ***
## percent_bc_gt_75  2.159e-04  1.423e-05  15.169  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03372 on 9992 degrees of freedom
## Multiple R-squared:  0.4154, Adjusted R-squared:  0.4149
## F-statistic:  1014 on 7 and 9992 DF,  p-value: < 2.2e-16
```

Model 5

```
attach(Loan)

model <- lm(formula= int_rate ~  log(loan_amnt) + revol_util + dti +
            tot_hi_cred_lim +  total_bc_limit + term + percent_bc_gt_75,
            data = Loan)

summary(model)
```

```
##
## Call:
## lm(formula = int_rate ~ log(loan_amnt) + revol_util + dti + tot_hi_cred_lim +
##     total_bc_limit + term + percent_bc_gt_75, data = Loan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12662 -0.02477 -0.00352  0.02312  0.19189
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.418e-01  5.663e-03  25.038  < 2e-16 ***
## log(loan_amnt)   -2.920e-03  6.276e-04  -4.653 3.32e-06 ***
## revol_util        3.311e-02  2.171e-03  15.249  < 2e-16 ***
## dti               2.490e-04  4.544e-05   5.480 4.35e-08 ***
## tot_hi_cred_lim  -4.035e-08  2.298e-09 -17.560  < 2e-16 ***
## total_bc_limit   -4.061e-07  2.115e-08 -19.200  < 2e-16 ***
## term 60 months    4.894e-02  8.552e-04  57.228  < 2e-16 ***
## percent_bc_gt_75  2.159e-04  1.423e-05  15.169  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03372 on 9992 degrees of freedom
## Multiple R-squared:  0.4154, Adjusted R-squared:  0.4149
## F-statistic:  1014 on 7 and 9992 DF,  p-value: < 2.2e-16
```

Model 6

```
attach(Loan)
```

```
## The following objects are masked from Loan (pos = 3):
```

```
##
##     acc_now_delinq, addr_state, annual_inc, avg_cur_bal,
##     bc_open_to_buy, bc_util, chargeoff_within_12_mths,
##     collection_recovery_fee, collections_12_mths_ex_med,
##     delinq_2yrs, delinq_amnt, dti, earliest_cr_line, emp_length,
##     emp_title, emp_title_cluster, funded_amnt, funded_amnt_inv,
##     grade, home_ownership, homeF, id, ID_unit,
##     initial_list_status, inq_last_6mths, installment, int_rate,
##     issue_d, last_credit_pull_d, last_pymnt_amnt, last_pymnt_d,
##     loan_amnt, loan_status, mo_sin_old_rev_tl_op,
##     mo_sin_rcnt_rev_tl_op, mo_sin_rcnt_tl, mort_acc,
##     mths_since_recent_bc, mths_since_recent_inq,
##     num_accts_ever_120_pd, num_actv_bc_tl, num_actv_rev_tl,
##     num_bc_sats, num_bc_tl, num_il_tl, num_op_rev_tl,
##     num_rev_accts, num_rev_tl_bal_gt_0, num_sats,
##     num_tl_120dpd_2m, num_tl_30dpd, num_tl_90g_dpd_24m,
##     num_tl_op_past_12m, pct_tl_nvr_dlq, percent_bc_gt_75, Prob,
##     pub_rec, pub_rec_bankruptcies, purpose, recoveries, revol_bal,
##     revol_util, Stratum, sub_grade, tax_liens, term, title,
##     tot_coll_amt, tot_cur_bal, tot_hi_cred_lim, total_acc,
##     total_bal_ex_mort, total_bc_limit, total_il_high_credit_limit,
##     total_rev_hi_lim, verification_status, X, X.1, zip_code,
##     zip_new
```

```
Loan$purposeF = factor(Loan$purpose)
Loan = within(Loan, purposeF<-relevel(purposeF,ref="credit_card"))

model6 <- lm(formula= log(int_rate) ~  log(loan_amnt) + revol_util + dti +
            log(tot_hi_cred_lim) + log(total_bc_limit) + term +purposeF, data = Loan)

summary(model6)
```

```
##
## Call:
## lm(formula = log(int_rate) ~ log(loan_amnt) + revol_util + dti +
##     log(tot_hi_cred_lim) + log(total_bc_limit) + term + purposeF,
##     data = Loan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12897 -0.14720  0.00856  0.15295  0.92625
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -1.1566101  0.0422354 -27.385  < 2e-16 ***
## log(loan_amnt)              0.0315160  0.0045932   6.862 7.22e-12 ***
## revol_util                  0.4345859  0.0109184  39.803  < 2e-16 ***
## dti                         0.0043846  0.0003116  14.072  < 2e-16 ***
## log(tot_hi_cred_lim)       -0.0696411  0.0024753 -28.135  < 2e-16 ***
## log(total_bc_limit)        -0.0812760  0.0028993 -28.033  < 2e-16 ***
## term 60 months              0.3205601  0.0058434  54.858  < 2e-16 ***
## purposeFcar                 0.0320237  0.0268659   1.192  0.23330
## purposeFdebt_consolidation  0.0581684  0.0056048  10.378  < 2e-16 ***
## purposeFhome_improvement    0.0693304  0.0114073   6.078 1.26e-09 ***
## purposeFhouse               0.1331312  0.0331613   4.015 6.00e-05 ***
```

```
## purposeFmajor_purchase      0.0477691  0.0176704   2.703  0.00688 **
## purposeFmedical             0.2947464  0.0293694  10.036  < 2e-16 ***
## purposeFmoving              0.3609366  0.0329391  10.958  < 2e-16 ***
## purposeFother               0.3008862  0.0121695  24.725  < 2e-16 ***
## purposeFrenewable_energy    0.2757938  0.1324057   2.083  0.03728 *
## purposeFsmall_business      0.3284137  0.0254994  12.879  < 2e-16 ***
## purposeFvacation            0.2810771  0.0371926   7.557 4.48e-14 ***
## purposeFwedding             0.3132694  0.0324871   9.643  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2291 on 9981 degrees of freedom
## Multiple R-squared:  0.4915, Adjusted R-squared:  0.4906
## F-statistic:   536 on 18 and 9981 DF,  p-value: < 2.2e-16
```