STAT 628 Credit Risk Project: Installment 2 (technical)

Bohyoon Lee and Salsabila Mahdi

2025-09-18

```r
installment2_id01 <- read.csv("installment2_id01.csv")
str(installment2_id01)
```

```
'data.frame':   2271 obs. of  11 variables:
 $ PRSM           : num  0.948 0.963 1.14 0.407 0.925 ...
 $ FICO           : int  839 736 729 599 710 680 691 770 850 691 ...
 $ TotalAmtOwed   : int  197639 224181 32346 63525 175181 248385 258654 280182 248993 298062 .
 $ Volume         : int  82782 58002 11564 30373 54820 43211 241356 94187 181704 87855 ...
 $ Stress         : num  0.199 0.322 0.233 0.174 0.266 ...
 $ Num_Delinquent : int  3 4 4 5 4 4 4 4 3 4 ...
 $ Num_CreditLines: int  12 11 12 11 11 11 12 13 9 8 ...
 $ WomanOwned     : int  1 1 1 0 1 1 0 1 1 1 ...
 $ CorpStructure  : chr  "Sole" "Corp" "Partner" "Sole" ...
 $ NAICS          : int  444140 458210 722410 444240 445230 722514 445292 445250 722513 722330
 $ Months         : int  15 17 22 12 18 23 14 16 19 19 ...
```

```r
summary(installment2_id01)
```

```
      PRSM              FICO        TotalAmtOwed         Volume       
 Min.   :-0.7804   Min.   :476.0   Min.   :  10153   Min.   :   2057  
 1st Qu.: 0.6111   1st Qu.:659.0   1st Qu.:  96146   1st Qu.:  41172  
 Median : 0.7930   Median :698.0   Median : 194844   Median :  86807  
 Mean   : 0.8066   Mean   :700.7   Mean   : 231844   Mean   : 153598  
 3rd Qu.: 0.9834   3rd Qu.:740.0   3rd Qu.: 299862   3rd Qu.: 170504  
 Max.   : 2.6988   Max.   :850.0   Max.   :1992741   Max.   :4434782  
     Stress         Num_Delinquent  Num_CreditLines   WomanOwned    
 Min.   :0.00627   Min.   :3.000   Min.   : 8.00   Min.   :0.0000  
 1st Qu.:0.11243   1st Qu.:4.000   1st Qu.: 9.00   1st Qu.:0.0000  
 Median :0.18454   Median :4.000   Median :10.00   Median :0.0000  
 Mean   :0.19800   Mean   :4.083   Mean   :10.26   Mean   :0.4469  
 3rd Qu.:0.26115   3rd Qu.:4.000   3rd Qu.:12.00   3rd Qu.:1.0000  
 Max.   :0.65406   Max.   :8.000   Max.   :13.00   Max.   :1.0000  
 CorpStructure          NAICS            Months      
 Length:2271        Min.   :441120   Min.   :  5.00  
 Class :character   1st Qu.:445230   1st Qu.: 15.00  
 Mode  :character   Median :458210   Median : 18.00  
                    Mean   :509739   Mean   : 18.47  
                    3rd Qu.:459910   3rd Qu.: 21.00  
                    Max.   :722514   Max.   :116.00  
```

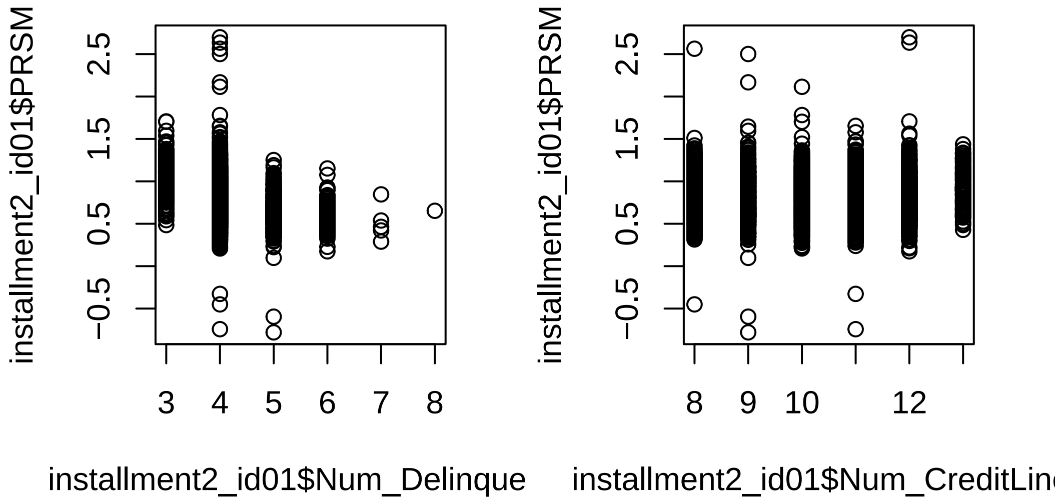
```r
installment2_id01 <- installment2_id01 %>%
  mutate(NAICS = as.factor(NAICS),
         WomanOwned = as.factor(WomanOwned),
         FICO = case_when(
           (300 <= FICO)&(FICO <= 579) ~ "Poor",
           (580 <= FICO)&(FICO <= 669) ~ "Fair",
           (670 <= FICO)&(FICO <= 739) ~ "Good",
           (740 <= FICO)&(FICO <= 799) ~ "Very Good",
           (800 <= FICO)&(FICO <= 850) ~ "Excellent",
         ))
table(installment2_id01$FICO)
```

```
Excellent      Fair      Good      Poor Very Good 
      212       592      1004       102       361 
```

```r
table(installment2_id01$NAICS)
```

```
441120 444140 444240 445110 445131 445230 445240 445250 445291 445292 445320 
    95    110     99    110    109    116     99    104    103     77    113 
458210 458310 459210 459310 459910 722330 722410 722511 722513 722514 
   111    116    202    104    104    111     92     97    103     96 
```

```r
par(mfrow=c(1,2))
plot(installment2_id01$Num_Delinquent, installment2_id01$PRSM)
plot(installment2_id01$Num_CreditLines, installment2_id01$PRSM)
```



```
# TODO: remove or transform to Num_Delinquent/Num_CreditLines
```

```
# TODO: outliers
```

```
str(installment2_id01)
```

```
'data.frame':   2271 obs. of  11 variables:
 $ PRSM           : num  0.948 0.963 1.14 0.407 0.925 ...
 $ FICO           : chr  "Excellent" "Good" "Good" "Fair" ...
 $ TotalAmtOwed   : int  197639 224181 32346 63525 175181 248385 258654 280182 248993 298062 .
 $ Volume         : int  82782 58002 11564 30373 54820 43211 241356 94187 181704 87855 ...
 $ Stress         : num  0.199 0.322 0.233 0.174 0.266 ...
 $ Num_Delinquent : int  3 4 4 5 4 4 4 4 3 4 ...
 $ Num_CreditLines: int  12 11 12 11 11 11 12 13 9 8 ...
 $ WomanOwned     : Factor w/ 2 levels "0","1": 2 2 2 1 2 2 1 2 2 2 ...
 $ CorpStructure  : chr  "Sole" "Corp" "Partner" "Sole" ...
 $ NAICS          : Factor w/ 21 levels "441120","444140",..: 2 12 18 3 6 21 10 8 20 17 ...
 $ Months         : int  15 17 22 12 18 23 14 16 19 19 ...
```

```
fullmodel <- lm(PRSM ~ ., data = installment2_id01)
summary(fullmodel)
```

```
Call:
lm(formula = PRSM ~ ., data = installment2_id01)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.34986 -0.07112 -0.00122  0.06816  1.98673

Coefficients:
                            Estimate      Std. Error t value
(Intercept)           0.566380773911  0.047154353587  12.011
FICOFair             -0.209916298873  0.017797038499 -11.795
FICOGood             -0.129033790801  0.014680423702  -8.790
FICOPoor             -0.203484272193  0.033894797455  -6.003
FICOVery Good        -0.103046976592  0.015661664599  -6.580
TotalAmtOwed          0.000000505994  0.000000020242  24.997
Volume               -0.000000008736  0.000000017514  -0.499
Stress                0.478055129827  0.032101865877  14.892
Num_Delinquent       -0.007488610681  0.010711632079  -0.699
Num_CreditLines      -0.001291709101  0.002166358789  -0.596
WomanOwned1           0.270815054539  0.006818746488  39.716
CorpStructureLLC      0.229819212627  0.008611429527  26.688
CorpStructurePartner  0.148254208672  0.008794536161  16.858
CorpStructureSole    -0.025504025449  0.008763167199  -2.910
NAICS444140          -0.053781821473  0.020409448897  -2.635
NAICS444240          -0.008562180044  0.020952292792  -0.409
NAICS445110          -0.029525621480  0.020419305368  -1.446
NAICS445131          -0.039542092390  0.020455677998  -1.933
NAICS445230          -0.045429285100  0.020200987590  -2.249
NAICS445240          -0.032622018736  0.020912665159  -1.560
NAICS445250          -0.036610762166  0.020674020557  -1.771
NAICS445291          -0.037610179846  0.020732373483  -1.814
NAICS445292          -0.030545962461  0.022401722048  -1.364
NAICS445320          -0.007210200863  0.020270703785  -0.356
NAICS458210          -0.047146094269  0.020358944807  -2.316
NAICS458310          -0.018646654283  0.020175176155  -0.924
NAICS459210          -0.008796687929  0.018139587024  -0.485
NAICS459310          -0.032172157958  0.020700184000  -1.554
NAICS459910          -0.035145851627  0.020677863051  -1.700
NAICS722330          -0.047655936883  0.020353972262  -2.341
NAICS722410          -0.023552687894  0.021344333491  -1.103
NAICS722511          -0.033309871545  0.021030786313  -1.584
NAICS722513          -0.031730578900  0.020718656502  -1.531
NAICS722514          -0.038757251012  0.021074697404  -1.839
Months                0.001538887409  0.000396450037   3.882
                                 Pr(>|t|)
(Intercept)          < 0.0000000000000002 ***
FICOFair             < 0.0000000000000002 ***
FICOGood             < 0.0000000000000002 ***
FICOPoor                  0.0000000022485 ***
FICOVery Good             0.0000000000586 ***
```

```
TotalAmtOwed         < 0.0000000000000002 ***
Volume                           0.617964    
Stress               < 0.0000000000000002 ***
Num_Delinquent                   0.484556    
Num_CreditLines                  0.551063    
WomanOwned1          < 0.0000000000000002 ***
CorpStructureLLC     < 0.0000000000000002 ***
CorpStructurePartner < 0.0000000000000002 ***
CorpStructureSole                0.003646 ** 
NAICS444140                      0.008468 ** 
NAICS444240                      0.682835    
NAICS445110                      0.148327    
NAICS445131                      0.053355 .  
NAICS445230                      0.024618 *  
NAICS445240                      0.118921    
NAICS445250                      0.076720 .  
NAICS445291                      0.069799 .  
NAICS445292                      0.172845    
NAICS445320                      0.722102    
NAICS458210                      0.020662 *  
NAICS458310                      0.355462    
NAICS459210                      0.627763    
NAICS459310                      0.120279    
NAICS459910                      0.089329 .  
NAICS722330                      0.019301 *  
NAICS722410                      0.269945    
NAICS722511                      0.113366    
NAICS722513                      0.125788    
NAICS722514                      0.066042 .  
Months                           0.000107 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1454 on 2236 degrees of freedom
Multiple R-squared:  0.7224,    Adjusted R-squared:  0.7182 
F-statistic: 171.2 on 34 and 2236 DF,  p-value: < 0.00000000000000022
```

```
stepmodel <- step(fullmodel)
```

```
Start:  AIC=-8724.62
PRSM ~ FICO + TotalAmtOwed + Volume + Stress + Num_Delinquent + 
    Num_CreditLines + WomanOwned + CorpStructure + NAICS + Months

                  Df Sum of Sq    RSS     AIC
- NAICS           20     0.488 47.735 -8741.3
- Volume           1     0.005 47.253 -8726.4
- Num_CreditLines  1     0.008 47.255 -8726.3
- Num_Delinquent   1     0.010 47.258 -8726.1
<none>                         47.247 -8724.6
- Months           1     0.318 47.566 -8711.4
- FICO             4     3.938 51.185 -8550.8
- Stress           1     4.686 51.933 -8511.9
- TotalAmtOwed     1    13.203 60.451 -8167.0
- CorpStructure    3    25.290 72.538 -7757.0
- WomanOwned       1    33.331 80.578 -7514.3

Step:  AIC=-8741.29
PRSM ~ FICO + TotalAmtOwed + Volume + Stress + Num_Delinquent +
    Num_CreditLines + WomanOwned + CorpStructure + Months

                  Df Sum of Sq    RSS     AIC
- Volume           1     0.005 47.740 -8743.1
- Num_CreditLines  1     0.006 47.742 -8743.0
- Num_Delinquent   1     0.012 47.747 -8742.7
<none>                         47.735 -8741.3
- Months           1     0.334 48.069 -8727.5
- FICO             4     3.938 51.673 -8569.3
- Stress           1     4.711 52.447 -8529.5
- TotalAmtOwed     1    13.412 61.148 -8180.9
- CorpStructure    3    25.502 73.238 -7775.2
- WomanOwned       1    33.661 81.397 -7531.3

Step:  AIC=-8743.05
PRSM ~ FICO + TotalAmtOwed + Stress + Num_Delinquent + Num_CreditLines +
    WomanOwned + CorpStructure + Months

                  Df Sum of Sq    RSS     AIC
- Num_CreditLines  1     0.006 47.747 -8744.8
- Num_Delinquent   1     0.012 47.752 -8744.5
<none>                         47.740 -8743.1
- Months           1     0.335 48.075 -8729.2
- FICO             4     3.938 51.678 -8571.0
- Stress           1     6.450 54.191 -8457.2
- TotalAmtOwed     1    22.111 69.851 -7880.7
- CorpStructure    3    25.516 73.257 -7776.6
- WomanOwned       1    33.719 81.460 -7531.6

Step:  AIC=-8744.75
PRSM ~ FICO + TotalAmtOwed + Stress + Num_Delinquent + WomanOwned +
    CorpStructure + Months

                 Df Sum of Sq    RSS     AIC
- Num_Delinquent  1     0.012 47.758 -8746.2
<none>                        47.747 -8744.8
```

```
- Months          1     0.333 48.079 -8731.0
- FICO            4     3.994 51.740 -8570.3
- Stress          1     6.444 54.191 -8459.2
- TotalAmtOwed    1    22.110 69.857 -7882.5
- CorpStructure   3    25.514 73.261 -7778.5
- WomanOwned      1    33.775 81.522 -7531.9

Step:  AIC=-8746.2
PRSM ~ FICO + TotalAmtOwed + Stress + WomanOwned + CorpStructure + 
    Months

                Df Sum of Sq    RSS     AIC
<none>                       47.758 -8746.2
- Months         1     0.330 48.089 -8732.5
- Stress         1     6.471 54.229 -8459.6
- FICO           4     7.245 55.003 -8433.5
- TotalAmtOwed   1    22.102 69.860 -7884.5
- CorpStructure  3    25.504 73.262 -7780.5
- WomanOwned     1    33.780 81.539 -7533.4
```

```
summary(stepmodel)
```

```
Call:
lm(formula = PRSM ~ FICO + TotalAmtOwed + Stress + WomanOwned + 
    CorpStructure + Months, data = installment2_id01)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.37041 -0.07239 -0.00084  0.06862  2.01551 

Coefficients:
                           Estimate     Std. Error t value             Pr(>|t|)
(Intercept)           0.49679821732  0.01601877178  31.014 < 0.0000000000000002
FICOFair             -0.21662092896  0.01239823731 -17.472 < 0.0000000000000002
FICOGood             -0.13370415601  0.01118923399 -11.949 < 0.0000000000000002
FICOPoor             -0.22298802621  0.01828880873 -12.193 < 0.0000000000000002
FICOVery Good        -0.10913794789  0.01266198664  -8.619 < 0.0000000000000002
TotalAmtOwed          0.00000050181  0.00000001552  32.333 < 0.0000000000000002
Stress                0.48572702665  0.02776333701  17.495 < 0.0000000000000002
WomanOwned1           0.27089171498  0.00677687001  39.973 < 0.0000000000000002
CorpStructureLLC      0.22875971245  0.00855856217  26.729 < 0.0000000000000002
CorpStructurePartner  0.14762366668  0.00875050512  16.870 < 0.0000000000000002
CorpStructureSole    -0.02665358379  0.00872273558  -3.056              0.00227
Months                0.00156412299  0.00039560061   3.954            0.0000793

(Intercept)          ***
```

```
FICOFair             ***
FICOGood             ***
FICOPoor             ***
FICOVery Good        ***
TotalAmtOwed         ***
Stress               ***
WomanOwned1          ***
CorpStructureLLC     ***
CorpStructurePartner ***
CorpStructureSole    **
Months               ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1454 on 2259 degrees of freedom
Multiple R-squared:  0.7194,    Adjusted R-squared:  0.7181
F-statistic: 526.6 on 11 and 2259 DF,  p-value: < 0.00000000000000022
```

```
stress_model <- lm(PRSM ~ Stress, data = installment2_id01)
summary(stress_model)
```

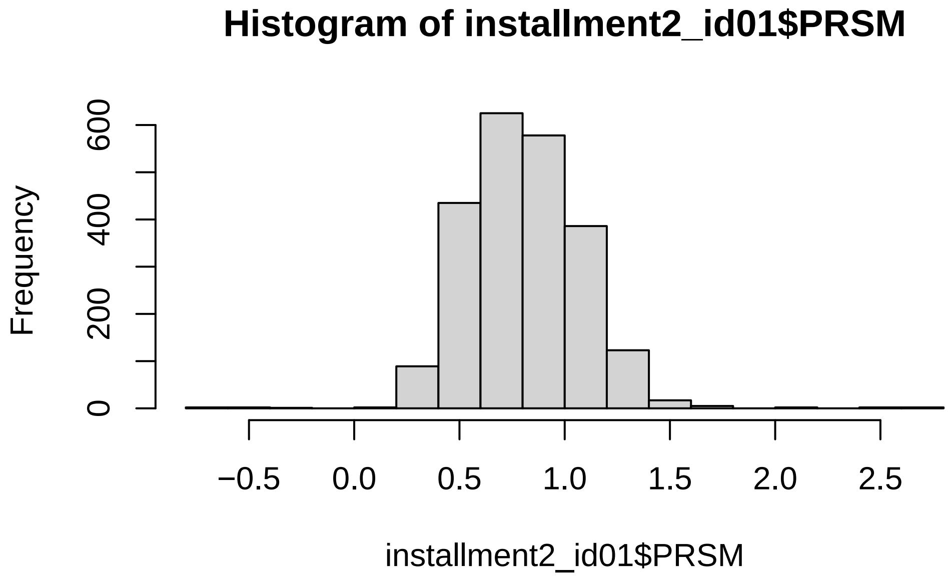
```
Call:
lm(formula = PRSM ~ Stress, data = installment2_id01)

Residuals:
     Min       1Q   Median       3Q      Max
-1.54561 -0.19343 -0.01609  0.18025  1.89805

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)  0.70887    0.01160  61.130 <0.0000000000000002 ***
Stress       0.49342    0.05119   9.639 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2685 on 2269 degrees of freedom
Multiple R-squared:  0.03933,   Adjusted R-squared:  0.03891
F-statistic:  92.9 on 1 and 2269 DF,  p-value: < 0.00000000000000022
```

```
#plot(installment2_id01[,c("PRSM", "TotalAmtOwed", "Volume",
#                           "Stress", "Num_Delinquent", "Num_CreditLines",
#                           "Months")])
#plot(log(installment2_id01$Months), installment2_id01$PRSM)
#plot(installment2_id01$NAICS, as.numeric(installment2_id01$PRSM))
#plot(fullmodel)
```

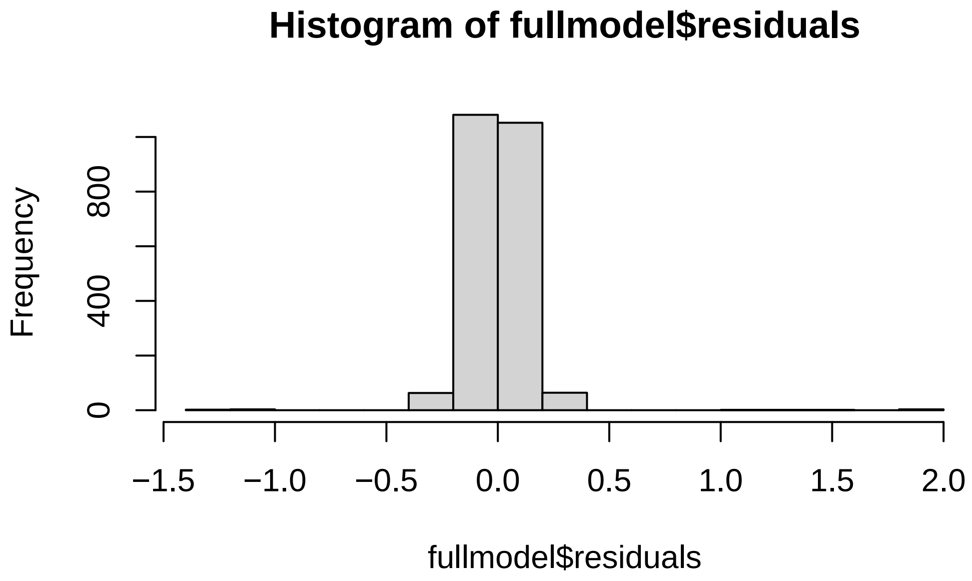
```r
# y: assumptions on residuals. cant compare AIC etc
# x: linearity, outlier
constant <- abs(min(installment2_id01$PRSM)) + 0.01
fullmodel_shifted <- lm(PRSM + constant ~ ., data = installment2_id01)
library(MASS)
```

```
Attaching package: 'MASS'


The following object is masked from 'package:dplyr':

    select
```

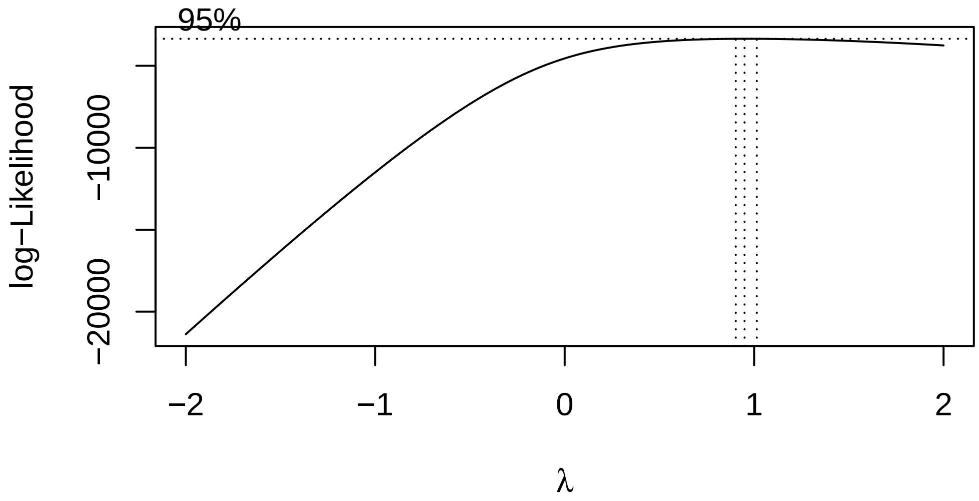
```r
# maybe try log
hist(installment2_id01$PRSM)
```



```r
hist(fullmodel$residuals)
```



```r
boxcox(fullmodel_shifted, lambda = seq(-2, 2, 0.1))
```



```r
library(car)
```

```
Loading required package: carData
```

```
Attaching package: 'car'


The following object is masked from 'package:dplyr':

    recode
```

```
#crPlots(fullmodel)
#boxTidwell(PRSM + constant ~ TotalAmtOwed + Volume + Stress,
#           data = installment2_id01)
#plot(installment2_id01$PRSM, installment2_id01$FICO)
transformedmodel <- lm(PRSM ~ TotalAmtOwed + Volume + FICO + Stress,
           data = installment2_id01)
summary(transformedmodel)
```

```
Call:
lm(formula = PRSM ~ TotalAmtOwed + Volume + FICO + Stress, data = installment2_id01)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.32733 -0.14609 -0.00511  0.14450  1.90794 

Coefficients:
                   Estimate     Std. Error t value            Pr(>|t|)    
(Intercept)    0.85871332846  0.01874418150  45.812 <0.0000000000000002 ***
TotalAmtOwed   0.00000050787  0.00000003048  16.660 <0.0000000000000002 ***
Volume        -0.00000003841  0.00000002637  -1.456               0.145    
FICOFair      -0.39402140298  0.01763862076 -22.339 <0.0000000000000002 ***
FICOGood      -0.22571951914  0.01663559212 -13.568 <0.0000000000000002 ***
FICOPoor      -0.44758980095  0.02651211116 -16.882 <0.0000000000000002 ***
FICOVery Good -0.17392555341  0.01903162337  -9.139 <0.0000000000000002 ***
Stress         0.43567477038  0.04833368443   9.014 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2198 on 2263 degrees of freedom
Multiple R-squared:  0.3575,	Adjusted R-squared:  0.3555 
F-statistic: 179.9 on 7 and 2263 DF,  p-value: < 0.00000000000000022
```