Regression of Credit

171840094 Feichi Lu

April 1, 2020

1 Is there a relationship between balance and all other 10 potential predictors?

The question can be addressed by fitting a multiple regression model of Balance over Income, Limit, Rating, Cards, Age, Education, Gender, Student, Married and Ethnicity. We have to test the null hypothesis

$$H_0 : \beta_{Limit} = \beta_{Rating} = \cdots = \beta_{EthnicityAsian} = \beta_{EthnicityCaucasian} = 0$$

The F-statistic is 487.5, which is far larger than 1. What's more, the p-value is less than $2.2e^{-16}$, which is very close to 0. Thus, the null hypothesis is rejected. It means that there exists a relationship between balance and all other 10 potential predictors.

2 How strong is the relationship?

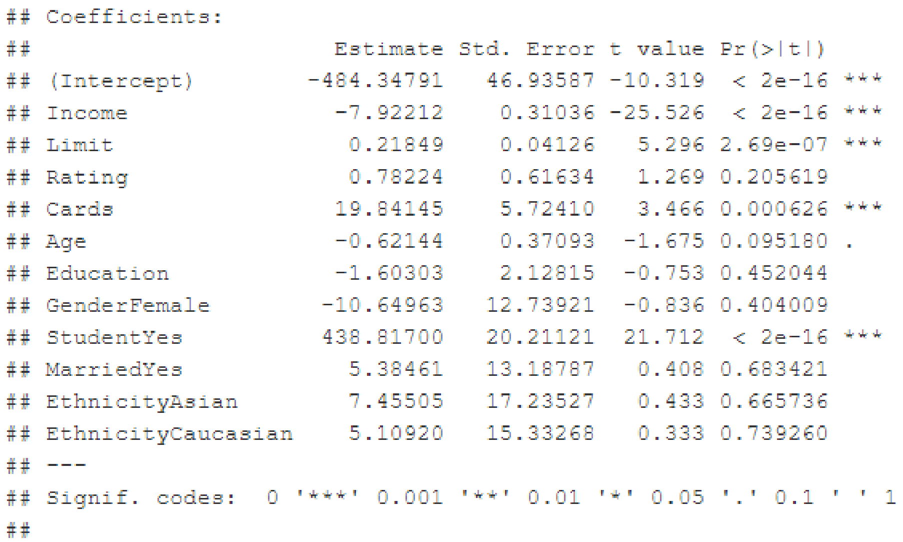
1. RSE

   The RSE estimates the standard deviation of the residuals, so it estimates the deviation of the response from the population regression line.

   $$\hat{\sigma} = RSE = \sqrt{\frac{RSS}{n-p-1}}$$

   Here, the RSE of Balance over all 10 potential predictors is 99.41, degrees of freedom is $238 = 400 - 11 - 1$. The mean of the response is 514.864, indicating a percentage error of 19.3%.

2. R^2



```
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       -484.34791   46.93587 -10.319  < 2e-16 ***
## Income              -7.92212    0.31036 -25.526  < 2e-16 ***
## Limit                0.21849    0.04126   5.296 2.69e-07 ***
## Rating               0.78224    0.61634   1.269 0.205619    
## Cards               19.84145    5.72410   3.466 0.000626 ***
## Age                 -0.62144    0.37093  -1.675 0.095180 .  
## Education           -1.60303    2.12815  -0.753 0.452044    
## GenderFemale       -10.64963   12.73921  -0.836 0.404009    
## StudentYes         438.81700   20.21121  21.712  < 2e-16 ***
## MarriedYes           5.38461   13.18787   0.408 0.683421    
## EthnicityAsian       7.45505   17.23527   0.433 0.665736    
## EthnicityCaucasian   5.10920   15.33268   0.333 0.739260    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Figure 1: The p-value of each predictor

The R^2 statistic records the percentage of variability in the response that is explained by the combination of predictors.

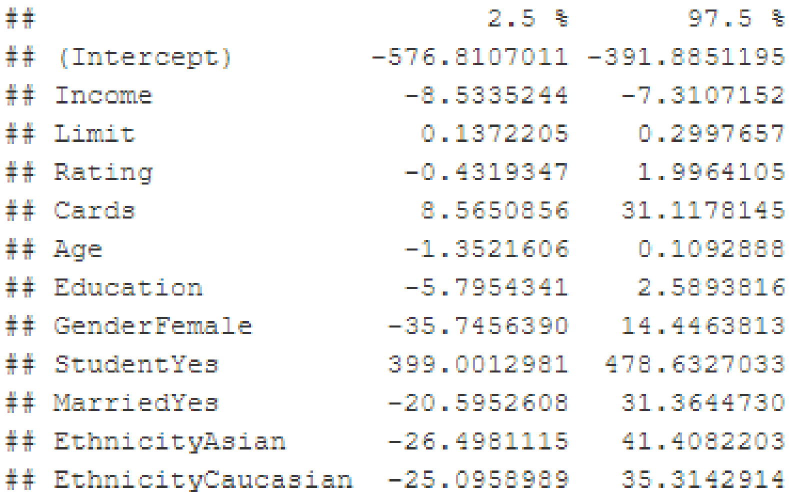
$$R^2 = COV(Y, \hat{Y}) = 1 - \frac{RSS}{TSS}$$

Here, $R^2 = 0.9575$, so the predictors explain 95.75% of the variance in the response.

3 Which predictors contribute to the response Balance?

1. First, we can examine the p-value associated with the t-statistic of each predictor (Figure1). Notice that the p-values of Income, Limit, Cards and Student are very small, smaller than 0.001. The p-value of Age is smaller than 0.1, while the p-values of Rating, Education, Gender, Married and Ethnicity are relatively large. Thus, we can conclude that on one hand, Income, Limit, Cards and Student contribute significantly to Balance. Age also contribute to Balance, but not as much as the four predictors listed before. On the other hand, Education, Gender, Married and Ethnicity doesn't contribute significantly to Balance.

2. We can also use forward selection, backward selection and mixed selection to determine exactly which predictors contribute to the response. Back-



```
##                           2.5 %       97.5 %
## (Intercept)        -576.8107011 -391.8851195
## Income               -8.5335244   -7.3107152
## Limit                 0.1372205    0.2997657
## Rating               -0.4319347    1.9964105
## Cards                 8.5650856   31.1178145
## Age                  -1.3521606    0.1092888
## Education            -5.7954341    2.5893816
## GenderFemale        -35.7456390   14.4463813
## StudentYes          399.0012981  478.6327033
## MarriedYes          -20.5952608   31.3644730
## EthnicityAsian      -26.4981115   41.4082203
## EthnicityCaucasian  -25.0958989   35.3142914
```

Figure 2: The confidence interval of the coefficients

ward and mixed selection both show the same result as what was suggested by the p-values, while forward selection has Rating added as a significant predictor. However, we can see the significant collinearity between Rating and Limit which is discussed in section 3. Because the p-value of Limit is smaller than that of Rating, we keep Limit while delete Rating.

To conclude, the predictors contribute to the response are Income, Limit, Cards, Age and Student.

4 How large is the effect of each predictor on Balance?

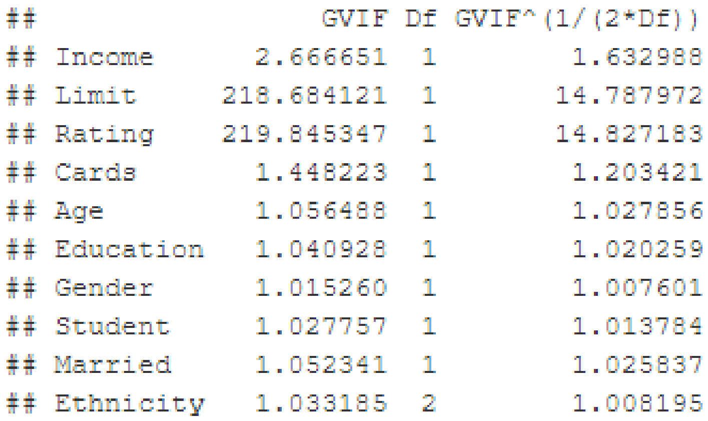
4.1 Confidence Interval

The standard error of $\hat{\beta}_j$ can be used to construct construct confidence interval for β_j. The confidence interval for β_j is

$$[\hat{\beta}_j - 2SE(\hat{\beta}_j), \hat{\beta}_j + 2SE(\hat{\beta}_j)]$$

The 95% confidence interval of the coefficients are displayed in Figure 2.

Notice that the interval of Income and Limit are narrow and far from 0, the interval of Cards and Student, though very large, are even far from 0. Thus, these predictors are effective on Balance. Moreover, the interval of Age, though



```
##                    GVIF Df GVIF^(1/(2*Df))
## Income      2.666651  1        1.632988
## Limit     218.684121  1       14.787972
## Rating    219.845347  1       14.827183
## Cards       1.448223  1        1.203421
## Age         1.056488  1        1.027856
## Education   1.040928  1        1.020259
## Gender      1.015260  1        1.007601
## Student     1.027757  1        1.013784
## Married     1.052341  1        1.025837
## Ethnicity   1.033185  2        1.008195
```

Figure 3: The VIF of the predictors

going through zero, is narrow and tend to the negative side obviously. So Age can also be recognized as slightly effective. On the contrary, all the other intervals include 0 and are large, with no obvious tendency to positive or negative side. Thus, all the other predictors can be viewed as not statistically significant effective to Balance.

4.2 Collinearity

We can check whether the large interval come from collinearity by calculating the VIF of each predictor. The result is displayed in Fiure 3.

We notice that Limit and Rating has a significant collinearity. We keep Limit while delete Rating. Fit the model again, and we find that Age is no more significant. The effective analysis is the same as before, Income, Limit, Cards and Student are still effective, while the rest are less effective.

4.3 Simple Linear Regression

In order to assess the association of each medium individually on sales, we can perform 10 separate simple linear regressions. We can find that there is a strong association between Balance and Income, Balance and Limit, Balance and Rating, Balance and Student. Cards and Age, in this case, lose their significant association with Balance.

ID	predictors
lm.fit1	.-ID
lm.fit2	Income+Limit+Cards+Age+Student
lm.fit3	.-ID-Rating
lm.fit4	Income+Limit+log(Limit)+Cards+Age+Student
lm.fit5	Income+log(Income)+Limit+log(Limit)+Limit^2+Cards+Age+Student
lm.fit6	Limit*Income+Student+Cards+Age
lm.fit7	Limit*Income+log(Limit)+Student+Cards+Age

Table 1: Model ID and the predictors relating to the response Balance. Here, ’.’ means all the potential predictors,’ a*b’ means a+b+a:b

ID	R^2	RSS_test
lm.fit1	0.9575	581684.8
lm.fit2	0.9569	555694
lm.fit3	0.9572	591577.4
lm.fit4	0.9764	353946.2
lm.fit5	0.9794	293361.8
lm.fit6	0.96	510149.7
lm.fit7	0.9804	312659.8

Table 2: R^2 and the RSS for test set of each model

5 How accurately can we predict future Balance?

The response can be predicted by

$$\hat{y} = \hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_i x_i$$

Now, we use the model derived from training data set to predict the Balance in test set. We can use the prediction interval for an individual response and confidence interval for an average response. Here, we use RSS of the prediction and the real test Balance to asses the accuracy of the model.

The RSS_test of several models(listed in Table 1) are listed in Table 2:

We can summarize from Table 2 that the best model is ’lm.fit5’ with the least test set RSS and ’lm.fit7’ with the least R^2.

We can still improve the R^2 by eliminating the outliers and leverage points (section 8).

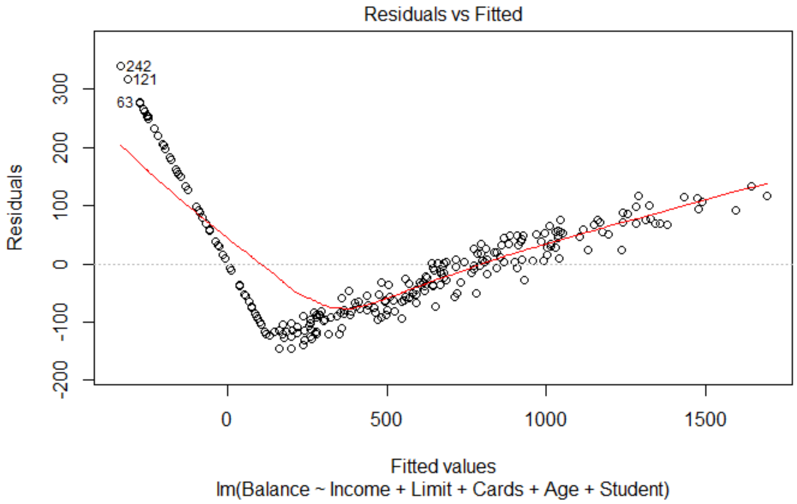


Figure 4: The Residual vs Fitted plot of 'lm.fit2'

6 Is the relationship linear?

The residual plots(Figure 4)can be used in order to identify non-linearity. Here we check the regression of Balance over Income, Limit, Cards, Age and Student('lm.fit2'). From the obvious pattern of the residual plot, we can conclude that the relationship is not linear. By adding square root, log, square to all the predictors, we find the best model with small p-value for every predictor is Balance over Income, log(Income), Limit, log(Limit), Limit^2, Cards, Age and Student('lm.fit5'). Among all the non-linear predictors added, log(Limit) is the crucial one to cancel the non-linearity effect('lm.fit4'), the residual plot of 'lm.fit4' is displayed in Figure 5, we can see that the pattern is smoother than Figure 4.

We can also use the function anova() to test whether the model fit the data equally well. In this case, we can see that 'lm.fit4' (the corresponding predictors can be found in Table 1)is significantly better than 'lm.fit2', while 'lm.fit5' is significantly better than 'lm.fit4'.

7 Is there synergy among the predictors?

From the pairs plot of Income and Limit we notice the relationship of interaction between the two. So we add the term Income:Limit on the model of section 6 and get a larger R^2. However, some of the non-linear predictors become insignificant. And after deleting those insignificant predictors, we get the best model: Balance

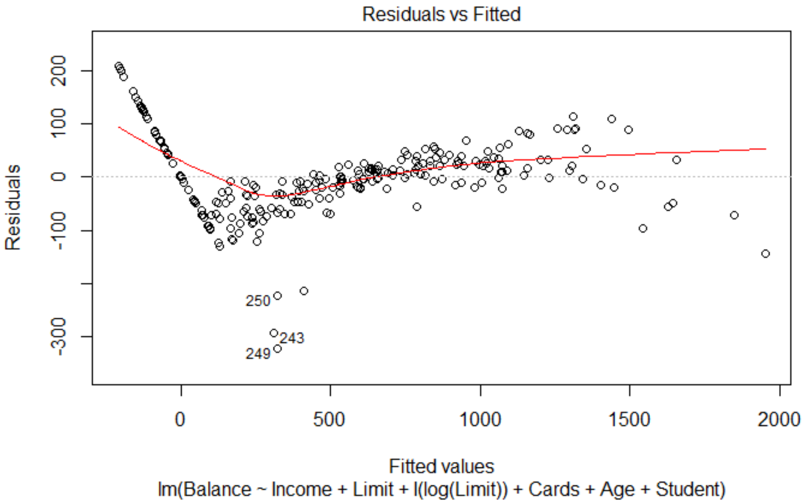


Figure 5: The Residual vs Fitted plot of 'lm.fit4'

over Income, Limit, Income:Limit, log(Limit), Cards, Age and student. All the predictors are significant and $R^2 = 98.04\%$.

8 Other problems to concern about

8.1 High Leverage Points

We can notice that in the Residual vs Leverage plot of model 'lm.fit7' (Figure 6), there is a high leverage point 29. Using the function hatvalues(), we can also find that observation 29 is a high leverage point. Thus, we exclude this point from the training set and use the same model to fit the data. In this way, $R^2 = 0.98$, RSS_test= 340742.4. It seems that the new model is no better than the previous one.

8.2 Outliers

We can also see from Figure 6 that there are 2 outliers. Using the function outlierTest(), we can find that observation 243 and 249 are the outliers. Delete these two from the training data and use 'lm.fit7' to fit. Then we find that $R^2 = 0.9843$ and RSS_test= 329894.3. R^2 is larger than the previous model, which means that after eliminating the outliers, the model fits better.

Using the same method, we delete the high leverage points and outliers from and use the model 'lm.fit5' to fit the new training set. After deleting the leverage

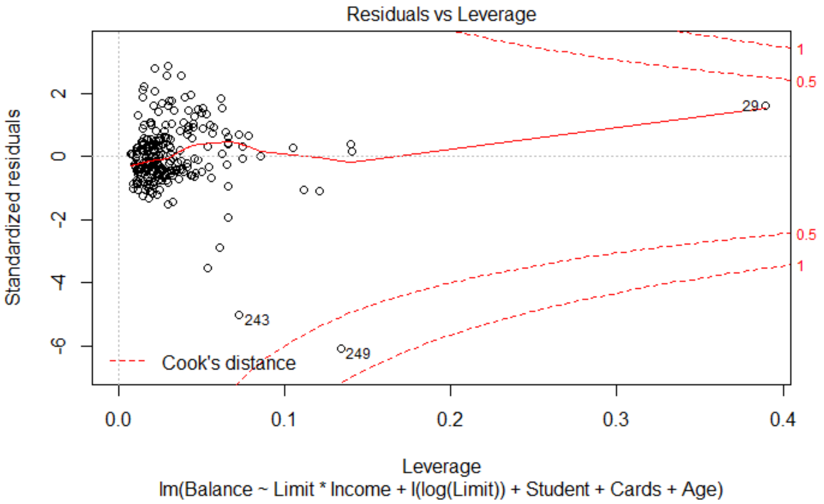


Figure 6: The Residual vs Leverage plot of model 'lm.fit7'

point 29, $R^2 = 0.9788$ and RSS_test= 293248.4. After deleting the outliers 243 and 249, $R^2 = 0.9843$ and RSS_test= 329894.3.

So, if we want to find the model that made the RSS of test set small, we should choose Balance$\sim$Income+log(Income)+Limit+log(Limit)+Limit^2+Cards+Age+Student over the training set without observation 29. If we want to find the model that made R^2 large, we should choose Balance $\sim$ Income+log(Income)+Limit+log(Limit)+Limit^2+Cards+Age+Stu or the model Balance $\sim$ Income*Limit+log(Limit)+Cards+Age+Student over the training set without observation $29, 243, 249$.