Chen, Te-Jung
Lee, Matthias
Wang, Chenxu
Wang, Chi
Wang, Roujian

Predicting Late Debt Payments with Logistic Regression

I. Executive Summary

The objective of this project is to identify the characteristics that best predict people who have late debt payments. As such, we decided to use the data from the 2013 Survey of Consumer Finances (SCF)[1]. The dataset contains a variable, *LATE60* that denotes if a household had a debt payment 60 days or more past due within the last year and is used as the dependent variable. The 12 predictor variables chosen are based on what the team thinks are simple, generic questions that could be asked by a bank to quickly determine if a person will have late debt payments. The predictor variables include level of education, number of kids in a household, race, spending habits within the past year, overall household expenses within the last year, income, if the respondent or his/her spouse has been turned down for credit or received only partial credit by a lender within the last 5 years, if the household owns any financial assets (defined in the Appendix), debt-to-income ratio, ratio of equity to normal income, and percentage of a home that is paid off. Our final model uses 9 of the 12 predictor variables including: level of education, number of kids in the household, spending habits within the past year, percentage of a home that is paid off, if a household has been turned down by a creditor, ratio of equity to normal income, debt-to-income ratio.

The odds ratio obtained from these variables support most of our initial presumptions. However, the model reveals a counter-intuitive aspect of two explanatory variables. Holding all other variables in the final model constant, a person with a college degree is more likely to have late debt payments relative to a person with no high school diploma. In addition, an individual who has a higher pencentage of a house paid off tends to have a higher odds of having late debt payments.

II. Data Characteristics:

***Table 1** –Definitions of Categorical Variables with Simple Summary Statistics*

Variables	Description	% Of Sample Data
Late60	Whether the person were ever behind in his/her payments by two months or more	
	0. No	93.92
	1. Yes	6.08
EDCL		
	1 = no high school diploma/GED	9.04
	2 = high school diploma/GED	26.62
	3 = some college	17.09
	4 = college degree	47.25
RACE		
	1 = white non-Hispanic,	73.53
	2 = black/African-American	12.42
	3 = Hispanic	9.23

[1] Data source: https://www.federalreserve.gov/econres/scfindex.htm

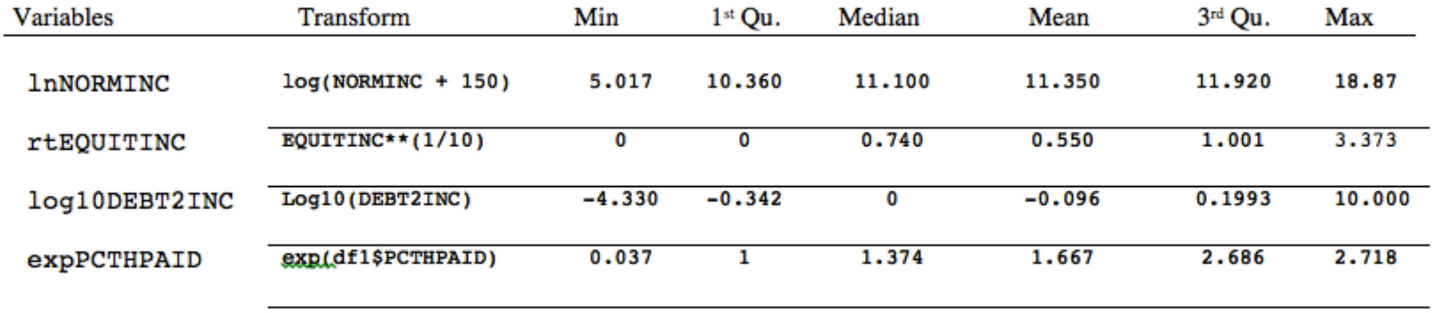
	5 = Others	4.82
WSAVED		
	1 = spending > income (within past year)	13.6
	2 = spending = income	27.1
	3 = spending < income	59.3
EXPENSHILO		
	1 = Overall household expenses unusually high	30.71
	2 = Overall household expenses unusually low	4.42
	3 = Overall household expenses is normal	64.87
KIDS		
	0 = 0 kids	56.79
	1 = 1 kid	17.54
	2 = 2 kids	15.21
	3 = 3 kids	7.05
	4 = 4 or more kids [2]	3.41
TURNDOWN		
	If the person (or his/her spouse) has been turned down or received partial credit by a creditor in the past five years	
	0 = No	84.74
	1 = Yes	15.26

Table II *– Definitions of Continuous Variables with Summary Statistics*

Variables	Description	Min	1st Qu.	Median	Mean	3rd Qu.	Max
NORMINC	Normal Income	0	31,450	65,940	707,700	150,300	157,300,000
EQUITINC	Equity-to-Normal Income Ratio	0	0	0.05	34.39	1.01	190,400
DEBT2INC	Debt-to-Income Ratio	0	0	0.38	1.42	1.58	549.90
PCTHPAID	Percentage of house paid off	3.31	0	0.32	0.41	0.99	1

Since most of the numeric continuous variables are quite skewed (refer to Fig 1. in Appendix), we transform them as follows:

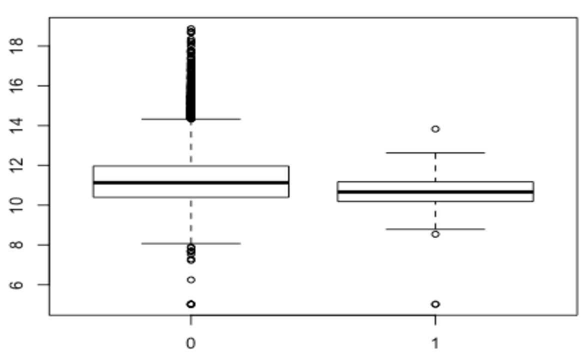
Table III *– Definitions of Transformed Continuous Variables with Summary Statistics*

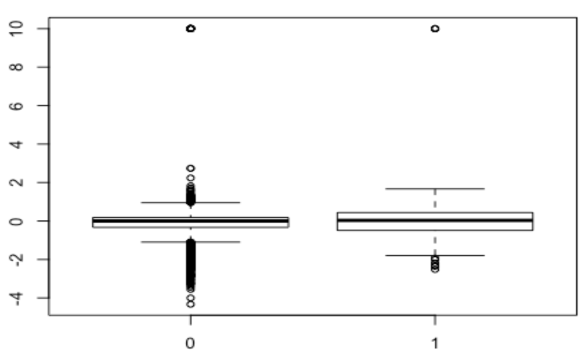


Variables	Transform	Min	1st Qu.	Median	Mean	3rd Qu.	Max
lnNORMINC	log(NORMINC + 150)	5.017	10.360	11.100	11.350	11.920	18.87
rtEQUITINC	EQUITINC**(1/10)	0	0	0.740	0.550	1.001	3.373
log10DEBT2INC	Log10(DEBT2INC)	-4.330	-0.342	0	-0.096	0.1993	10.000
expPCTHPAID	exp(df1$PCTHPAID)	0.037	1	1.374	1.667	2.686	2.718

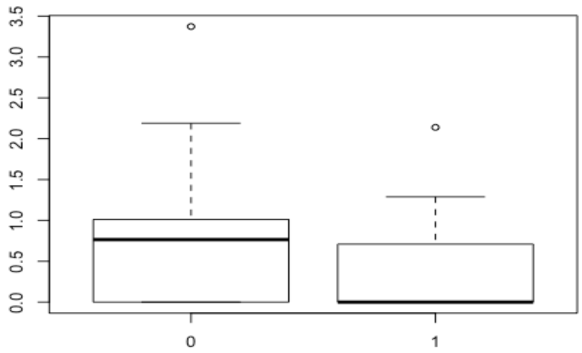
The plots of the transformed continous variables are shown below.

[2] Since households that have 4,5,6,7 and 8 kids around 3.4% of the entire data set, so we merged it into one as more than 4 kids.

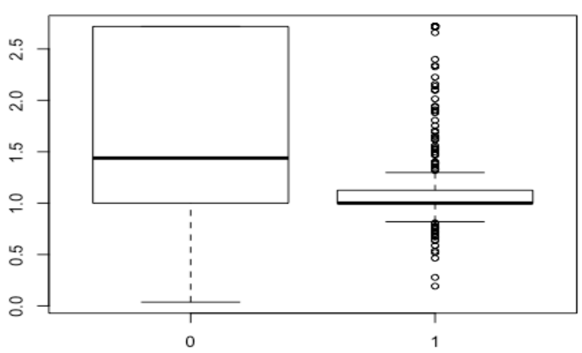




LATE60 vs rtEQUITINC



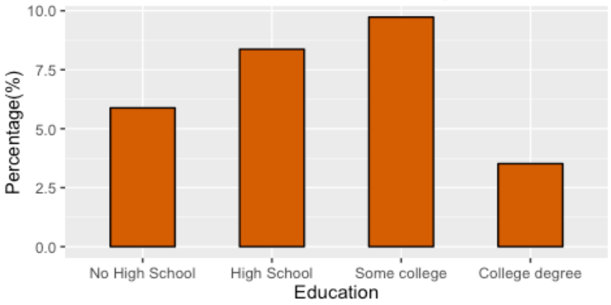
LATE60 vs expPCTHPAID



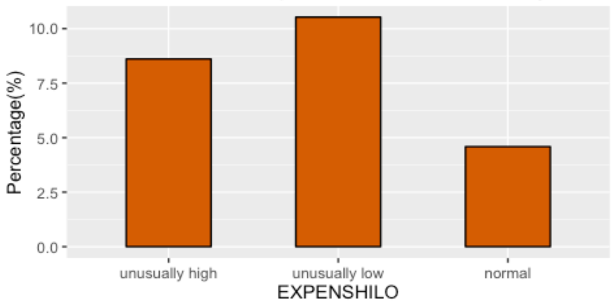
Visualizing Frequency Table:

We decided to visualize the frequency table to get a clearer picture. However, a frequency table can be obtained from Fig 3. in the Appendix. Observing the Spending Habits and Late Debt Payments plot suggests that individuals whose spending exceeds income has a higher percentage of having a late debt payment. However, the other cases aren't as significant as the large difference in WSAVED variable.

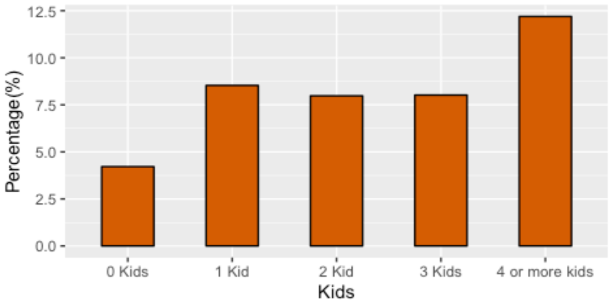
Education and Late Debt Payments



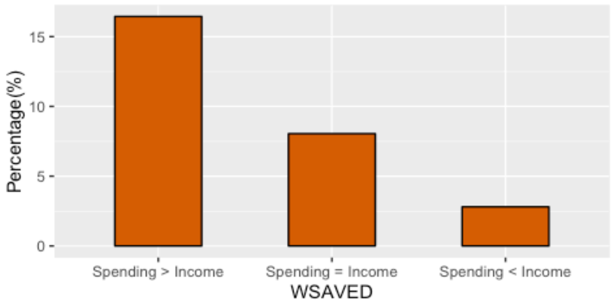
Household overall expenses and Late Debt Payments



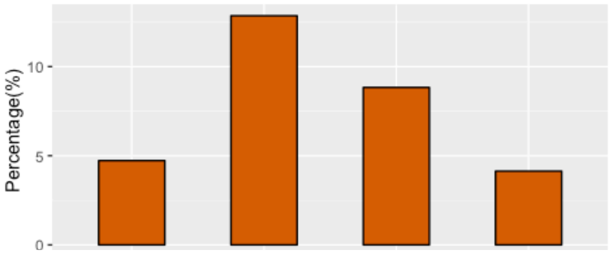
Kids and Late Debt Payments



Spending Habits and Late Debt Payments

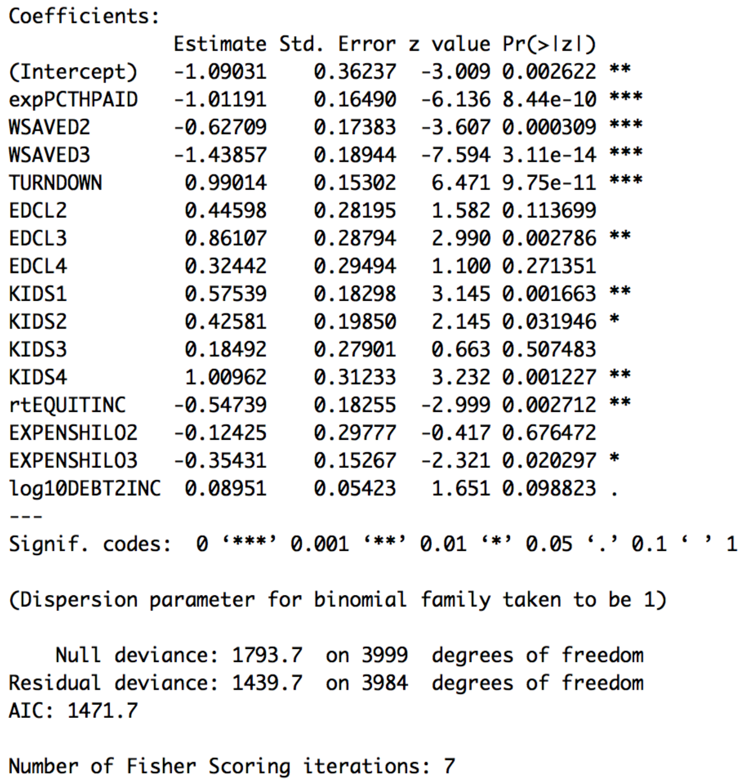


Race and Late Debt Payments



III. Outline of Analyses and Interpretation

After making a model based on these transformed data, running both backward and forward elimination, we have an optimized model with a better AIC value.



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.09031    0.36237  -3.009 0.002622 **
expPCTHPAID   -1.01191    0.16490  -6.136 8.44e-10 ***
WSAVED2       -0.62709    0.17383  -3.607 0.000309 ***
WSAVED3       -1.43857    0.18944  -7.594 3.11e-14 ***
TURNDOWN       0.99014    0.15302   6.471 9.75e-11 ***
EDCL2          0.44598    0.28195   1.582 0.113699
EDCL3          0.86107    0.28794   2.990 0.002786 **
EDCL4          0.32442    0.29494   1.100 0.271351
KIDS1          0.57539    0.18298   3.145 0.001663 **
KIDS2          0.42581    0.19850   2.145 0.031946 *
KIDS3          0.18492    0.27901   0.663 0.507483
KIDS4          1.00962    0.31233   3.232 0.001227 **
rtEQUITINC    -0.54739    0.18255  -2.999 0.002712 **
EXPENSHILO2   -0.12425    0.29777  -0.417 0.676472
EXPENSHILO3   -0.35431    0.15267  -2.321 0.020297 *
log10DEBT2INC  0.08951    0.05423   1.651 0.098823 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1793.7  on 3999  degrees of freedom
Residual deviance: 1439.7  on 3984  degrees of freedom
AIC: 1471.7

Number of Fisher Scoring iterations: 7
```

In order to interpret the coefficients of the logistic regression, we have to exponentiate them. Let’s take a look at one of the explanatory variable with multiple categories WSAVED2 that has a negative coefficient.. It’s coefficient is -0.622 and exponentiating it:

$$e^{-0.6271} = 0.5341$$

Holding all other variables constant and when WSAVED = 2, a household that has spending equal to income is about 0.53 times as likely to have a late debt payment relative to the baseline category, a household that has spending more than income. Similarly, for variable WSAVED3 *(spending less than income)* with coefficient of -1.4386 and exponentiating it:

$$e^{-1.4386} = 0.2373$$

We see that the odds ratio is lower than WSAVED2 and also means that a household whose spending is less than income is 0.23 as likely to have a late debt payment than compared to a household whose spending is more than income (baseline category)

On the other hand, the binary explanatory variable TURNDOWN has a positive coeffient of 0.9901 and exponentiating it:

$$e^{0.9901} = 2.6916$$

This implies that a household that got turned down or offered partial credit by a lender in the past 5 years is 2.69 times as likely to have a late debt payment as a household which did not get turned down. Alternatively, if we do the following:

$$\frac{1}{2.6916} = 0.3715$$

We can say that a household who did not get turned down by a creditor is 0.3715 times less likely to have a late debt payment to those who got turned down.

Correlation between log10DEBT2INC and LATE60 is 0.039. This suggests that a higher debt-to-income ratio does indeed have a higher probability of late debt payments. However, its correlation with the response variable is 0.039. Looking at the coefficient of log10DEBT2INC, we can see that it is not significant at the .05 level. It is also seen that the correlation 0.039 is very small and implies the less significant it is with the model. This is show to be true by observing the Beta of log10DEBT2INC on how significant it is to the model.

Also, notice that expPCTHPAID and LATE60 are negative correlated, which implies the more money one has paid for his/her house has a higher probability of a late debt payment. At first, this may seem counter-intuitive, but we must remember that the coefficients of the logistic regression models are only true when all other variables are held constant. Thus, if were to think about this a little more, it makes some sense. Particularly, comparing two households with the same normal income and debt-to-income ratio, an individual that owns more of its house implies that there must be other sources of debt. Perhaps, it is safe to assume that having other sources of debt correlates with higher odds of making a late debt payment.

As for the other variables, the odds ratio are as expected. The more kids a household has, the likelihood of a late debt payment increases; if a household has an unsually low or normal household expense within the past year, the probability of a late debt payment is less likely, etc. However with an exception for education level, the higher the education, the odds of having late debt payment is higher. This is in contrast with our initial assumption that those who have a college degree will have less debt payment. It is interesting to note that out of 6,015 respondents to the survey, 47% of them have college degrees.

Based on the misclassification rate table (Fig. 8 in the Appendix), with threshold of 0.5, 0.3, and 0.1, the lowest AIC fit1 with 9 explanatory variables seems to have a slightly better prediction model when observing the in-sample misclassification. However, in-sample misclassification rates tend to be underestimates. The out-of-sample misclassification confirms that fit1 does indeed have lower misclassification rates. With multiple models tested, fit1 is indeed the better prediction model based on it's lowest AIC and misclassification rates.

IV. Contribution Section:

Our team was formed from Piazza. Most of the team discussions were on Facebook and had a few physical meetings to facilitate team discussions.

[Order of the names are in Alphabetical order by last name]

Chen, Te-Jung**:** Join discussion about the ideas on how to approach the project. Understand the end goal, identify clear roles, collaborate with teammates, break down the problem and analyse the data in R-code such as misclassification. The major part of my role is to interoperate the betas of both categorical and numeric variables of the best model. For example, interoperating the Education categorical variable including the dummy variables and the rtEQUINTINC, ratio of equity to normal income, numeric variable. Also, formed an equation of the late debt payment (LATE60) of all the variables and the coefficients.

Lee, Matthias**:** Found the dataset and came up with the basic ideas of our project. Co-wrote most of the code and analysis with Wang, Chi.

Wang, Chenxu**:** Participated in the discussion and the decision of the main topic. She put forward the idea of using Q-Q plot to choose the transformation of variables and wrote the codes for this. She took part in the discussion of AIC and misclassification to choose the best model. She also participated in the interpretation of the coefficients of the final best model. Helped with the interpretation of the coefficients.

Wang, Chi: Co-wrote most of the code and analysis with Matthias Lee.

Wang, Roujian: Discuss about how to choose the model that is better, discuss about if the model is good or bad, analysis the possible reason that caused a bad model, discuss about the comparison of models, did some researches about how to form a statistical report, how to interpret data, variables and model.

IV. Conclusion

Overall, we must conclude that a linear classification model such as logistic regression may not necessarily be the best classification model to predict late debt payments. This can be concluded from a reasonably high out-of-sample and in-sample misclassification rate (with threshold 0.1) of about 0.65 and 0.53 respectively. Perhaps non-linear classification models like SVMs may be better suited for this dataset.

There are other methods of validating / optimizing the model such as performing LOOCV or k-fold cross validation to further validate the model and performing chi-square test to make this make this model more robust which are not carried out in this project. While the effects of many of these variables are intuitive, the coefficient of the EDCL and expPCTHPAID were counter-intuitive. One of the reason for the high misclassification rate can also be attributed to the low records of response variable being yes's, which amount to only about 6% of 6,015 observations.

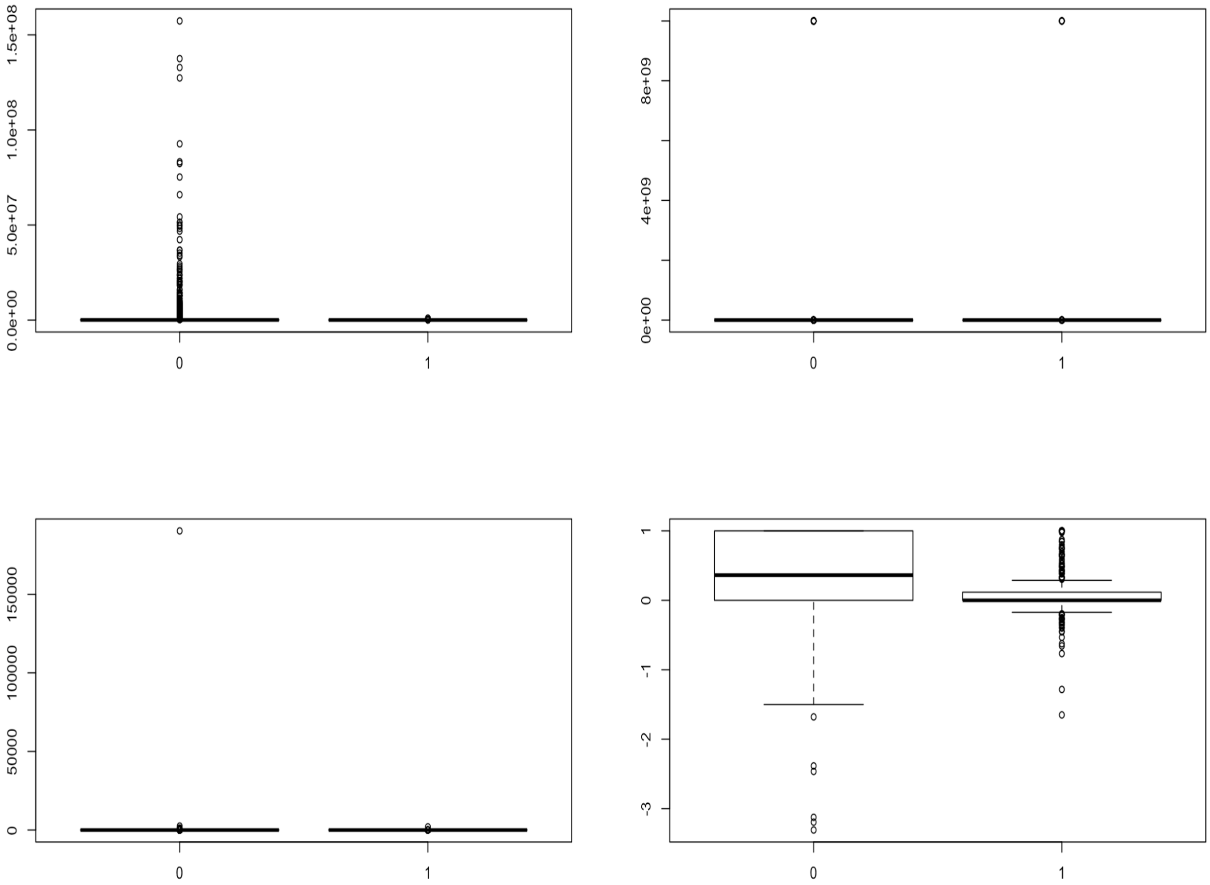
Nevertheless as a team, we learned about the process behind developing a logistic regression model, how to validate it, and most importantly interpreting the output of the model.

Appendix:

Variable Transformation:

```
#Plots of LATE60 vs numeric continuous variables before transformation
plot(factor(df1$LATE60), df1$NORMINC); plot(factor(df1$LATE60), df1$DEBT2INC);
plot(factor(df1$LATE60), df1$EQUITINC); plot(factor(df1$LATE60), df1$PCTHPAID)
```

Figure 1.



Correlation of explanatory variables to LATE60:

```
#Converting to a matrix matrix form to find correlation based on the training dataset
train_matrix_form <- data.matrix(train)
cor(train_matrix_form[,1],train_matrix_form[,c(2:12)])
```

Figure 2.

EDCL	KIDS	RACE	WSAVED	EXPENSHILO	lnNORMINC	TURNDOWN
-0.053	-0.833	0.049	-0.204	-0.088	-0.114	0.195

rtEQUITINC	HOTHFIN	expPCTHPAID	log10DEBT2INC
-0.135	0.011	-0.174	0.039

Frequency tables:

Figure 3.

print(table(LATE60, df3$KIDS))

	0	1	2	3	4
0	3272	965	842	390	180
1	144	90	73	34	25

print(table(df2$LATE60, df2$EDCL))

	1	2	3	4
0	512	1467	928	2742
1	32	134	100	100

print(table(df2$LATE60, df2$RACE))

	1	2	3	5
0	4214	651	506	278
1	209	96	49	12

print(table(df2$LATE60, df2$WSAVED))

	1	2	3
0	686	1479	3466
1	135	131	100

print(table(df2$LATE60, df2$TURNDOWN))

	0	1
0	4886	763
1	211	155

print(table(df2$LATE60, df2$EXPENSHILO))

	1	2	3
0	1688	238	3723
1	159	28	179

Variable Selection:

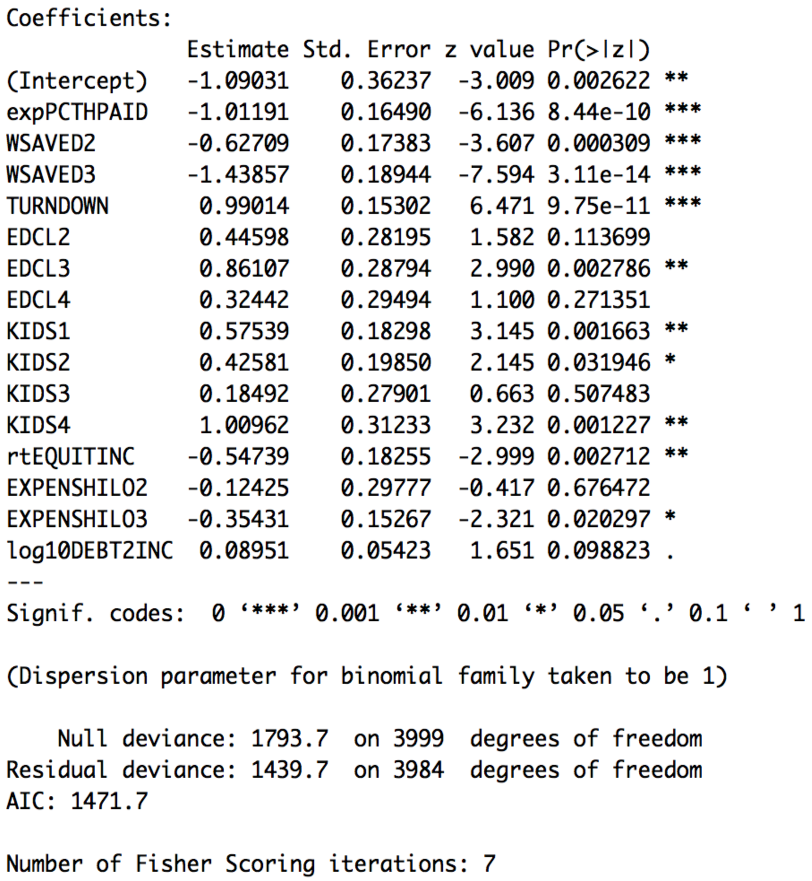
```
full.model <-
glm(LATE60~EDCL+KIDS+RACE+WSAVED+EXPENSHILO+TURNDOWN+HOTHFIN+expPCTHPAID+lnNORMINC+log10DEBT2INC+lnEQUITINC, family = "binomial", data = train)
null.model <- glm(LATE60~1,family="binomial", data = train)

#Step function variable selection chooses both => backwards and forwards elimination:
fit <-step(null.model, scope=list(lower=null.model, upper=full.model),direction="both")
```

<u>Testing different models:</u>

#fit1 is the the best AIC model that includes 8 explanatory variables KIDS, EDCL, WSAVED, #EXPENSHILO, TURNDOWN, expPCTHPAID, lnNORMINC

```
fit1 <-
glm(LATE60~expPCTHPAID+WSAVED+TURNDOWN+EDCL+KIDS+EXPENSHILO+lnEQUITINC+lnNORMINC+log10DEBT
2INC, family = "binomial", data = train)
summary(fit1)
```

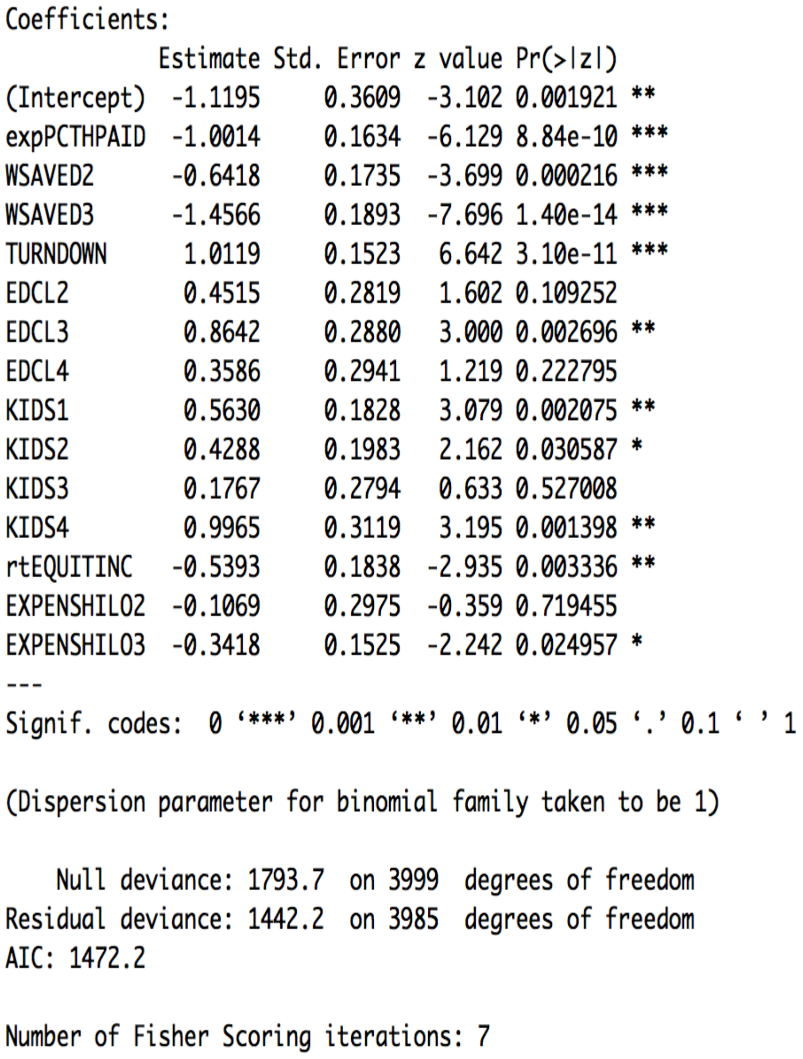
Figure 4.



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)   -1.09031    0.36237  -3.009 0.002622 ** 
expPCTHPAID   -1.01191    0.16490  -6.136 8.44e-10 ***
WSAVED2       -0.62709    0.17383  -3.607 0.000309 ***
WSAVED3       -1.43857    0.18944  -7.594 3.11e-14 ***
TURNDOWN       0.99014    0.15302   6.471 9.75e-11 ***
EDCL2          0.44598    0.28195   1.582 0.113699    
EDCL3          0.86107    0.28794   2.990 0.002786 ** 
EDCL4          0.32442    0.29494   1.100 0.271351    
KIDS1          0.57539    0.18298   3.145 0.001663 ** 
KIDS2          0.42581    0.19850   2.145 0.031946 *  
KIDS3          0.18492    0.27901   0.663 0.507483    
KIDS4          1.00962    0.31233   3.232 0.001227 ** 
rtEQUITINC    -0.54739    0.18255  -2.999 0.002712 ** 
EXPENSHILO2   -0.12425    0.29777  -0.417 0.676472    
EXPENSHILO3   -0.35431    0.15267  -2.321 0.020297 *  
log10DEBT2INC  0.08951    0.05423   1.651 0.098823 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1793.7  on 3999  degrees of freedom
Residual deviance: 1439.7  on 3984  degrees of freedom
AIC: 1471.7

Number of Fisher Scoring iterations: 7
```

```
#fit2 has 8 explanatory variables ,deleting log10DEBT2INC with backwards elimination:
fit2 <-
glm(LATE60~expPCTHPAID+WSAVED+TURNDOWN+EDCL+KIDS+EXPENSHILO+lnEQUITINC+lnNORMINC,
family = "binomial", data = train)
summary(fit2)
```

Figure 5.

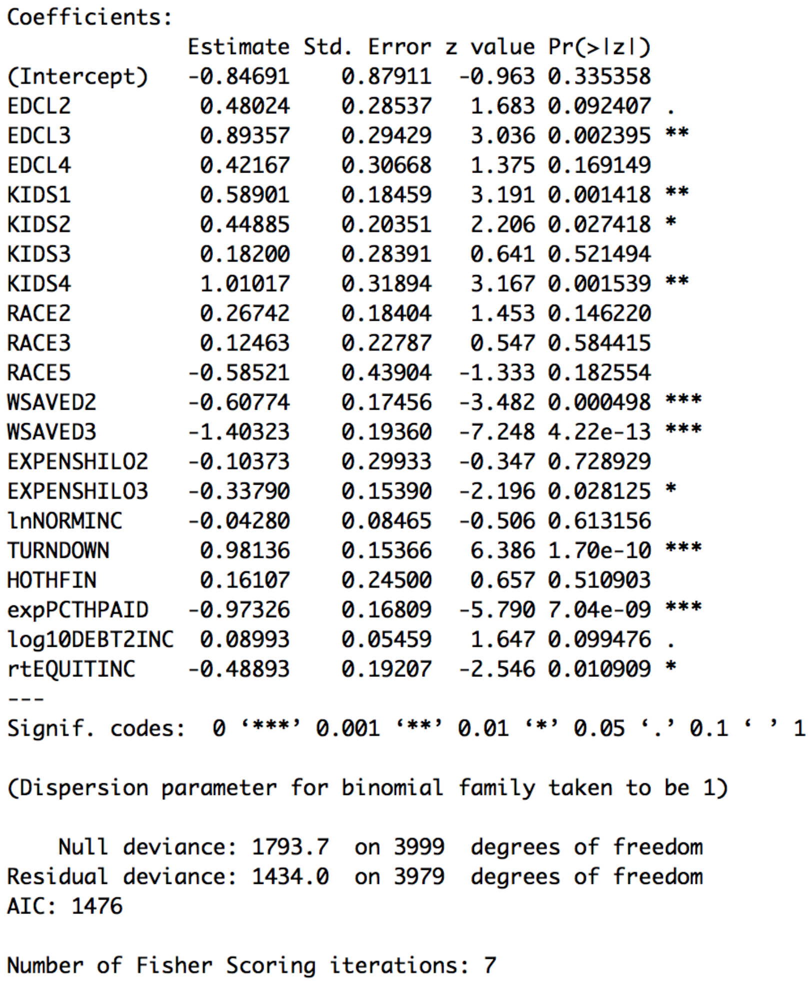


```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1195     0.3609  -3.102 0.001921 **
expPCTHPAID  -1.0014     0.1634  -6.129 8.84e-10 ***
WSAVED2      -0.6418     0.1735  -3.699 0.000216 ***
WSAVED3      -1.4566     0.1893  -7.696 1.40e-14 ***
TURNDOWN      1.0119     0.1523   6.642 3.10e-11 ***
EDCL2         0.4515     0.2819   1.602 0.109252
EDCL3         0.8642     0.2880   3.000 0.002696 **
EDCL4         0.3586     0.2941   1.219 0.222795
KIDS1         0.5630     0.1828   3.079 0.002075 **
KIDS2         0.4288     0.1983   2.162 0.030587 *
KIDS3         0.1767     0.2794   0.633 0.527008
KIDS4         0.9965     0.3119   3.195 0.001398 **
rtEQUITINC   -0.5393     0.1838  -2.935 0.003336 **
EXPENSHILO2  -0.1069     0.2975  -0.359 0.719455
EXPENSHILO3  -0.3418     0.1525  -2.242 0.024957 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1793.7  on 3999  degrees of freedom
Residual deviance: 1442.2  on 3985  degrees of freedom
AIC: 1472.2

Number of Fisher Scoring iterations: 7
```

#fit3 model the orignal model

fit3 <-
glm(LATE60~EDCL+KIDS+RACE+WSAVED+EXPENSHILO+TURNDOWN+HOTHFIN+expPCTHPAID+lnNORMINC+log10DEBT2INC+lnEQUITINC, family = "binomial", data = train)
summary(fit3)

Figure 6.



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)   -0.84691    0.87911  -0.963 0.335358    
EDCL2          0.48024    0.28537   1.683 0.092407 .  
EDCL3          0.89357    0.29429   3.036 0.002395 ** 
EDCL4          0.42167    0.30668   1.375 0.169149    
KIDS1          0.58901    0.18459   3.191 0.001418 ** 
KIDS2          0.44885    0.20351   2.206 0.027418 *  
KIDS3          0.18200    0.28391   0.641 0.521494    
KIDS4          1.01017    0.31894   3.167 0.001539 ** 
RACE2          0.26742    0.18404   1.453 0.146220    
RACE3          0.12463    0.22787   0.547 0.584415    
RACE5         -0.58521    0.43904  -1.333 0.182554    
WSAVED2       -0.60774    0.17456  -3.482 0.000498 ***
WSAVED3       -1.40323    0.19360  -7.248 4.22e-13 ***
EXPENSHILO2   -0.10373    0.29933  -0.347 0.728929    
EXPENSHILO3   -0.33790    0.15390  -2.196 0.028125 *  
lnNORMINC     -0.04280    0.08465  -0.506 0.613156    
TURNDOWN       0.98136    0.15366   6.386 1.70e-10 ***
HOTHFIN        0.16107    0.24500   0.657 0.510903    
expPCTHPAID   -0.97326    0.16809  -5.790 7.04e-09 ***
log10DEBT2INC  0.08993    0.05459   1.647 0.099476 .  
rtEQUITINC    -0.48893    0.19207  -2.546 0.010909 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1793.7  on 3999  degrees of freedom
Residual deviance: 1434.0  on 3979  degrees of freedom
AIC: 1476

Number of Fisher Scoring iterations: 7
```

Misclassification tables on fit1 and fit2 models:

Figure 7.

pred1 <- predict(fit1,type='response');
summary(pred1)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.001624	0.009399	0.026880	0.059000	0.072100	0.670600

pred2 <- predict(fit2, type = 'response');
summary(pred2)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.001697	0.009684	0.027240	0.059000	0.072040	0.679500

Figure 8.

In-Sample Misclassification tables

	π <= 0.5	π > 0.5	misclass	π <= 0.3	π > 0.3	misclass	π <= 0.5	π > 0.5	misclass
y= 0 (fit1)	3754	10	0.0027	3691	73	0.0194	3210	554	0.1472
y =1 (fit1)	228	8	0.9661	191	45	0.8093	94	142	0.3983
y = 0 (fit2)	3755	9	0.0024	3690	74	0.0197	3212	552	0.1467
y = 1 (fit2)	230	6	0.9746	192	44	0.8136	91	145	0.3856

Out-of-sample Misclassification tables

	π <= 0.5	π > 0.5	misclass	π <= 0.3	π > 0.3	misclass	π <= 0.5	π > 0.5	misclass
y= 0 (fit1)	1881	4	0.0021	1834	51	0.0271	1590	295	0.1565
y =1 (fit1)	125	5	0.9615	109	21	0.8384	66	64	0.5077
y = 0 (fit2)	1881	4	0.0021	1832	53	0.0281	1592	293	0.1554
y =1 (fit2)	126	4	0.9692	112	18	0.8615	66	64	0.50769