DATA 621 Business Analytics & Data Mining

Homework 4 Linear & Binary Logistic Regression

Kyle Gilde

4/16/2018

Contents

Overview

Your objective is to build **multiple linear regression** and **binary logistic regression models** on the training data to predict the probability that a person will crash their car and also the amount of money it will cost if the person does crash their car. You can only use the variables given to you (or variables that you derive from the variables provided).

1. DATA EXPLORATION

Examine the cases & variables

Data Dictionary

In this homework assignment, you will explore, analyze and model a data set containing 8161 records representing a customer at an auto insurance company. As the KIDSDRIV variable implies, each record represents a customer that could have multiple drivers on the account.

After excluding the `INDEX` variable, there are a total of 25 variables, 2 of which are response variables:

- `TARGET_FLAG` is a binary variable where a "1" means that the person was in a car crash. A "0" means that the person was not in a car crash.

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO	None
TARGET_AMT	If car was in a crash, what was the cost	None
AGE	Age of Driver	Very young people tend to be risky. Maybe very old people also.
BLUEBOOK	Value of Vehicle	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase probability of collision
CLM_FREQ	# Claims (Past 5 Years)	The more claims you filed in the past, the more you are likely to file in the future
EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend to drive more safely
HOMEKIDS	# Children at Home	Unknown effect
HOME_VAL	Home Value	In theory, home owners tend to drive more responsibly
INCOME	Income	In theory, rich people tend to get into fewer crashes
JOB	Job Category	In theory, white collar jobs tend to be safer
KIDSDRIV	# Driving Children	When teenagers drive your car, you are more likely to get into crashes
MSTATUS	Marital Status	In theory, married people drive more safely
MVR_PTS	Motor Vehicle Record Points	If you get lots of traffic tickets, you tend to get into more crashes
OLDCLAIM	Total Claims (Past 5 Years)	If your total payout over the past five years was high, this suggests future payouts will be high
PARENT1	Single Parent	Unknown effect
RED_CAR	A Red Car	Urban legend says that red cars (especially red sports cars) are more risky. Is that true?
REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably are a more risky driver.
SEX	Gender	Urban legend says that women have less crashes then men. Is that true?
TIF	Time in Force	People who have been customers for a long time are usually more safe.
TRAVTIME	Distance to Work	Long drives to work usually suggest greater risk
URBANICITY	Home/Work Area	Unknown
YOJ	Years on Job	People who stay at a job for a long time are usually more safe

Figure 1:

- `TARGET_AMT` is the cost of the accident. It is zero if the person did not crash their car.

Statistical Summary

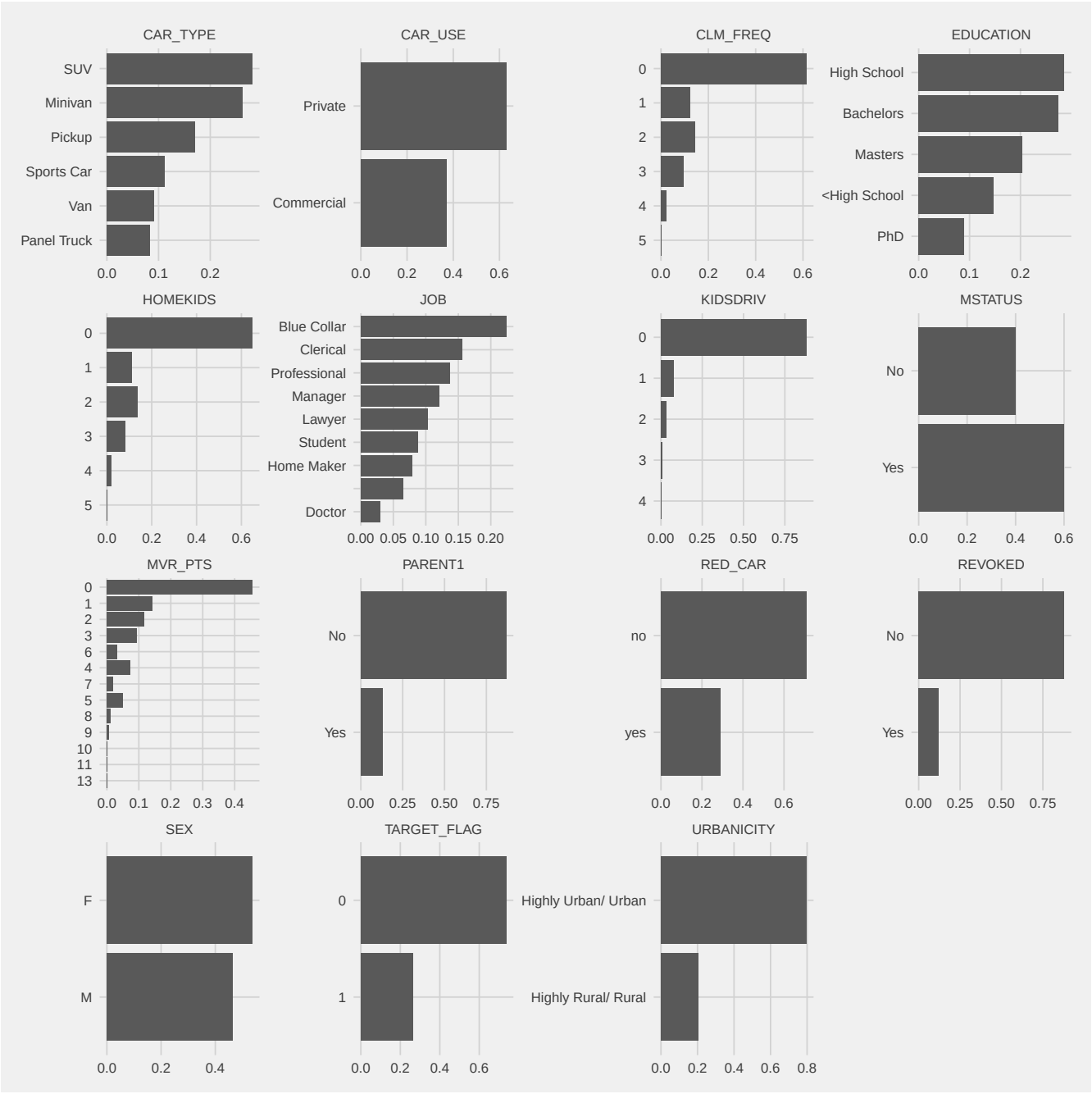
Among the numerical variables, we notice the following:

- `CAR_AGE`, `YOJ` & `AGE` appear to be missing values.
- Additionally, `CAR_AGE` has a minimum value of -3. We should consider imputing a value for this observation since that value isn't impossible.
- The cost of the accident `TARGET_AMT` has large skew and kurtosis values, which means that it may be a candidate for transformation.
- The small values for `INCOME`, `HOME_VAL` & `BLUEBOOK` indicate that this data set is a few decades old.

	n uni	que_values min	Q.1	st med	ian	mean Q.3	rd	max	range	sd ske	w
TARGET_FLAG	8,161	2	0	0	0	0.3	1	1	1	0.4	1.1
TARGET_AMT	8,161	1,949	0	0	0	1,504.3	1,036	107,586.1	107,586.1	4,704	8.7
KIDSDRIV	8,161	5	0	0	0	0.2	0	4	4	0.5	3.4
AGE	8,155	61	16	39	45	44.8	51	81	65	8.6	0
HOMEKIDS	8,161	6	0	0	0	0.7	1	5	5	1.1	1.3
YOJ	7,707	22	0	9	11	10.5	13	23	23	4.1	-1.
INCOME	8,161	6,613	1	926	2,817	2,875.6	4,701	6,613	6,612	2,090.7	0.1
HOME_VAL	8,161	5,107	1	2	1,245	1,684.9	3,164	5,107	5,106	1,697.4	0.5
TRAVTIME	8,161	97	5	22	33	33.5	44	142	137	15.9	0.4
BLUEBOOK	8,161	2,789	1	478	1,124	1,283.6	2,234	2,789	2,788	893.5	0.2
TIF	8,161	23	1	1	4	5.4	7	25	24	4.1	0.9
OLDCLAIM	8,161	2,857	1	1	1	552.3	1,015	2,857	2,856	862.2	1.3
CLM_FREQ	8,161	6	0	0	0	0.8	2	5	5	1.2	1.2
MVR_PTS	8,161	13	0	0	1	1.7	3	13	13	2.1	1.3
CAR_AGE	7,651	31	-3	1	8	8.3	12	28	31	5.7	0.3

Visualizations

Categorical & Discrete variables



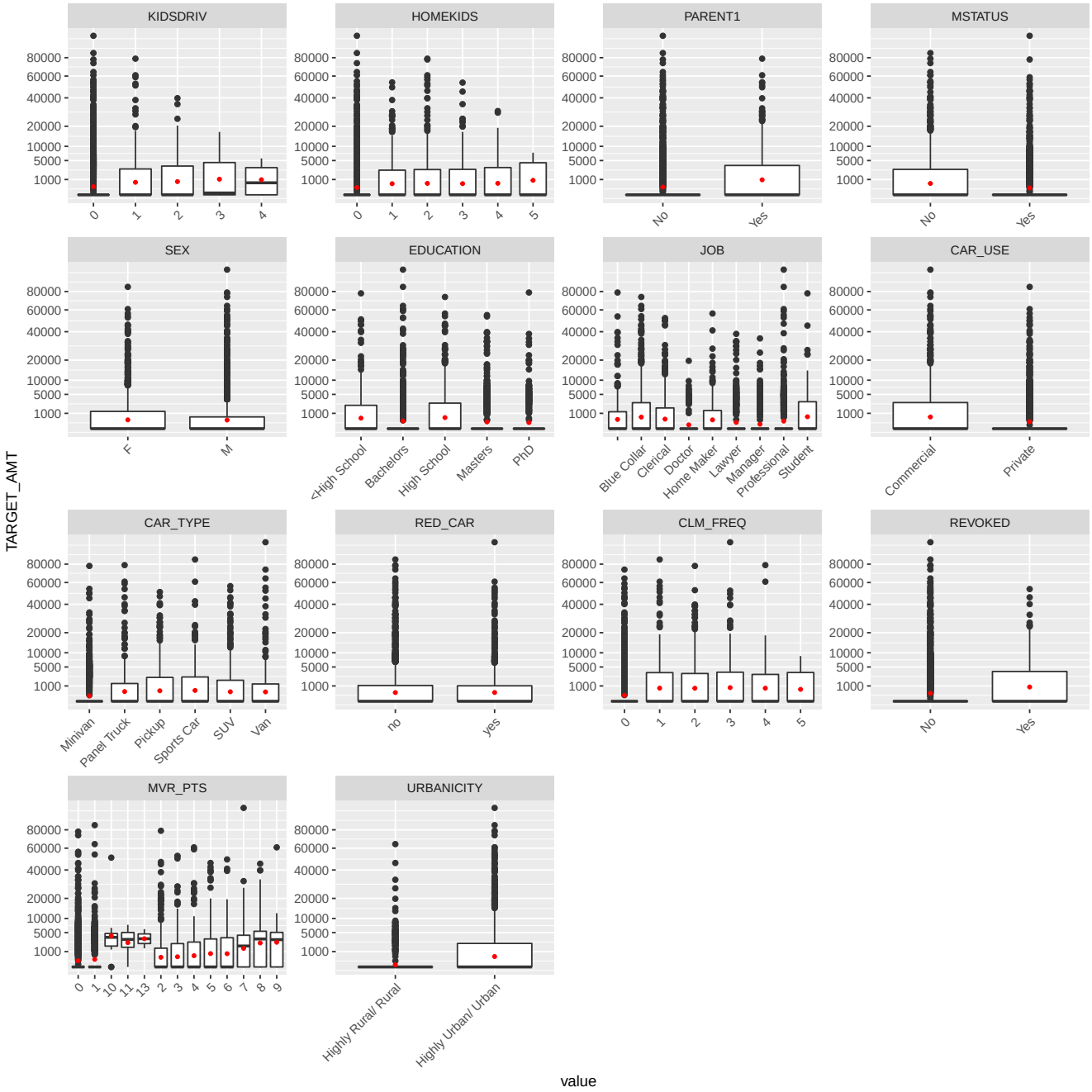
Side-by-Side Boxplots

The box plots contain the square root of `TARGET_AMT` distributions for each of the categorical and discrete variable values. We can see several several noteworthy differences in distributions that will likely inform some variable transformations.

The following characteristics significantly decrease the center of accident cost distribution compared to the other variable characteristics. We would these coeffients to be negative in our models.

- Having no kids at home in `HOMEKIDS` or no kids driving in `KIDSDRIV`
- Not being a single parent (`PARENT1`)
- Being married (`MSTATUS`)
- Having a Bachelors, Masters or PhD in `EDUCATION`

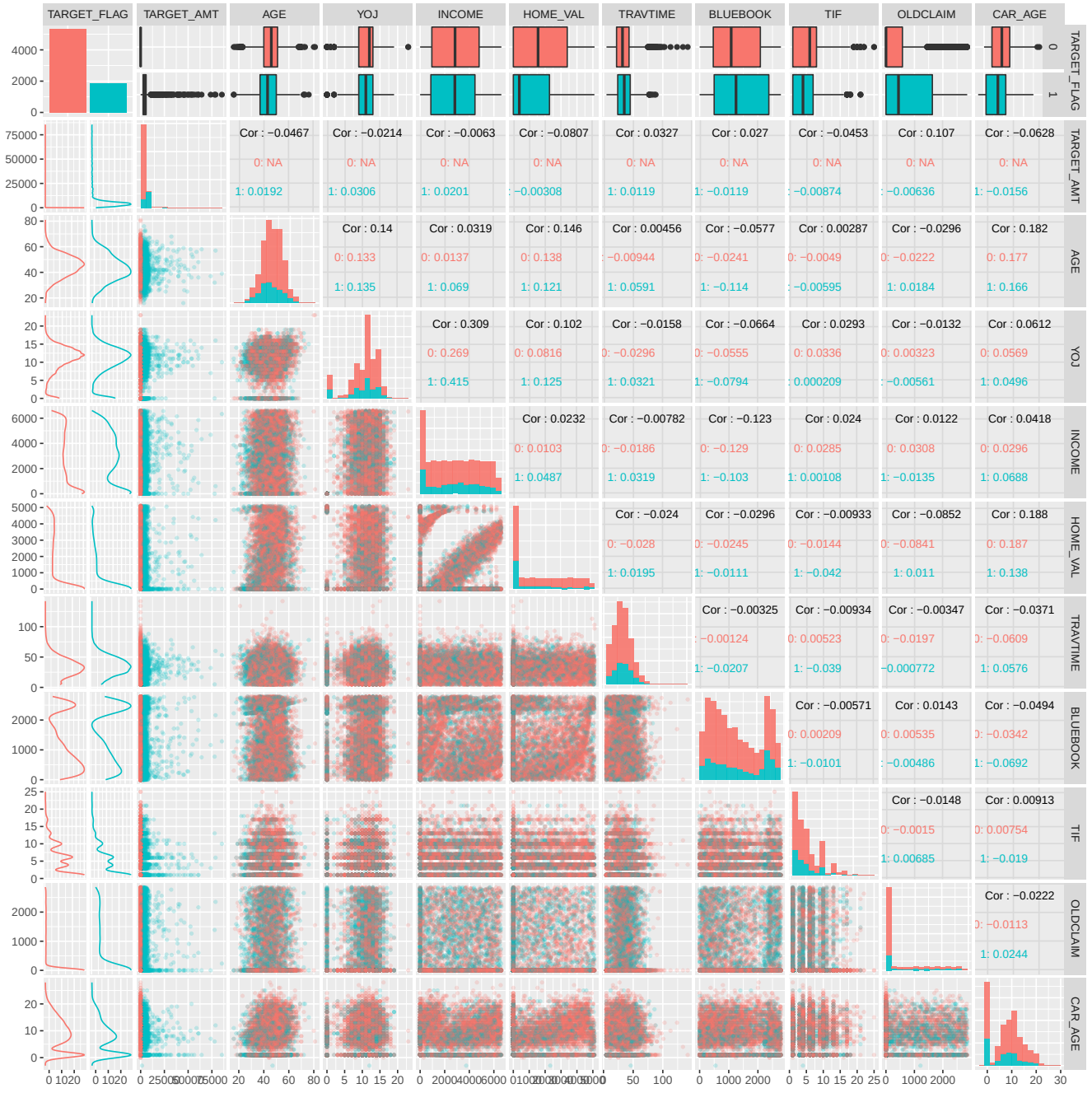
- Being a doctor, lawyer, manager or professional in `JOB`
- Not using the car for commercial purposes (`CAR_USE`)
- Driving a minivan `CAR_TYPE` & living in a rural area `URBANCITY`
- Having no claims in the last five years (`CLM_FRQ`) or not having had your licence revoked (`REVOKED`)
- Having 1 or less motor vehicle points (`MVR_PTS`)



Scatterplots, Histograms & Density Plots

Let's take a look at the relationships between our continuous variables. I have to give credit to classmate Jaan Bernberg for showing me the following `GGally` plot.

- First, the side-by-side **box plots** along the top row show surprisingly little difference in the variance and distribution between customers without an accident (in red) and those with an accident (in light blue). The only variable with moderately different different distributions is `OLDCLAIM`. Customers without accidents have smaller amounts of claim payouts in the last 5 years. Those who have not been in an accident also have a higher range of home values (`HOME_VAL`)
- In the histograms along the diagonal, other than `OLDCLAIM`, there are not severely different distributions between the 2 types of customers represented by `TARGET_FLAG`. The variables `BLUEBOOK` and `CAR_AGE` are bimodal & may benefit from transformations.
- The side-by-side density plots in the 1st column reveal that `INCOME` appears to be higher for customers who have had an accident. In our models, let's see if this variable defies conventional wisdom and is positively correlated with accidents.
- Next, let's take a look at the scatterplots in the 2nd column from the left where our continuous response variable `TARGET_AMT` is plotted against the other predictor variables. We don't see any pronounced positive or negative relationships. However, we see evidence that we may be able to make these relationships more linear by transforming `TARGET_AMT`.
- Lastly, we notice that the correlation values in the upper triangal section seem to be very small. We will confirm this in the next section.



Correlations

As we noted above, none of the variable pairs appear to have strong linear correlations. `YOJ` and `INCOME` are the most correlated at .31. We will likely not have collinearity among our original variables.

Table 2: Top Correlated Variable Pairs

	Var1	Var2	Correlation	Rsquared
1	YOJ	INCOME	0.31	0.10
2	HOME_VAL	CAR_AGE	0.19	0.04
3	AGE	CAR_AGE	0.18	0.03
4	AGE	HOME_VAL	0.15	0.02
5	AGE	YOJ	0.14	0.02

Var1	Var2	Correlation	Rsquared

The correlation coefficents with the response variable `TARGET_AMT` are even weaker. `OLDCLAIM` is only correlated with the response variable at .11.

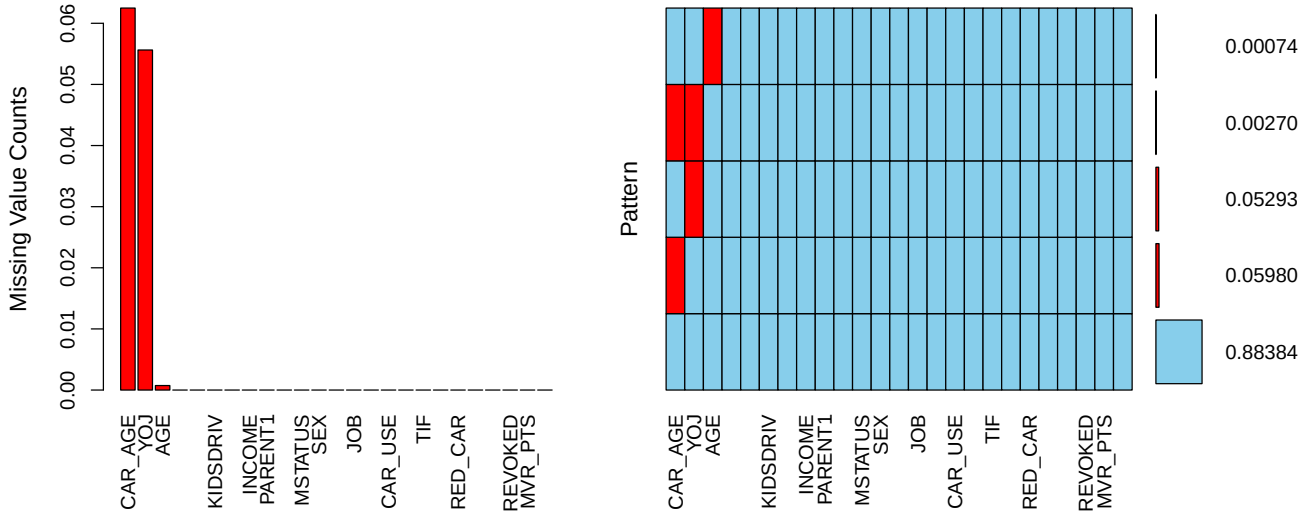
Table 3: Top Corrrelations with the Response Variable

	Var1	Var2	Correlation	Rsquared
1	TARGET_AMT	OLDCLAIM	0.11	0.01
2	TARGET_AMT	HOME_VAL	-0.08	0.01
3	TARGET_AMT	CAR_AGE	-0.06	0.00
4	TARGET_AMT	AGE	-0.05	0.00
5	TARGET_AMT	TIF	-0.05	0.00

Missing Values

Let's see how our missing variables are distributed.

- The variables for the age of the car `CAR_AGE` and years on the job `YOJ` are missing about 6% and 5% of their values. The customer age variable (`AGE`) is missing only 6 values. Otherwise, 88% of the cases are complete.
- Upon closer inspection, the `JOB` variable actually does not contain `NA`s. It does have a "blank" value. It's unclear whether the observation is missing. We will consider either replacing it with "Unknown" or imputing it.
- We should be able to use the other demographic variables to make reasonable imputations for these missing values.



2. DATA PREPARATION

Imputing the Missing Values

Let's use the `missForest` package to do nonparametric missing-value imputation using Random Forest on both the training and evaluation sets. We will set the impossible -3 `CAR_AGE` value to `NA` as well.

When we use the out-of-the-bag error in order to calculate the normalized root mean-squares error, we see NRMSE values closer to zero than not, which indicates that we have well-fitted imputations.

Variable	Count	MSE	NRMSE
CAR_AGE	510	16.20	0.14
YOJ	454	6.79	0.11
AGE	6	46.26	0.10

Variable Transformations

We will use the distribution differences seen in the side-by-side box plots to create the following new dummy variables. In order to avoid singularity issues, these new variables will replace their corresponding many-value categorical variables. We will run models using both the original variable set and the new variable data set to see which performs better.

- `NOHOMEKIDS`, `NOKIDSDRIV`, `HASCOLLEGE`, `ISPROFESSIONAL`, `ISMINIVAN`

3. BUILD MODELS

Linear Regression

LM #1: Original Variables with BIC Selection

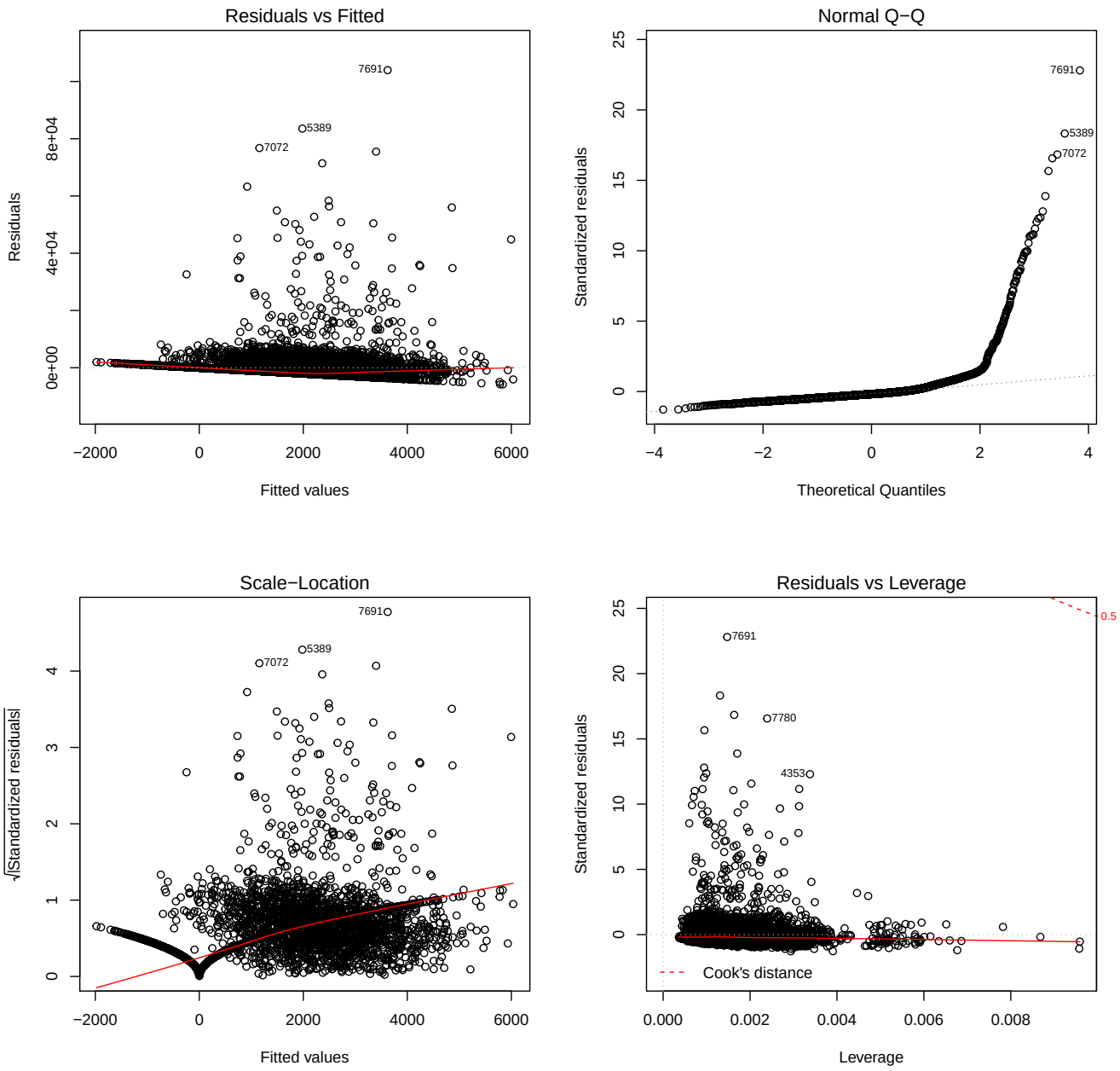
In our first model, we will use BIC forward and backward selection on the original variables with the missing values imputed. Twenty-six of the model variables were removed, leaving a model with eleven statically significant variables. In this model, being a single parent `PARENT1Yes` on average increases the cost of the accidents the most, and only using the insured vehicle for non-commercial purposes decreases the response variable the most. Despite having the highest correlation with the response variable `OLDCLAIM` was not statically significant. However, the model's adjusted R-squared is incredibly small at .06. This model does not account for very much of the variation in the cost of the accident `TARGET_AMT`.

```
##
## Call:
## lm(formula = TARGET_AMT ~ KIDSDRIV + PARENT1 + HOME_VAL + MSTATUS +
##     TRAVTIME + CAR_USE + TIF + REVOKED + MVR_PTS + CAR_AGE +
##     URBANICITY, data = dplyr::select(imputed_train$ximp, -TARGET_FLAG))
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5836  -1664   -827    275 103964
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  941.13962  222.44217   4.231 2.35e-05 ***
## KIDSDRIV                     377.38211  102.28462   3.690 0.000226 ***
## PARENT1Yes                   703.34692  176.07951   3.994 6.54e-05 ***
## HOME_VAL                      -0.10412    0.03304  -3.151 0.001634 **
## MSTATUSYes                  -438.26773  126.71698  -3.459 0.000546 ***
## TRAVTIME                      13.09238    3.22434   4.060 4.94e-05 ***
## CAR_USEPrivate              -848.17685  105.11429  -8.069 8.09e-16 ***
## TIF                          -46.86535   12.19943  -3.842 0.000123 ***
## REVOKEDYes                   495.89795  155.26068   3.194 0.001409 **
## MVR_PTS                      212.13760   24.06572   8.815  < 2e-16 ***
## CAR_AGE                      -51.03092    9.42700  -5.413 6.36e-08 ***
```

```
## URBANICITYHighly Urban/ Urban 1502.70221  131.29805  11.445  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4563 on 8149 degrees of freedom
## Multiple R-squared:  0.06052,    Adjusted R-squared:  0.05926
## F-statistic: 47.73 on 11 and 8149 DF,  p-value: < 2.2e-16
```

```
## [1] "removed variable(s): 26"
##  [1] "AGE"                   "HOMEKIDS"              "YOJ"
##  [4] "INCOME"                "SEXM"                  "EDUCATIONBachelors"
##  [7] "EDUCATIONHigh School" "EDUCATIONMasters"     "EDUCATIONPhD"
## [10] "JOBBlue Collar"        "JOBClerical"           "JOBDoctor"
## [13] "JOBHome Maker"         "JOBLawyer"             "JOBManager"
## [16] "JOBProfessional"       "JOBStudent"            "BLUEBOOK"
## [19] "CAR_TYPEPanel Truck"  "CAR_TYPEPickup"        "CAR_TYPESports Car"
## [22] "CAR_TYPESUV"           "CAR_TYPEVan"           "RED_CARyes"
## [25] "OLDCLAIM"              "CLM_FREQ"
```

In the model's diagnostic plots, the Residuals vs Fitted plot shows we have some nonconstant variance. The Normal Q-Q plot shows that the right side of the standardized residuals deviates from normality. The Leverage plot shows that we do not have any influential points.



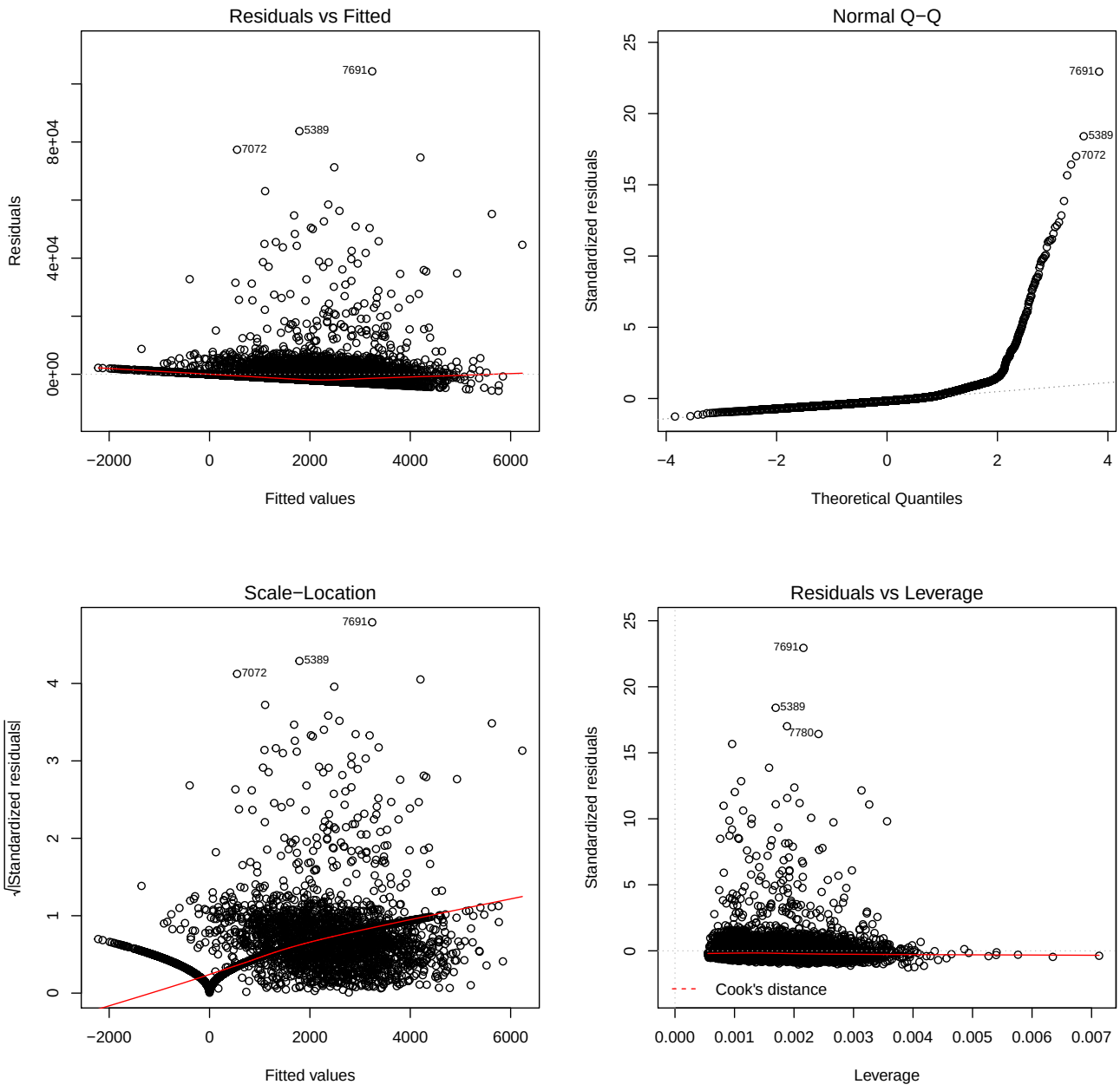
LM #2: Transformed Variables with BIC Selection

Our second model applied the same BIC selection processed to the transformed variable set with the five new variables. This version of the BIC model contains one more variable than the previous model, and it found three of the new variables to be statistically significant. However, adjusted R-squared is only slightly higher than the first model and still remains tiny.

```
##
## Call:
## lm(formula = TARGET_AMT ~ PARENT1 + MSTATUS + TRAVTIME + CAR_USE +
##     TIF + REVOKED + MVR_PTS + CAR_AGE + URBANICITY + NOKIDSDRIV +
##     ISPROFESSIONAL + ISMINIVAN, data = train_transformed)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5759  -1672   -790    272 104345
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    1604.767    274.723   5.841 5.38e-09 ***
## PARENT1Yes                      605.479    177.581   3.410 0.000654 ***
## MSTATUSYes                     -619.283    119.384  -5.187 2.18e-07 ***
## TRAVTIME                         12.778      3.219   3.969 7.27e-05 ***
## CAR_USEPrivate                 -592.297    113.393  -5.223 1.80e-07 ***
## TIF                             -47.497     12.175  -3.901 9.64e-05 ***
## REVOKEDYes                      487.327    154.874   3.147 0.001658 **
## MVR_PTS                         208.074     24.002   8.669  < 2e-16 ***
## CAR_AGE                         -42.060      9.901  -4.248 2.18e-05 ***
## URBANICITYHighly Urban/ Urban 1599.193    133.373  11.990  < 2e-16 ***
## NOKIDSDRIV                     -690.387    161.994  -4.262 2.05e-05 ***
## ISPROFESSIONAL                 -498.163    122.858  -4.055 5.06e-05 ***
## ISMINIVAN                      -539.684    117.309  -4.601 4.28e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4553 on 8148 degrees of freedom
## Multiple R-squared:  0.06451,    Adjusted R-squared:  0.06313
## F-statistic: 46.82 on 12 and 8148 DF,  p-value: < 2.2e-16
```

The diagnostic plots for the second linear model retain the nonconstant variance and normality issues. The leverage plot does not show any influential points.



LM #3: BIC Selection with Response Variable Transformation

Given the nonnormality and kurtosis we saw earlier, let's see what the car packages `powerTransform` function recommends for a negative Box-Cox transformation on `TARGET_AMT`. The negative Box-Cox transformation can accommodate all of the zero values that the variable has. For the sake of simplicity, let's add one to `TARGET_AMT` and take the natural log.

```
## Skew Power transformation to Normality
##
## Estimated power, lambda
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1   -0.2441      -0.244      -0.2511      -0.2371
##
```

```
## Estimated location, gamma
##    Est gamma Std Err. Wald Lower Bound Wald Upper Bound
## Y1    0.0101      NaN              NaN              NaN
##
## Likelihood ratio tests about transformation parameters
##                               LRT df pval
## LR test, lambda = (0)   5942.814  1    0
## LR test, lambda = (1) 143950.207  1    0
```

Our third model will use log-transformed `TARGET_AMT` with the newly created dummy variables. This model has two more statistically significant variables than the previous model, including all five of the transformed dummy variables we created. Moreover, adjusted R-squared is much higher at .21. This model explains much more of the variance in the response variable `TARGET_AMT`.

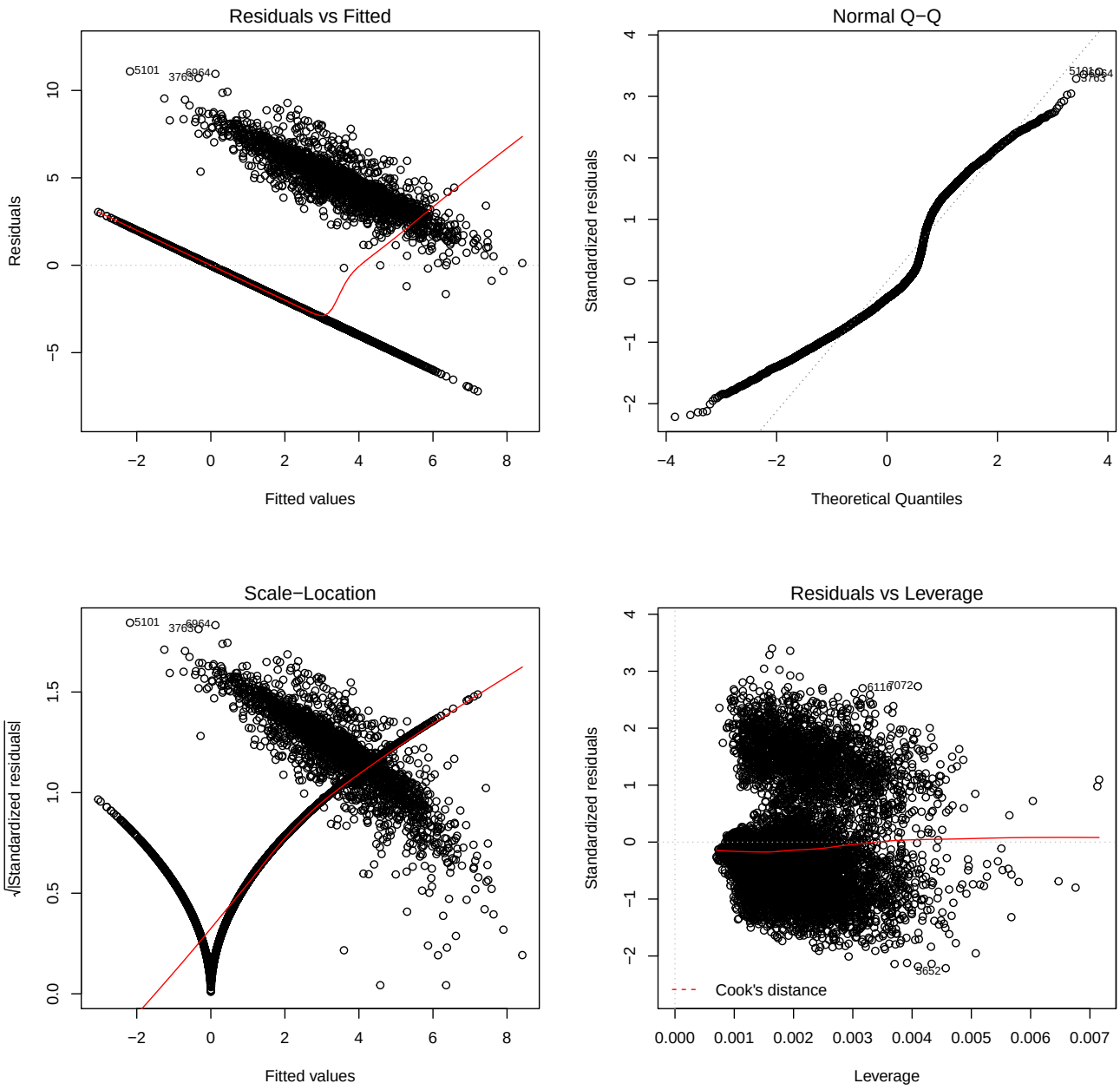
```
##
## Call:
## lm(formula = TARGET_AMT ~ YOJ + PARENT1 + HOME_VAL + MSTATUS +
##     TRAVTIME + CAR_USE + TIF + CLM_FREQ + REVOKED + MVR_PTS +
##     URBANICITY + NOHOMEKIDS + NOKIDSDRIV + HASCOLLEGE + ISPROFESSIONAL +
##     ISMINIVAN, data = train_transformed_BCN)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2088 -2.3507 -0.9222  2.2972 11.0897
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    2.510e+00  2.224e-01  11.283  < 2e-16 ***
## YOJ                           -4.119e-02  9.304e-03  -4.427 9.69e-06 ***
## PARENT1Yes                     5.178e-01  1.556e-01   3.328 0.000878 ***
## HOME_VAL                      -1.166e-04  2.432e-05  -4.793 1.67e-06 ***
## MSTATUSYes                    -6.164e-01  9.777e-02  -6.304 3.05e-10 ***
## TRAVTIME                       1.761e-02  2.309e-03   7.629 2.64e-14 ***
## CAR_USEPrivate                -6.164e-01  8.272e-02  -7.451 1.02e-13 ***
## TIF                           -6.591e-02  8.727e-03  -7.552 4.76e-14 ***
## CLM_FREQ                       2.099e-01  3.497e-02   6.002 2.04e-09 ***
## REVOKEDYes                     1.105e+00  1.111e-01   9.950  < 2e-16 ***
## MVR_PTS                        1.908e-01  1.848e-02  10.323  < 2e-16 ***
## URBANICITYHighly Urban/ Urban  2.292e+00  9.803e-02  23.380  < 2e-16 ***
## NOHOMEKIDS                    -3.371e-01  1.080e-01  -3.121 0.001811 **
## NOKIDSDRIV                    -7.080e-01  1.269e-01  -5.578 2.51e-08 ***
## HASCOLLEGE                    -7.479e-01  8.953e-02  -8.355  < 2e-16 ***
## ISPROFESSIONAL                -5.841e-01  9.799e-02  -5.961 2.61e-09 ***
## ISMINIVAN                     -7.681e-01  8.430e-02  -9.112  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.263 on 8144 degrees of freedom
## Multiple R-squared:  0.2118, Adjusted R-squared:  0.2102
## F-statistic: 136.8 on 16 and 8144 DF,  p-value: < 2.2e-16
```

After attempting to inverse the model coefficients, they still are not within the range of values we would expect given the distribution of the variables. This may require further research.

var	transformed_coef_value	inversed_coef_value
(Intercept)	2.51	11.3
URBANICITYHighly Urban/ Urban	2.29	8.89
REVOKEDYes	1.11	2.02

var	transformed_coef_value	inversed_coef_value
PARENT1Yes	0.52	0.68
CLM_FREQ	0.21	0.23
MVR_PTS	0.19	0.21
TRAVTIME	0.02	0.02
HOME_VAL	0	0
YOJ	-0.04	-0.04
TIF	-0.07	-0.06
NOHOMEKIDS	-0.34	-0.29
ISPROFESSIONAL	-0.58	-0.44
CAR_USEPrivate	-0.62	-0.46
MSTATUSYes	-0.62	-0.46
NOKIDSDRIV	-0.71	-0.51
HASCOLLEGE	-0.75	-0.53
ISMINIVAN	-0.77	-0.54

In the diagnostic plots for model #3, the log transformation did improve the normality of our residuals. However, it made the nonconstant variance much worse.



Binary Regression Models

Binary Regression Model #1: Backward Elimination with Original Variables

For our first model, let's use all of the original variables with a backward elimination process that removes the predictor with the highest p-value until all of the remaining p-values are statistically significant. This process removed seven predictors, leaving the model with 16 variables. `OLDCLAIM` has the most practical significance to the model with the largest coefficient. We can interpret the effect of of the variable as a dollar increase in `OLDCLAIM`, with the other variables held constant, increases the log-odds of the customer having had a car accident by 9.9. At -7.6, `CAR_USEPrivate` decreases the odds of a customer car accident the most.

```
##
##
```

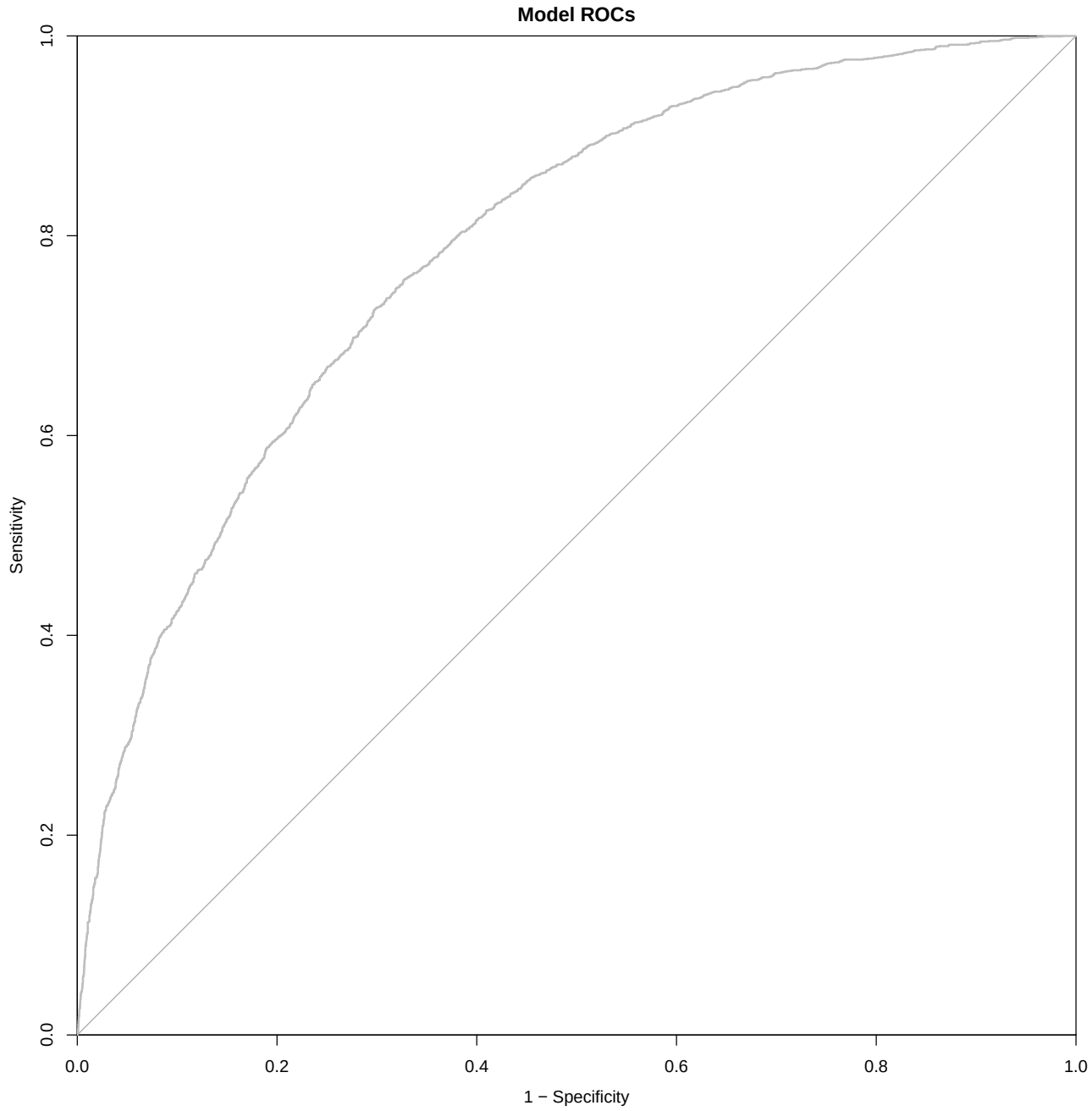
```
## removed_vars    removed_pvalues
## -------------  ----------------
## RED_CAR                  0.888
## EDUCATION                0.854
## BLUEBOOK                 0.618
## INCOME                   0.342
## JOB                      0.341
## CAR_TYPE                 0.733
## AGE                      0.224
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + TRAVTIME + CAR_USE + TIF + OLDCLAIM +
##     CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY, family = "binomial",
##     data = dplyr::select(imputed_train$ximp, -TARGET_AMT))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4323  -0.7444  -0.4460   0.7174   2.9576
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -1.969e+00  1.613e-01 -12.208  < 2e-16 ***
## KIDSDRIV                       3.140e-01  5.827e-02   5.389 7.08e-08 ***
## HOMEKIDS                       1.249e-01  3.310e-02   3.773 0.000161 ***
## YOJ                           -4.433e-02  7.108e-03  -6.237 4.45e-10 ***
## PARENT1Yes                     3.401e-01  1.058e-01   3.216 0.001300 **
## HOME_VAL                      -1.355e-04  1.903e-05  -7.124 1.05e-12 ***
## MSTATUSYes                    -3.727e-01  7.481e-02  -4.982 6.28e-07 ***
## SEXM                          -2.836e-01  6.017e-02  -4.714 2.43e-06 ***
## TRAVTIME                       1.470e-02  1.831e-03   8.028 9.94e-16 ***
## CAR_USEPrivate                -7.647e-01  6.033e-02 -12.676  < 2e-16 ***
## TIF                           -5.124e-02  7.143e-03  -7.174 7.28e-13 ***
## OLDCLAIM                       9.902e-05  4.152e-05   2.385 0.017079 *
## CLM_FREQ                       1.235e-01  3.130e-02   3.946 7.94e-05 ***
## REVOKEDYes                     7.666e-01  7.803e-02   9.824  < 2e-16 ***
## MVR_PTS                        1.129e-01  1.337e-02   8.440  < 2e-16 ***
## CAR_AGE                       -4.725e-02  5.347e-03  -8.837  < 2e-16 ***
## URBANICITYHighly Urban/ Urban  2.126e+00  1.107e-01  19.209  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7642.5  on 8144  degrees of freedom
## AIC: 7676.5
##
## Number of Fisher Scoring iterations: 5
```

The initial model has a cross-validation accuracy of .78 and area under the receiver operating characteristic curve of .79. Having no variance inflation factor greater than four `VIF_gt_4`, this model has no substantial collinearity, and a large p-value for the Hosmer-Lemeshow goodness-of-fit test, which means that we fail to reject the null hypothesis of the model having a good fit.

model_name	n_vars	model_pvalue	residual_deviance	H_L_pvalue	VIF_gt_4	CV_accuracy	AUC

model_name	n_vars	model_pvalue	residual_deviance	H_L_pvalue	VIF_gt_4	CV_accuracy	AUC
bk_elim_orig_vars	16	0.00e+00	7642.47	0.841	0	0.778	0.788



Binary Regression Model #2: Backward Elimination with Transformed Variables

Next, let's apply the same backward elimination process to the transformed set of variables. This process removed six variables, leaving the model with 17 statistically significant variables. All five of the the newly created dummy variables were included. `OLDCLAIM` still has the largest positive coefficient, but in this model,`HOME_VAL` followed by `ISMINIVAN` decreases odds of a customer car accident the most.

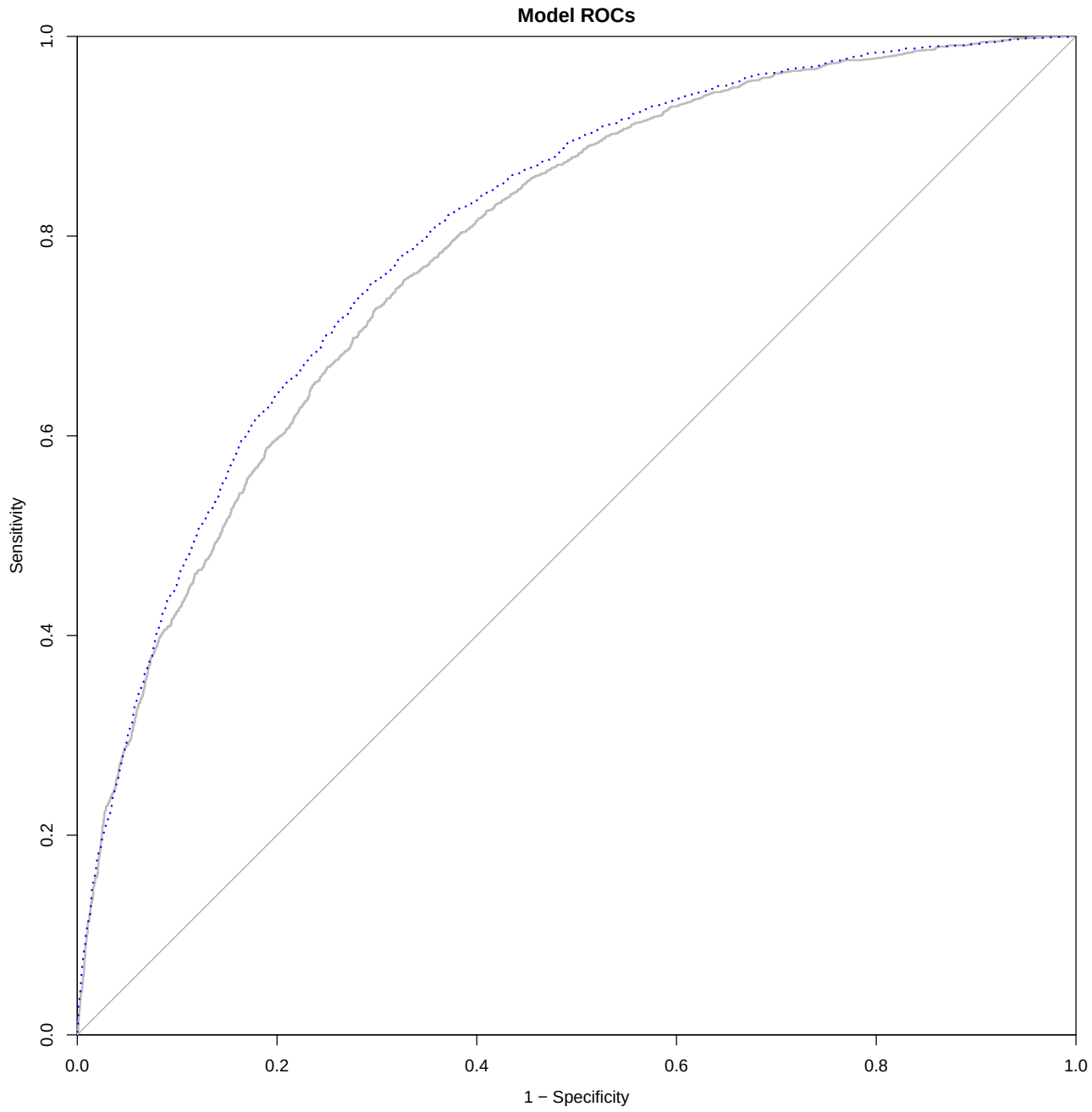
```
##
##
## removed_vars    removed_pvalues
## -------------  ----------------
## BLUEBOOK                  0.994
## AGE                       0.978
## RED_CAR                   0.927
## INCOME                    0.904
## SEX                       0.265
## CAR_AGE                   0.176
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ YOJ + PARENT1 + HOME_VAL + MSTATUS +
##     TRAVTIME + CAR_USE + TIF + OLDCLAIM + CLM_FREQ + REVOKED +
##     MVR_PTS + URBANICITY + NOHOMEKIDS + NOKIDSDRIV + HASCOLLEGE +
##     ISPROFESSIONAL + ISMINIVAN, family = "binomial", data = dplyr::select(train_transformed,
##     -TARGET_AMT))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3382  -0.7274  -0.4138   0.6820   3.2234
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -1.411e+00  1.868e-01  -7.553 4.26e-14 ***
## YOJ                          -3.300e-02  7.226e-03  -4.568 4.93e-06 ***
## PARENT1Yes                    2.363e-01  1.188e-01   1.988 0.046816 *
## HOME_VAL                     -9.406e-05  1.948e-05  -4.827 1.38e-06 ***
## MSTATUSYes                   -5.270e-01  7.914e-02  -6.658 2.77e-11 ***
## TRAVTIME                      1.510e-02  1.863e-03   8.106 5.25e-16 ***
## CAR_USEPrivate               -4.446e-01  6.295e-02  -7.063 1.63e-12 ***
## TIF                          -5.430e-02  7.237e-03  -7.503 6.23e-14 ***
## OLDCLAIM                      8.934e-05  4.204e-05   2.125 0.033556 *
## CLM_FREQ                      1.172e-01  3.181e-02   3.684 0.000229 ***
## REVOKEDYes                    7.742e-01  7.946e-02   9.743  < 2e-16 ***
## MVR_PTS                       1.123e-01  1.357e-02   8.279  < 2e-16 ***
## URBANICITYHighly Urban/ Urban  2.259e+00  1.110e-01  20.345  < 2e-16 ***
## NOHOMEKIDS                   -3.159e-01  8.741e-02  -3.614 0.000301 ***
## NOKIDSDRIV                   -5.277e-01  9.427e-02  -5.598 2.17e-08 ***
## HASCOLLEGE                   -5.808e-01  6.894e-02  -8.426  < 2e-16 ***
## ISPROFESSIONAL               -4.582e-01  7.636e-02  -6.001 1.96e-09 ***
## ISMINIVAN                    -7.028e-01  7.338e-02  -9.578  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7444.6  on 8143  degrees of freedom
## AIC: 7480.6
##
## Number of Fisher Scoring iterations: 5
```

In the second model,both the cross-validation accuracy and the area under the receiver operating characteristic curve have increased from the first model. The blue line of the area under the curve is closer to the upper right corner, indicating that this model is better at identifying true positives and negatives. The model has no substantial

collinearity, and a large p-value from the Hosmer-Lemeshow goodness-of-fit test means that we fail to reject the null hypothesis of the model having a good fit.

model_name	n_vars	model_pvalue	residual_deviance	H_L_pvalue	VIF_gt_4	CV_accuracy	AUC
bk_elim_orig_vars	16	0.00e+00	7642.470	0.841	0	0.778	0.788
bk_elim_transf_vars	17	0.00e+00	7444.605	0.766	0	0.782	0.803



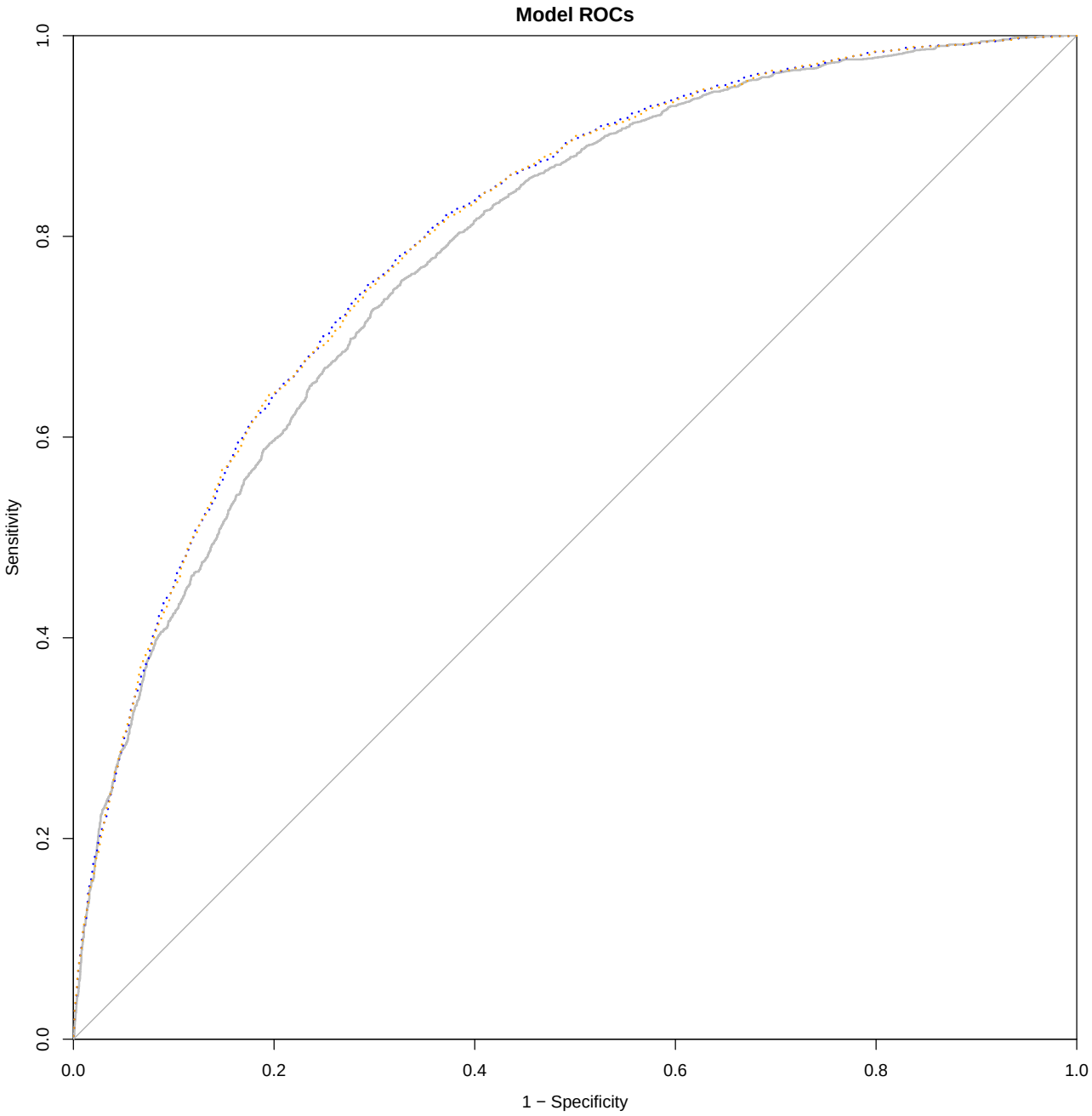
Binary Regression Model #3: Transformed Variables with BIC

For our third model, let's apply the BIC selection process to the transformed data set. This model has two variables less than the previous one and has 15 statistically significant variables. All five of the newly created dummy variables were included.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ YOJ + HOME_VAL + MSTATUS + TRAVTIME +
##     CAR_USE + TIF + CLM_FREQ + REVOKED + MVR_PTS + URBANICITY +
##     NOHOMEKIDS + NOKIDSDRIV + HASCOLLEGE + ISPROFESSIONAL + ISMINIVAN,
##     family = "binomial", data = dplyr::select(train_transformed,
##         -TARGET_AMT))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2641  -0.7257  -0.4158   0.6860   3.2336
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -1.247e+00  1.688e-01  -7.385 1.53e-13 ***
## YOJ                           -3.313e-02  7.219e-03  -4.589 4.46e-06 ***
## HOME_VAL                      -9.366e-05  1.944e-05  -4.817 1.46e-06 ***
## MSTATUSYes                    -6.203e-01  6.395e-02  -9.700  < 2e-16 ***
## TRAVTIME                       1.497e-02  1.861e-03   8.046 8.58e-16 ***
## CAR_USEPrivate                -4.428e-01  6.290e-02  -7.039 1.94e-12 ***
## TIF                           -5.438e-02  7.233e-03  -7.518 5.55e-14 ***
## CLM_FREQ                       1.572e-01  2.519e-02   6.242 4.31e-10 ***
## REVOKEDYes                     7.578e-01  7.911e-02   9.580  < 2e-16 ***
## MVR_PTS                        1.171e-01  1.339e-02   8.748  < 2e-16 ***
## URBANICITYHighly Urban/ Urban  2.262e+00  1.105e-01  20.462  < 2e-16 ***
## NOHOMEKIDS                    -4.253e-01  6.833e-02  -6.224 4.84e-10 ***
## NOKIDSDRIV                    -5.182e-01  9.385e-02  -5.521 3.37e-08 ***
## HASCOLLEGE                    -5.811e-01  6.892e-02  -8.432  < 2e-16 ***
## ISPROFESSIONAL                -4.568e-01  7.630e-02  -5.987 2.13e-09 ***
## ISMINIVAN                     -7.042e-01  7.328e-02  -9.610  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7453.2  on 8145  degrees of freedom
## AIC: 7485.2
##
## Number of Fisher Scoring iterations: 5
```

The third model essentially matches the cross-validation accuracy and the area under the receiver operating characteristic curve of the second model. The orange line of the area under the curve closely overlaps the blue line from the previous model. The model has no substantial collinearity, and it has a large p-value for the Hosmer-Lemeshow goodness-of-fit test, which means that we fail to reject the null hypothesis of the model having a good fit.

model_name	n_vars	model_pvalue	residual_deviance	H_L_pvalue	VIF_gt_4	CV_accuracy	AUC
bk_elim_orig_vars	16	0.00e+00	7642.470	0.841	0	0.778	0.788
bk_elim_transf_vars	17	0.00e+00	7444.605	0.766	0	0.782	0.803
BIC_transf_var_glm	15	0.00e+00	7453.214	0.693	0	0.782	0.803



4. SELECT MODELS

Evaluate Linear Models

Even though we still need to confirm how to interpret the coefficients, the third linear model performed the best of the three. Not only is the adjusted R-squared much higher, but the predicted R-squared is nearly as high, which indicates that we have not overfit the model.

Table 9: Model Summary Statistics

model_name	n_predictors	numdf	fstat	p.value	adj.r.squared	pre.r.squared
BIC_orig_var_lm	11	11	47.726	3.76e-102	0.059	0.058

model_name	n_predictors	numdf	fstat	p.value	adj.r.squared	pre.r.squared
BIC_transf_var_lm	12	12	46.821	1.11e-108	0.063	0.061
BIC_BCN_transf_var_lm	16	16	136.758	0.00e+00	0.210	0.208

The table below contains the diagnostic statistics for the models. All three do not have collinearity issues since none of their variables' variance inflation factors were over 4. Additionally, the moderate p-values for the Durbin-Watson test of independence indicate that we fail to reject the null hypothesis of no autocorrelation. However, the small p-values for the noncontant variance test from the `car` package and the Anderson-Darling test indicate that we would reject the null hypotheses of homoscedasticity and normality.

Table 10: Model Summary Statistics

DW.test	NCV.test	AD.test	VIF_gt_4
0.400	0	3.70e-24	0
0.432	0	3.70e-24	0
0.544	0	3.70e-24	0

Evaluate Binary Regression Models

The third BIC model's cross-validation accuracy and AUC performed has well as second model, but it was more parsimonious with two fewer variables.

Table 11: Model Summary Statistics

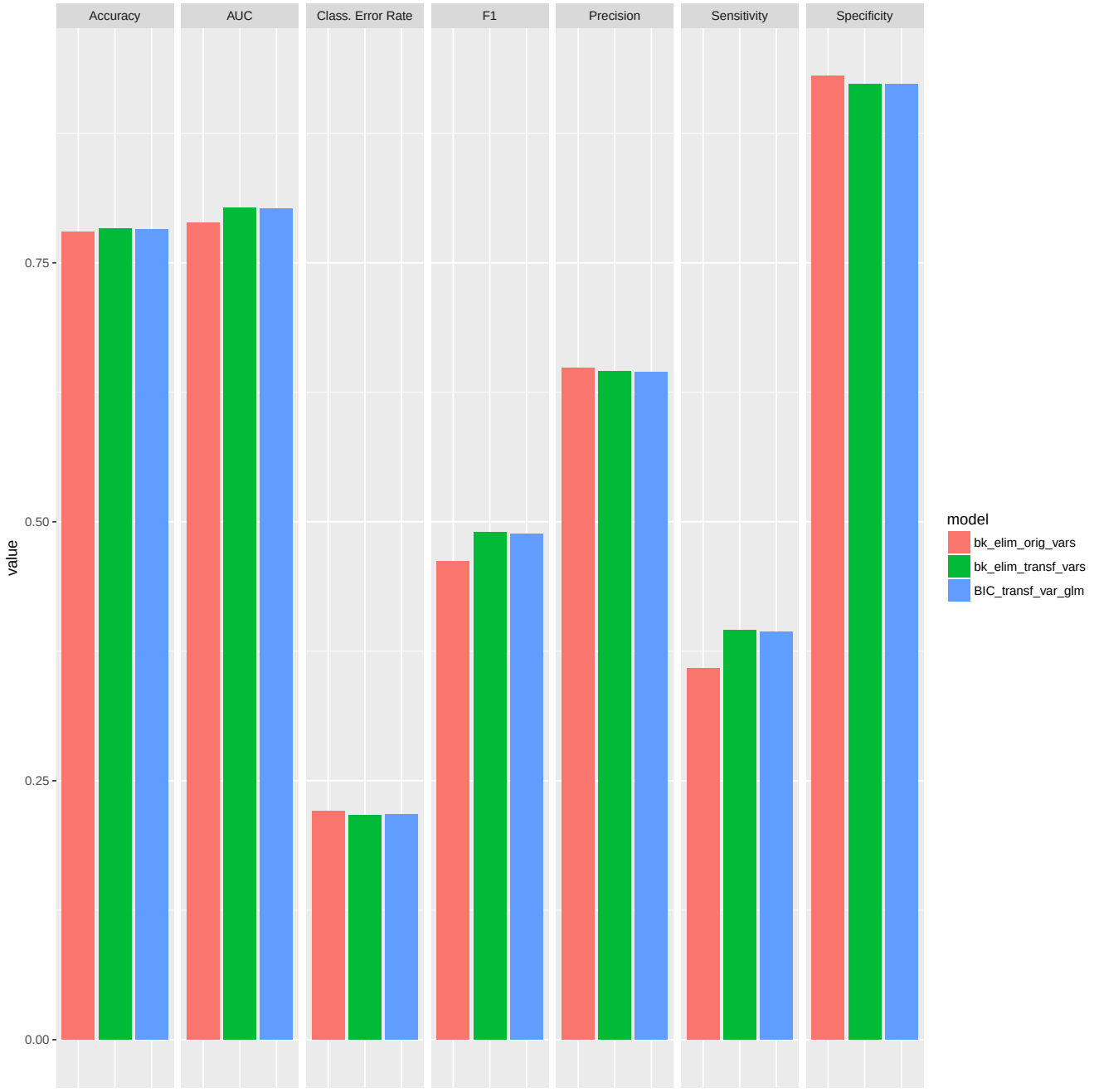
model_name	n_vars	model_pvalue	residual_deviance	H_L_pvalue	VIF_gt_4	CV_accuracy	AUC
bk_elim_orig_vars	16	0.00e+00	7642.470	0.841	0	0.778	0.788
bk_elim_transf_vars	17	0.00e+00	7444.605	0.766	0	0.782	0.803
BIC_transf_var_glm	15	0.00e+00	7453.214	0.693	0	0.782	0.803

Confusion Matrix Metrics

The second and third models performed nearly the same across the spectrum of classification metrics. We would only consider choosing the first model with its highest specificity if the issurance company was most concerned about avoiding false negatives.

The **strengths of the BIC model** include the following:

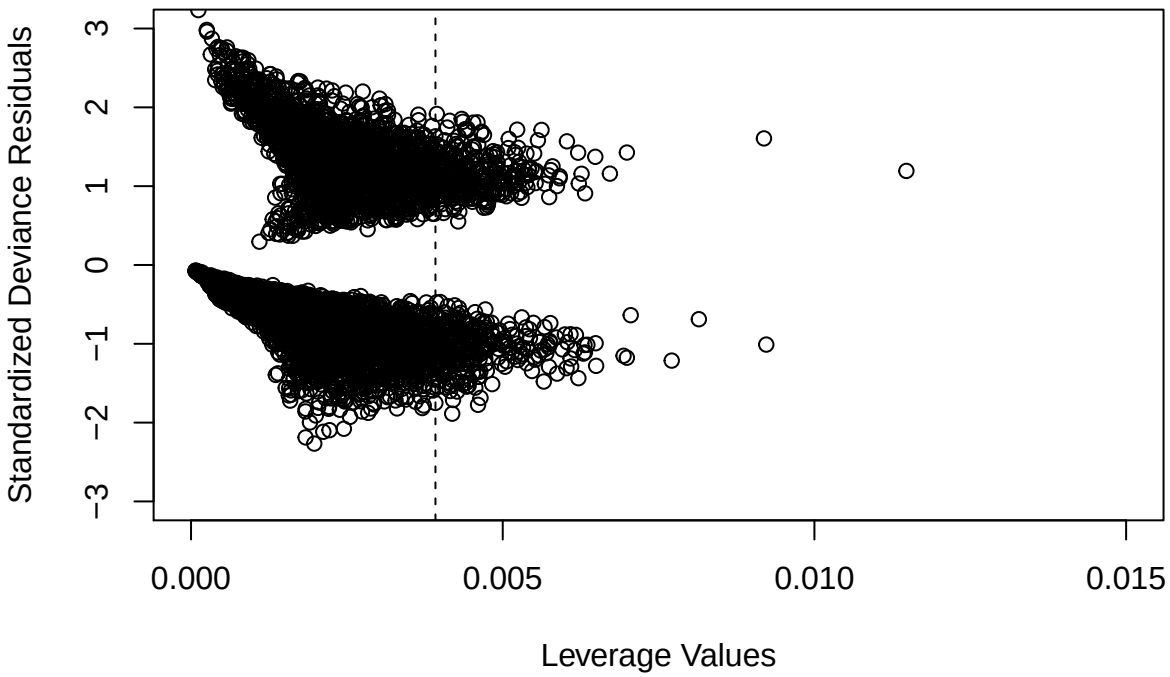
- With only 15 variables, it is the most parsimonious.
- It passes the Hosmer-Lemeshow goodness-of-fit test.
- It doesn't have multicollinearity issues.



Diagnostics

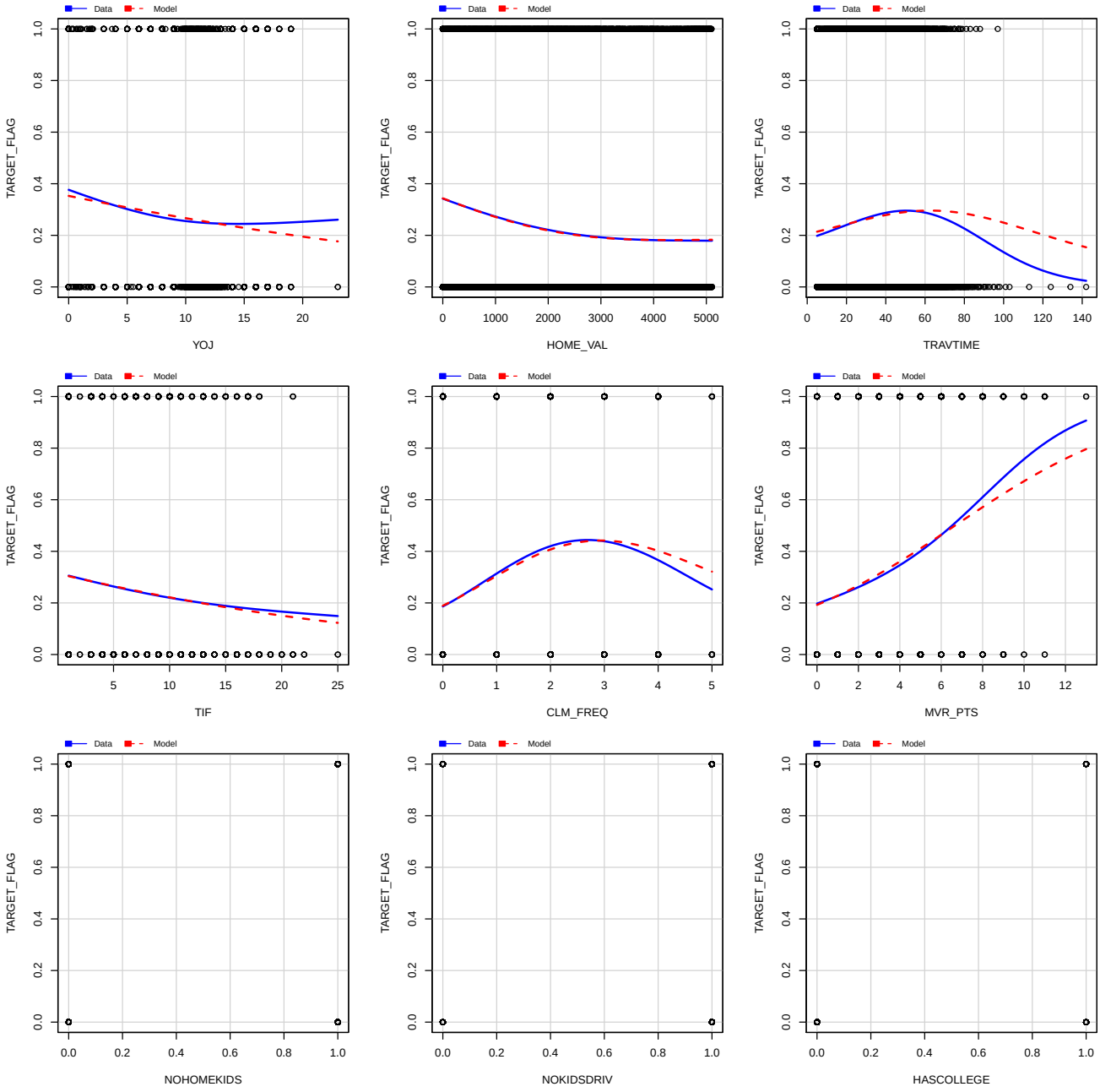
Influence Leverage Values

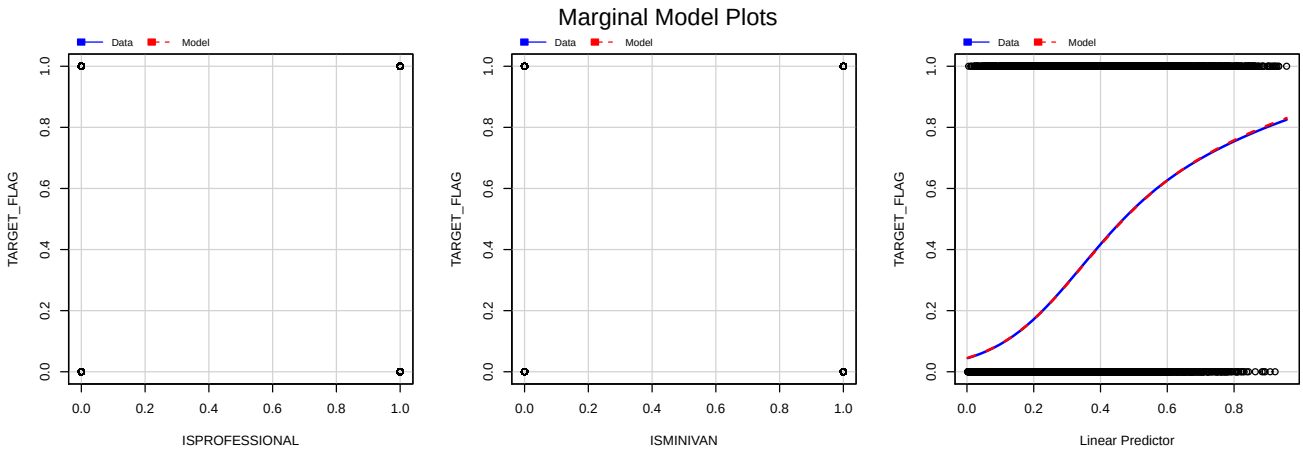
However, a plot of the Standardized Deviance Residuals against the leverage values shows that we have several observations greater than twice the average leverage value. This indicates that there may be other possible variable transformations that we have not considered.



Marginal Model Plots

The marginal model plots of the response variable versus the predictors and the fitted response values show that the model fits reasonably well - although there is some deviation on the right sides of `TRAVTIME` and `MVR_PTS`.


Data
Model
TARGET_FLAG
YOJ
HOME_VAL
TRAVTIME
TIF
CLM_FREQ
MVR_PTS
NOHOMEKIDS
NOKIDSDRIV
HASCOLLEGE



Finally, when we apply the BIC model to the evalution data, it predicts that there are 204 insurance customers that would have an auto accident and 1937 that would not.

```
##
##    0    1
## 1936  205
```

Code Appendix

```
knitr::opts_chunk$set(
                      error = F
                      , message = T
```

```
                      #,tidy = T
                      , cache = T
                      , warning = F
                      , echo = F
                      )
  # prettydoc::html_pretty:
  #   theme: cayman
    # highlight: github
#Install & load packages

load_install <- function(pkg){
  # Load packages. Install them if needed.
  # CODE SOURCE: https://gist.github.com/stevenworthington/3178163
  new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
  if (length(new.pkg)) install.packages(new.pkg, dependencies = TRUE)
  sapply(pkg, require, character.only = TRUE, quietly = TRUE, warn.conflicts = FALSE)
}

# required packages
packages <- c("prettydoc","tidyverse", "caret", "pROC", "DT", "knitr", "ggthemes", "Hmisc", "psych", "corr

#install_version("rmarkdown",version=1.8)

#table(load_install(packages))

data.frame(load_install(packages))

set.seed(5)
# 1. DATA EXPLORATION
train_data <- read.csv("https://raw.githubusercontent.com/kylegilde/D621-Data-Mining/master/HW4%20Auto%20A
  dplyr::select(-INDEX) %>%
  mutate(
    INCOME = as.numeric(INCOME),
    HOME_VAL = as.numeric(HOME_VAL),
    BLUEBOOK = as.numeric(BLUEBOOK),
    OLDCLAIM = as.numeric(OLDCLAIM),
    MSTATUS = as.factor(str_remove(MSTATUS, "^z_")),
    SEX = as.factor(str_remove(SEX, "^z_")),
    EDUCATION = as.factor(str_remove(EDUCATION, "^z_")),
    JOB = as.factor(str_remove(JOB, "^z_")),
    CAR_TYPE = as.factor(str_remove(CAR_TYPE, "^z_")),
    URBANICITY = as.factor(str_remove(URBANICITY, "^z_"))
  )

eval_data <- read.csv("https://raw.githubusercontent.com/kylegilde/D621-Data-Mining/master/HW4%20Auto%20Ac
    dplyr::select(-INDEX) %>%
  mutate(
    INCOME = as.numeric(INCOME),
    HOME_VAL = as.numeric(HOME_VAL),
    BLUEBOOK = as.numeric(BLUEBOOK),
    OLDCLAIM = as.numeric(OLDCLAIM),
    MSTATUS = as.factor(str_remove(MSTATUS, "^z_")),
    SEX = as.factor(str_remove(SEX, "^z_")),
    EDUCATION = as.factor(str_remove(EDUCATION, "^z_")),
    JOB = as.factor(str_remove(JOB, "^z_")),
    CAR_TYPE = as.factor(str_remove(CAR_TYPE, "^z_")),
```

```
    URBANICITY = as.factor(str_remove(URBANICITY, "^z_"))
  )


summary_metrics <- function(df){
  ###Creates summary metrics table
  metrics_only <- df[, sapply(df, is.numeric)]

  df_metrics <- psych::describe(metrics_only, quant = c(.25,.75))
  df_metrics$unique_values = rapply(metrics_only, function(x) length(unique(x)))
  df_metrics <-
    dplyr::select(df_metrics, n, unique_values, min, Q.1st = Q0.25, median, mean, Q.3rd = Q0.75,
    max, range, sd, skew, kurtosis
  )
  return(df_metrics)
}

metrics_df <- summary_metrics(train_data)

# datatable(round(metrics_df, 2), options = list(searching = F, paging = F))

kable(metrics_df, digits = 1, format.args = list(big.mark = ',', scientific = F, drop0trailing = T))

###Categorical & Discrete variables Frequencies
cat_discrete_vars <- train_data %>%
  mutate(
    TARGET_FLAG = as.factor(TARGET_FLAG),
    KIDSDRIV    = as.factor(KIDSDRIV),
    HOMEKIDS    = as.factor(HOMEKIDS),
    CLM_FREQ    = as.factor(CLM_FREQ),
    MVR_PTS = as.factor(MVR_PTS)
  )

cat_discrete_freq <-
  cat_discrete_vars[, sapply(cat_discrete_vars, is.factor)] %>%
  gather("var", "value") %>%
  group_by(var) %>%
  count(var, value) %>%
  mutate(prop = prop.table(n))

ggplot(data = cat_discrete_freq,
       aes(x = reorder(value, prop),
       y = prop)) +
  geom_bar(stat = "identity") +
  facet_wrap(~var, scales = "free") +
  coord_flip() +
  ggthemes::theme_fivethirtyeight()

# https://stackoverflow.com/questions/34860535/how-to-use-dplyr-to-generate-a-frequency-table?utm_medium=o

####Side-by-Side Boxplots
boxplot_data <- cat_discrete_vars %>%
  select_if(is.factor) %>%
  mutate(TARGET_AMT = cat_discrete_vars$TARGET_AMT) %>%
  dplyr::select(-TARGET_FLAG) %>%
  reshape2::melt(id.vars = "TARGET_AMT")
```

```
### Side-by-Side Boxplots
ggplot(data = boxplot_data, aes(x = value, y = TARGET_AMT)) +
  geom_boxplot() +
  facet_wrap( ~ variable, scales = "free") +
  stat_summary(fun.y=mean, geom="point", size=1, color = "red") +
  scale_y_sqrt(breaks = c(1000, 5000, 10000, 20000 * c(1:4))) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))# +

#Reference: https://stackoverflow.com/questions/14604439/plot-multiple-boxplot-in-one-graph?utm_medium=org

### Scatterplots, Histograms & Density Plots
continuous_vars <-
  cat_discrete_vars %>%
  mutate(TARGET_FLAG = as.numeric(TARGET_FLAG) - 1) %>%
  select_if(is.numeric) %>%
  mutate(TARGET_FLAG = as.factor(TARGET_FLAG))

memory.limit(size = 20000)

if (!exists("binary_plot")){
binary_plot <- GGally::ggpairs(
  na.omit(continuous_vars),
  mapping = ggplot2::aes(color = TARGET_FLAG),
  lower = list(continuous = wrap('points', size = .9, alpha = .2),
               combo = wrap('facetdensity', alpha = 1)),
  upper = list(continuous =  wrap("cor", size = 3, alpha = 1),
               combo = 'box'),
  diag =  list(continuous = wrap('barDiag', alpha = .9, bins = 15 ))
 )  +
  theme(panel.background = element_rect(fill = 'grey92', color = NA),
        panel.spacing = unit(3, "pt"),
        panel.grid = element_line(color = 'white'),
        strip.background = element_rect(fill = "grey85", color = NA),
        #plot.margin = margin(.1, .1, .1, .1, "cm"),
        panel.border = element_rect(color = "grey85", fill=NA, size= unit(.5, 'pt')))
}
binary_plot
#http://ggobi.github.io/ggally/#columns_and_mapping

# train_melted <- continuous_vars %>%
#   mutate(TARGET_FLAG = as.integer(TARGET_FLAG) - 1) %>%
#   dplyr::select(-TARGET_AMT) %>%
#   reshape::melt(id.vars = "TARGET_FLAG")
#
#
# na.omit(train_melted) %>%
#   ggplot(aes(x = value, y = TARGET_FLAG)) +
#   geom_point(position = position_jitter(height=.2, width=.2),
#              alpha = .1,
#              aes(color = TARGET_FLAG)
#              ) +
#   geom_smooth(method = "glm", method.args = list(family = "binomial")) +
#   facet_wrap(~ variable, scales = "free")
##CORRELATIONS
```

```r
cormatrix <-
  continuous_vars %>%
  dplyr::select(-TARGET_FLAG) %>%
  cor(use = "complete.obs")

#plot
#corrplot(cormatrix, method = "square", type = "upper")

#find the top correlations
correlations <- c(cormatrix[upper.tri(cormatrix)])

cor_df <-
  data.frame(Var1 = rownames(cormatrix)[row(cormatrix)[upper.tri(cormatrix)]],
             Var2 = colnames(cormatrix)[col(cormatrix)[upper.tri(cormatrix)]],
             Correlation = correlations,
             Rsquared = correlations^2
             ) %>%
  arrange(-Rsquared)

#Reference: https://stackoverflow.com/questions/28035001/transform-correlation-matrix-into-dataframe-with-

kable(head(cor_df, 5), digits = 2, row.names = T, caption = "Top Correlated Variable Pairs")
#Corrrelations with TARGET_AMT
TARGET_AMT_corr <- subset(cor_df, Var2 == "TARGET_AMT" | Var1 == "TARGET_AMT")
rownames(TARGET_AMT_corr) <- 1:nrow(TARGET_AMT_corr)

kable(head(TARGET_AMT_corr, 5), digits = 2, row.names = T, caption = "Top Corrrelations with the Response
## Missing Values
missing_plot <- VIM::aggr(train_data,  numbers=TRUE, sortVars=TRUE,
                      labels=names(train_data),
                      ylab=c("Missing Value Counts","Pattern"))

sum(train_data$JOB == "")

summary(missing_plot)

# 2. DATA PREPARATION
## Imputing the Missing Values
train_data$CAR_AGE[train_data$CAR_AGE == -3] <- NA

eval_data <- dplyr::select(eval_data, -c(TARGET_FLAG, TARGET_AMT))

## Imputing the Missing Values
if (!exists("imputed_train")){
  imputed_train <- missForest(train_data, variablewise = T)
  imputed_eval <- missForest(eval_data, variablewise = T)
}
#impute_results
impute_df <-
  summary(missing_plot)$missings %>%
  mutate(
    MSE = as.numeric(imputed_train$OOBerror),
    Max = sapply(imputed_train$ximp, function(x) tryCatch(max(x), error=function(err) NA)),
    Min = sapply(imputed_train$ximp, function(x) tryCatch(min(x), error=function(err) NA)),
    Range = Max - Min,
    NRMSE = sqrt(MSE)/Range
```

```
  ) %>%
  filter(MSE > 0) %>%
  dplyr::select(-c(Max, Min, Range)) %>%
  arrange(-NRMSE)


kable(impute_df, digits = 2)
#http://rcompanion.org/handbook/G_14.html
#https://stackoverflow.com/questions/14668972/catch-an-error-by-producing-na?utm_medium=organic&utm_source

## Variable Transformations
levels(imputed_train$ximp$JOB)[1] <- "Unknown"

train_transformed <-
  imputed_train$ximp %>%
  mutate(
    NOHOMEKIDS = as.integer(HOMEKIDS == 0),
    NOKIDSDRIV = as.integer(KIDSDRIV == 0),
    HASCOLLEGE = as.integer(EDUCATION %in% c("Bachelors", "Masters", "PhD")),
    ISPROFESSIONAL = as.integer(JOB %in% c("Doctor", "Lawyer", "Manager", "Professional")),
    ISMINIVAN = as.integer(CAR_TYPE == "Minivan")#,
    #sqrt.TARGET_AMT = sqrt(TARGET_AMT)
  ) %>%
  dplyr::select(-c(HOMEKIDS, KIDSDRIV, EDUCATION, JOB, CAR_TYPE))

# 3. BUILD MODELS
## Linear Regression
### LM #1: Original Variables with BIC

orig_var_lm <- lm(TARGET_AMT ~ ., data = dplyr::select(imputed_train$ximp, -TARGET_FLAG))

n <- nrow(orig_var_lm$model)
BIC_orig_var_lm <- step(orig_var_lm, k = log(n), trace = 0)

summary(BIC_orig_var_lm)

removed_variables <- function(larger_mod, smaller_mod){
  removed <- names(coef(larger_mod))[!names(coef(larger_mod)) %in%
names(coef(smaller_mod))]
    print(paste("removed variable(s):", length(removed)))
    print(removed)
}

removed_variables(orig_var_lm, BIC_orig_var_lm)


PRESS <- function(linear.model) {
  #source:  https://gist.github.com/tomhopper/8c204d978c4a0cbcb8c0#file-press-r
  #' calculate the predictive residuals
  pr <- residuals(linear.model)/(1 - lm.influence(linear.model)$hat)
  #' calculate the PRESS
  PRESS <- sum(pr^2)
  return(PRESS)
}
pred_r_squared <- function(linear.model) {
  #source: https://gist.github.com/tomhopper/8c204d978c4a0cbcb8c0#file-pred_r_squared-r
```

```
  #' Use anova() to get the sum of squares for the linear model
  lm.anova <- anova(linear.model)
  #' Calculate the total sum of squares
  tss <- sum(lm.anova$'Sum Sq')
  # Calculate the predictive R^2
  pred.r.squared <- 1 - PRESS(linear.model)/(tss)
  return(pred.r.squared)
}

lm_evaluation <- function(lmod) {
    ### Summarizes the model's key statistics in one row
    ### https://gist.github.com/stephenturner/722049#file-pvalue-from-lm-object-r-L5
    lm_summary <- summary(lmod)
    f <- as.numeric(lm_summary$fstatistic)

    df_summary <-
      data.frame(
        model_name = deparse(substitute(lmod)),
        n_predictors = ncol(lmod$model) - 1,
        numdf = f[2],
        fstat = f[1],
        p.value = formatC(pf(f[1], f[2], f[3], lower.tail = F), format = "e", digits = 2),
        adj.r.squared = lm_summary$adj.r.squared,
        pre.r.squared = pred_r_squared(lmod)#,
        #rmse = as.numeric(DMwR::regr.eval(lmod$model[1], fitted(lmod), stats = c("rmse")))

        )
    return(df_summary)
}

lm_diagnotics <- function(lmod){
  diag_df <- data.frame(
    DW.test = car::durbinWatsonTest(lmod)$p,
    NCV.test = car::ncvTest(lmod)$p,
    AD.test = formatC(nortest::ad.test(lmod$residuals)$p.value, format = "e", digits = 2),
    VIF_gt_4 = sum(car::vif(lmod) > 4)
  )
  return(diag_df)
}

#evaluate performance & diagnostics
lm_results <- lm_evaluation(BIC_orig_var_lm)
lm_results_diagnostics <- lm_diagnotics(BIC_orig_var_lm)


par(mfrow=c(2,2))
plot(BIC_orig_var_lm)
#http://data.library.virginia.edu/diagnostic-plots/
#http://analyticspro.org/2016/03/07/r-tutorial-how-to-use-diagnostic-plots-for-regression-models/
### LM #2: Transformed Variables with BIC Selection
transf_var_lmod <- lm(TARGET_AMT ~ .-TARGET_FLAG, data = train_transformed)

n <- nrow(transf_var_lmod$model)

BIC_transf_var_lm <- step(transf_var_lmod, k = log(n), trace = 0)
```

```r
summary(BIC_transf_var_lm)


#evaluate performance & diagnostics
model_eval <- lm_evaluation(BIC_transf_var_lm)
model_diag <- lm_diagnotics(BIC_transf_var_lm)
lm_results <- rbind(lm_results, model_eval)
lm_results_diagnostics <- rbind(lm_results_diagnostics, model_diag)
par(mfrow=c(2,2))
plot(BIC_transf_var_lm)

### LM #3: BIC Selection with Negative Box-Cox Transformation & Transformed Variables
PT <- car::powerTransform(BIC_transf_var_lm, family = "bcnPower")

#PT <- car::powerTransform(BIC_transf_var_lm, family = "yjPower")

summary(PT)
#load_install("VGAM")

train_transformed_BCN <-
  train_transformed %>%
  mutate(
    #TARGET_AMT = car::bcnPower(TARGET_AMT, lambda = PT$lambda, gamma = PT$gamma)
    #TARGET_AMT = VGAM::yeo.johnson(TARGET_AMT, PT$lambda)
    TARGET_AMT = log(TARGET_AMT + 1)
    #TARGET_AMT = (TARGET_AMT+PT$gamma)^PT$lambda
    #TARGET_AMT = TARGET_AMT^PT$lambda
  )

# Our third model will use the 2-parameter Box-Cox transformation of `TARGET_AMT` with the transformed var

BCN_lmod <- lm(TARGET_AMT~.-TARGET_FLAG, data = train_transformed_BCN)
BIC_BCN_transf_var_lm <- step(BCN_lmod, k = log(n), trace = 0)
summary(BIC_BCN_transf_var_lm)

model_eval <- lm_evaluation(BIC_BCN_transf_var_lm)
model_diag <- lm_diagnotics(BIC_BCN_transf_var_lm)
lm_results <- rbind(lm_results, model_eval)
lm_results_diagnostics <- rbind(lm_results_diagnostics, model_diag)

# Interpreting the coefficients after performing 2-parameter Box-Cox transformation can be challenging. Af
# Inverse the transformed coefficients
# BCN_inverse <- function(U, lambda, gamma){
#   #Calculates the inverse of the negative Box-Cox transformation
#   return((U * lambda + 1)^(1/lambda) - gamma)
#   #return((U)^(1/lambda) - gamma)
#   #return((U)^(1/lambda))
# }
#https://stats.stackexchange.com/questions/233611/reverse-boxcox-transformation-with-negative-values?utm_m
#https://stackoverflow.com/questions/18464491/transforming-data-to-normality-what-is-the-best-function-for

# inversed_coef_df <-
#   data.frame(var = names(coef(BIC_BCN_transf_var_lm))) %>%
#   mutate(
#     transformed_coef_value = coef(BIC_BCN_transf_var_lm),
#     inversed_coef_value = BCN_inverse(transformed_coef_value, PT$lambda, PT$gamma)
```

```
#   ) %>%
#   arrange(-transformed_coef_value)

inversed_coef_df <-
  data.frame(var = names(coef(BIC_BCN_transf_var_lm))) %>%
  mutate(
    transformed_coef_value = coef(BIC_BCN_transf_var_lm),
    inversed_coef_value = exp(transformed_coef_value) - 1
    #inversed_coef_value = VGAM::yeo.johnson(transformed_coef_value, PT$lambda, inverse = T)
  ) %>%
  arrange(-transformed_coef_value)

kable(inversed_coef_df, digits = 2, format.args = list(big.mark = ',', scientific = F, drop0trailing = T))

par(mfrow=c(2,2))
plot(BIC_BCN_transf_var_lm)

backward_elimination <- function(lmod){
  #performs backward elimination model selection
  #removes variables until all remaining ones are stat-sig
  removed_vars <- c()
  removed_pvalues <- c()

  #handles category dummy variables
  cat_levels <- unlist(lmod$xlevels)
  cat_vars <- str_sub(names(cat_levels), 1, nchar(names(cat_levels)) - 1)
  cat_var_df <- data.frame(cat_vars,
                           dummy_vars = str_c(cat_vars, cat_levels),
                           stringsAsFactors = F)
  # checks for p-values > .05 execpt for the intercept
  while (max(summary(lmod)$coefficients[2:length(summary(lmod)$coefficients[, 4]), 4]) > .05){
    # find insignificant pvalue
    pvalues <- summary(lmod)$coefficients[2:length(summary(lmod)$coefficients[, 4]), 4]
    max_pvalue <- max(pvalues)
    remove <- names(which.max(pvalues))
    #if categorical dummy variable, remove the variable
    dummy_var <- dplyr::filter(cat_var_df, dummy_vars == remove)
    remove <- ifelse(nrow(dummy_var) > 0, dummy_var[, 1], remove)
    #record the removed variables
    removed_vars <- c(removed_vars, remove)
    removed_pvalues <- c(removed_pvalues, max_pvalue)
    # update model
    lmod <- update(lmod, as.formula(paste0(".~.-`", remove, "`")))
  }

  print(kable(data.frame(removed_vars, removed_pvalues), digits = 3))
  return(lmod)
}

orig_var_glm <- glm(TARGET_FLAG ~ . , family = "binomial", data = dplyr::select(imputed_train$ximp, -TARGE

summary(bk_elim_orig_vars <- backward_elimination(orig_var_glm))


glm_performance <- function(model) {
  ### Summarizes the model's key statistics
```

```
  ### References: https://www.r-bloggers.com/predicting-creditability-using-logistic-regression-in-r-cross
  ### https://www.rdocumentation.org/packages/boot/versions/1.3-20/topics/cv.glm
  ### https://www.rdocumentation.org/packages/ResourceSelection/versions/0.3-1/topics/hoslem.test
  cost <- function(r, pi = 0) mean(abs(r - pi) > 0.5)

  df_summary <- data.frame(
   # model = bk_elim_orig_vars
    model_name = deparse(substitute(model)),
    n_vars = length(coef(model)) - 1,
    model_pvalue = formatC(pchisq(model$null.deviance - model$deviance, 1, lower=FALSE), format = "e", dig
    residual_deviance = model$deviance,
    H_L_pvalue = ResourceSelection::hoslem.test(model$y, fitted(model))$p.value,
    VIF_gt_4 = sum(car::vif(model) > 4),
    CV_accuracy = 1 - boot::cv.glm(model$model, model, cost = cost, K = 100)$delta[1],
    AUC = as.numeric(pROC::roc(model$y, fitted(model))$auc)
  )
  return(df_summary)
}



kable(glm_models <- glm_performance(bk_elim_orig_vars), digits = 3)

bk_elim_orig_vars_roc <- roc(bk_elim_orig_vars$y, fitted(bk_elim_orig_vars))
plot(bk_elim_orig_vars_roc, legacy.axes = T, main = "Model ROCs", col = "gray", xaxs = "i", yaxs = "i")

#https://web.archive.org/web/20160407221300/http://metaoptimize.com:80/qa/questions/988/simple-explanation
transf_var_glm <- glm(TARGET_FLAG ~ . , family = "binomial", data = dplyr::select(train_transformed, -TARG

summary(bk_elim_transf_vars <- backward_elimination(transf_var_glm))

glm_mod <- glm_performance(bk_elim_transf_vars)
kable(glm_models <- rbind(glm_models, glm_mod), digits = 3)


bk_elim_transf_vars_roc <- roc(bk_elim_transf_vars$y, fitted(bk_elim_transf_vars))
plot(bk_elim_orig_vars_roc, legacy.axes = T, main = "Model ROCs", col = "gray", xaxs = "i", yaxs = "i")
plot(bk_elim_transf_vars_roc, add = T, col = "blue", lty = 3)

### Binary Regression Model #3: Transformed Variables with BIC
BIC_transf_var_glm <- step(transf_var_glm, k = log(n), trace = 0)
summary(BIC_transf_var_glm)

glm_mod <- glm_performance(BIC_transf_var_glm)
kable(glm_models <- rbind(glm_models, glm_mod), digits = 3)

BIC_transf_var_glm_roc <- roc(BIC_transf_var_glm$y, fitted(BIC_transf_var_glm))

plot(bk_elim_orig_vars_roc, legacy.axes = T, main = "Model ROCs", col = "gray", xaxs = "i", yaxs = "i")
plot(bk_elim_transf_vars_roc, add = T, col = "blue", lty = 3)
plot(BIC_transf_var_glm_roc, add = T, col = "orange", lty = 3)
kable(lm_results, digits = 3, caption = "Model Summary Statistics")


kable(lm_results_diagnostics, digits = 3, caption = "Model Summary Statistics")
kable(glm_models, digits = 3, caption = "Model Summary Statistics")
```

```r
# 4. SELECT MODELS
all_models <- as.character(glm_models$model_name)

confusion_metrics <- data.frame(metric =  c("Accuracy", "Class. Error Rate", "Sensitivity", "Specificity",

for (i in 1:length(all_models)){
  model <- get(all_models[i])
  model_name <- all_models[i]
  predicted_values <- as.factor(as.integer(fitted(model) > .5))
  CM <- confusionMatrix(predicted_values, as.factor(model$y), positive = "1")
  caret_metrics <- c(CM$overall[1],
                     1 - as.numeric(CM$overall[1]),
                     CM$byClass[c(1, 2, 5, 7)],
                     get(paste0(model_name, "_roc"))$auc)
  confusion_metrics[, model_name] <- caret_metrics
}

confusion_metrics_melted <- confusion_metrics %>%
  reshape::melt(id.vars = "metric") %>%
  dplyr::rename(model = variable)

ggplot(data = confusion_metrics_melted, aes(x = model, y = value)) +
  geom_bar(aes(fill = model), stat='identity') +
  theme(axis.ticks.x=element_blank(),
        axis.text.x=element_blank(),
        axis.title.x=element_blank()) +
  facet_grid(~metric)

#https://stackoverflow.com/questions/18624394/ggplot-bar-plot-with-facet-dependent-order-of-categories/186
#https://stackoverflow.com/questions/18158461/grouped-bar-plot-in-ggplot?utm_medium=organic&utm_source=goo

#kable(confusion_metrics, digits = 3)

# influential leverage values
# MARR p291

hvalues <- influence(BIC_transf_var_glm)$hat
stanresDeviance <- residuals(BIC_transf_var_glm) / sqrt(1 - hvalues)
n_predictors <- length(names(BIC_transf_var_glm$model)) - 1
average_leverage <- (n_predictors + 1) / nrow(BIC_transf_var_glm$model)
plot(hvalues, stanresDeviance,
     ylab = "Standardized Deviance Residuals",
     xlab = "Leverage Values",
     ylim = c(-3, 3),
     xlim = c(0, 0.015))
abline(v = 2 * average_leverage, lty = 2)

#Reference: http://www.stat.tamu.edu/~sheather/book/docs/rcode/Chapter8.R

if(!exists("marginal_model_plots")){marginal_model_plots <- car::mmps(BIC_transf_var_glm)}

#http://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_templt_sect027.ht

levels(imputed_eval$ximp$JOB)[1] <- "Unknown"
```

```r
eval_transformed <-
  imputed_eval$ximp %>%
  mutate(
    NOHOMEKIDS = as.integer(HOMEKIDS == 0),
    NOKIDSDRIV = as.integer(KIDSDRIV == 0),
    HASCOLLEGE = as.integer(EDUCATION %in% c("Bachelors", "Masters", "PhD")),
    ISPROFESSIONAL = as.integer(JOB %in% c("Doctor", "Lawyer", "Manager", "Professional")),
    ISMINIVAN = as.integer(CAR_TYPE == "Minivan")#,
    #sqrt.TARGET_AMT = sqrt(TARGET_AMT)
  ) %>%
  dplyr::select(-c(HOMEKIDS, KIDSDRIV, EDUCATION, JOB, CAR_TYPE))


eval_results <- predict(BIC_transf_var_glm, newdata = eval_transformed)

table(as.integer(eval_results > .5))
```