# Predicting the Occurrence and Cost of Car Accidents

Biguzzi, Connin, Greenlee, Moscoe, Sooklall, Telab, and Wright

11/21/2021

## Introduction

This analysis predicts the probability that a given insurance customer will be involved in a car crash, as well as the cost of the crash to the insurance company. We begin with an exploration of the data to build an initial impression on the relationships, which will guide how we transform and select variables. We construct two models: a logistic regression for the binary target variable of Crash vs No Crash; and a least-squares model for the target dollar cost variable. Ultimately, we will integrate both results to help the insurer evaluate their financial risk.

In this report we will:

- Explore the data
- Transform data to address multicollinearity and meet variable distribution needs
- Compare different models and select the most accurate model
- Test our model on the evaluation data set

## Exploratory Data Analysis and Wrangling

As a first step in our EDA process, we applied a basic set of data cleaning operations at the onset. These included:

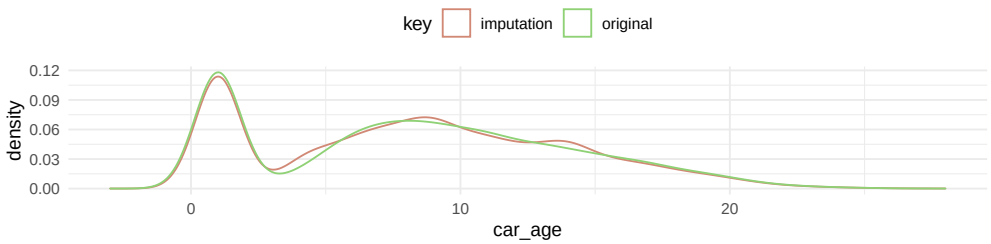
1. standardizing column names to remove special characters and convert to lower-case construction.
2. removing any empty rows or columns
3. updating column types where appropriate
4. removing special characters and symbols from column values
5. simplifying select character values for brevity and clarity
6. converting character column types to factor
7. creating factor levels for the 'education' variable

We then evaluated our data for missing values. Our assessment distinguished 5 numerical variables and 1 categorical variable with missing entries. The proportion of missingness across these variables was relatively low, ranging from ~ .07 to 6.5 percent.

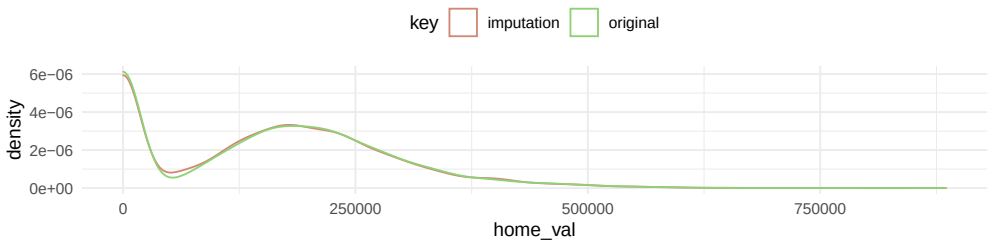
| variables | types | missing_count | missing_percent |
|---|---|---|---|
| job | factor | 526 | 6.4452886 |
| car_age | numeric | 510 | 6.2492342 |
| home_val | numeric | 464 | 5.6855777 |
| yoj | numeric | 454 | 5.5630437 |
| income | numeric | 445 | 5.4527631 |
| age | numeric | 6 | 0.0735204 |

We then imputed values for these entries using a combination of recursive partitioning and chained equations. These methods yielded superior results relative to other imputation procedures that we tested (e.g., mode, mean, median, knn). This view is supported by a comparison of the distribution densities for the original data and data with imputed values included for each variable with missing entries (see below).

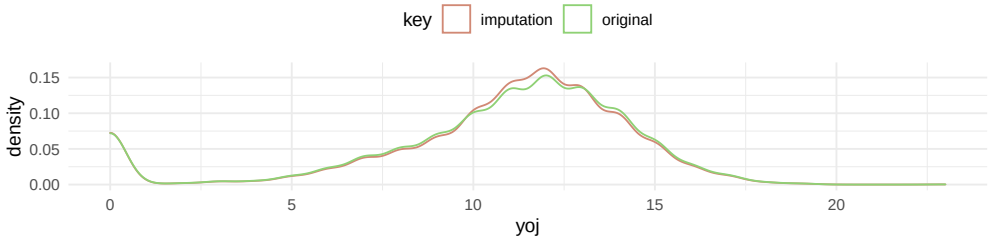
imputation method : rpart



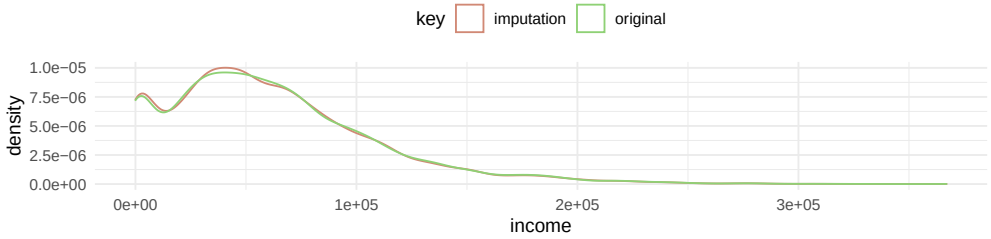

imputation method : rpart



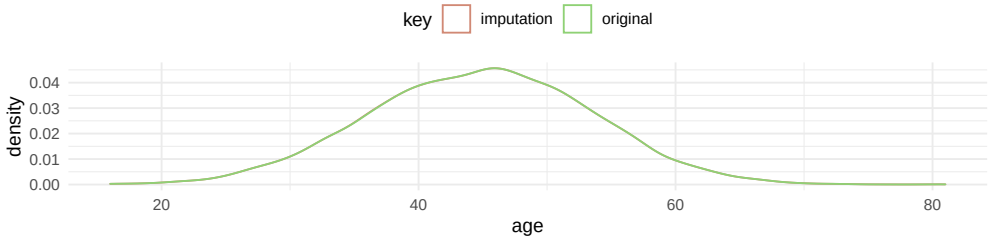

imputation method : rpart



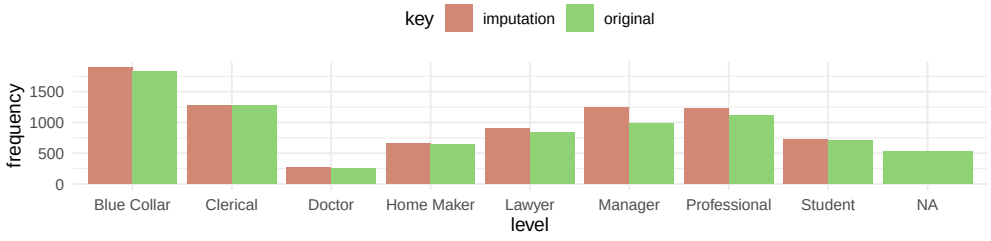

imputation method : rpart




imputation method : rpart




imputation method : mice (seed = 999)

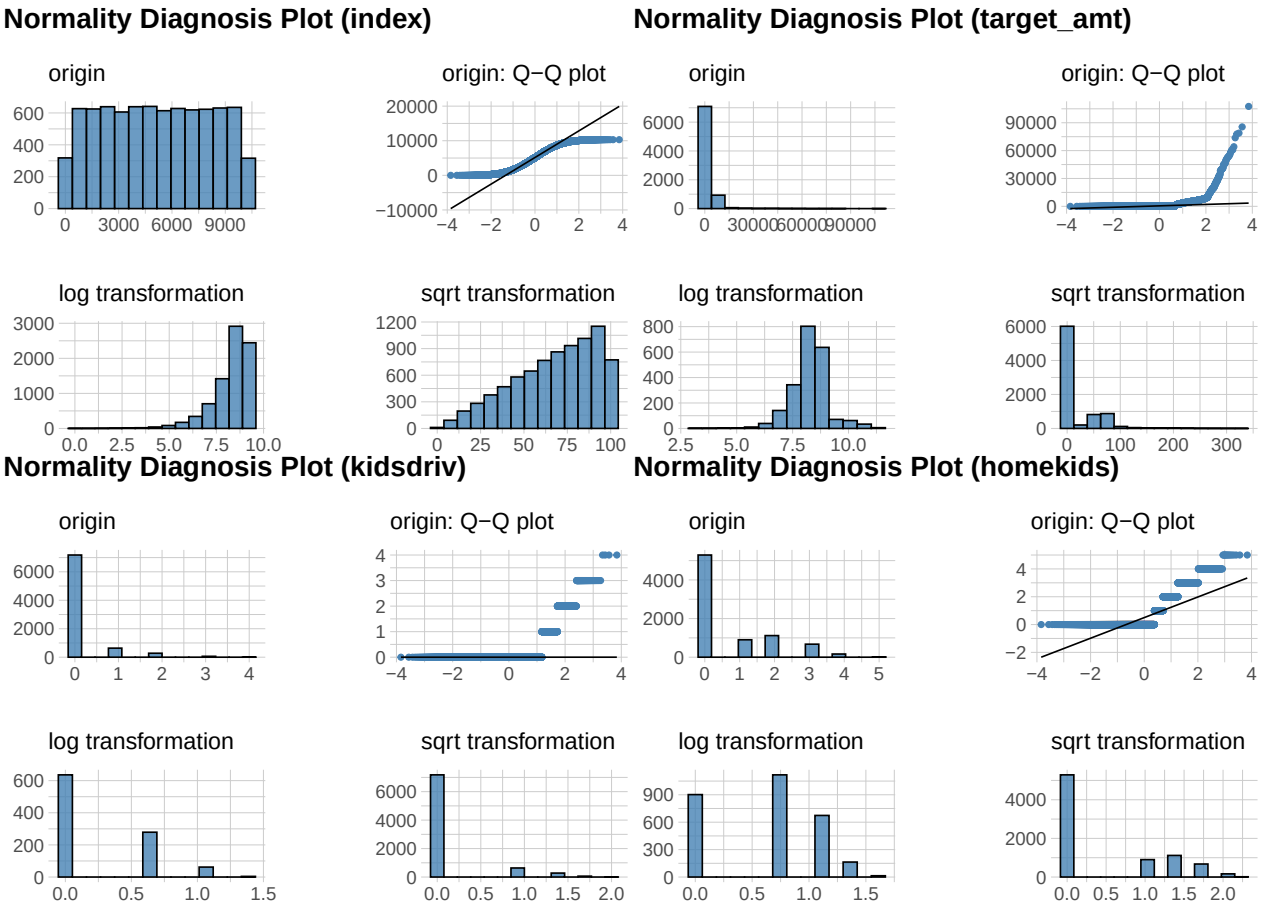


With a complete data set, we then evaluated our variable distributions relative to a normal model. The histograms and QQ-plots below highlight these distributions and provide alternative outcomes based on transformations (log or square-root) of the data for each variable. We can discern the following:

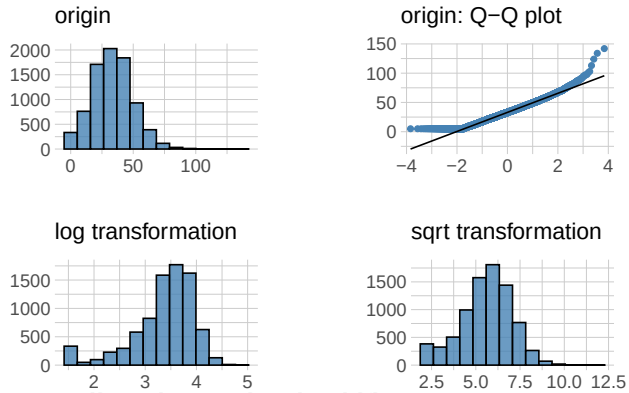
1. a number of variables are highly right skewed owing to zero inflation: target_amt, kidsdrive, homekids, tif, oldclaim, clm_freq, mvr_pts, car_age, home_val
2. others are right skewed without zero inflation: bluebook, income
3. several numerical variables are also bimodal: car_age, home_val, yoj

A Shapiro-Wilk can also be employed to identify deviations from a normal model. These statistics are included in the following table. We note that a significance level below 0.05 indicates that the data do not fit a normal distribution.

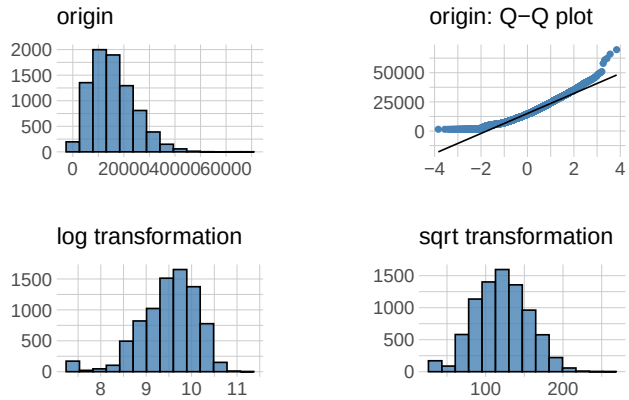
These assessments provide a basis for identifying potential variable transformations during the modeling stage to improve model fit. And we note that high zero counts may point to contamination in the data that isn't implicit in the variable desriptions. For example, a zero value in home-value may be related to external factors such as renting, under-age driver claims, etc.



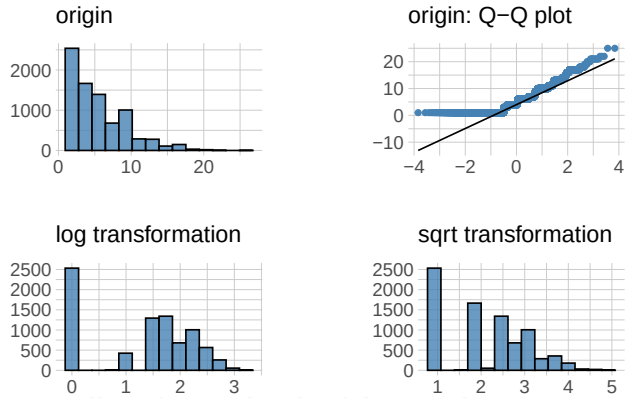
**Normality Diagnosis Plot (travtime)**



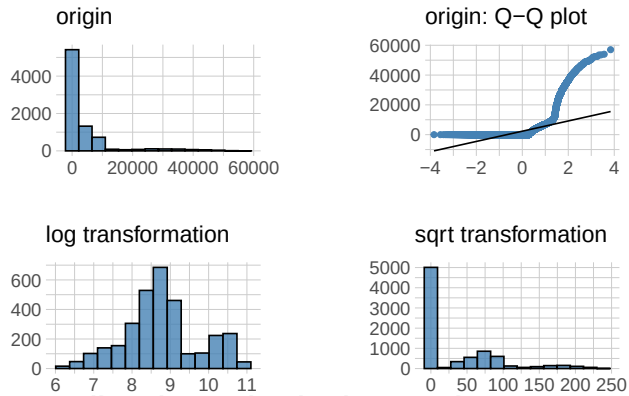

**Normality Diagnosis Plot (bluebook)**



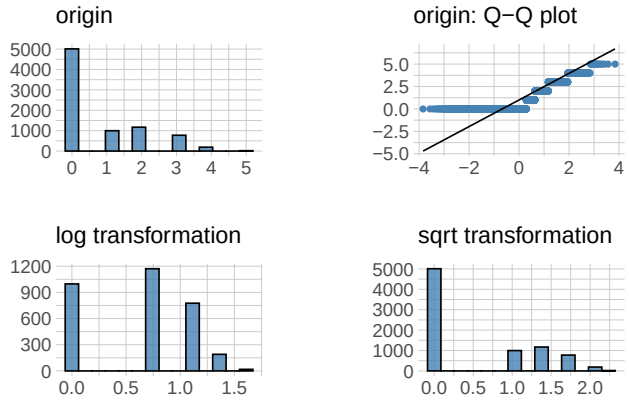

**Normality Diagnosis Plot (tif)**



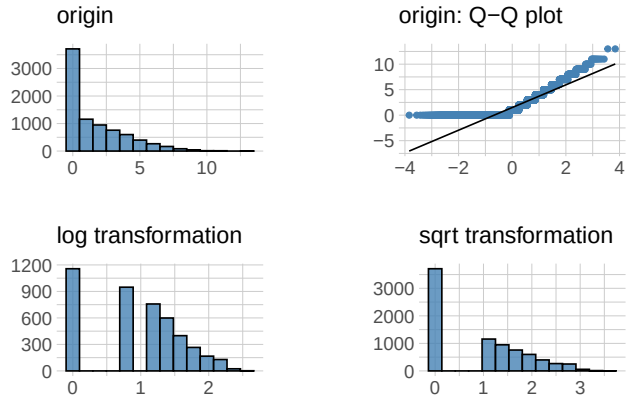

**Normality Diagnosis Plot (oldclaim)**

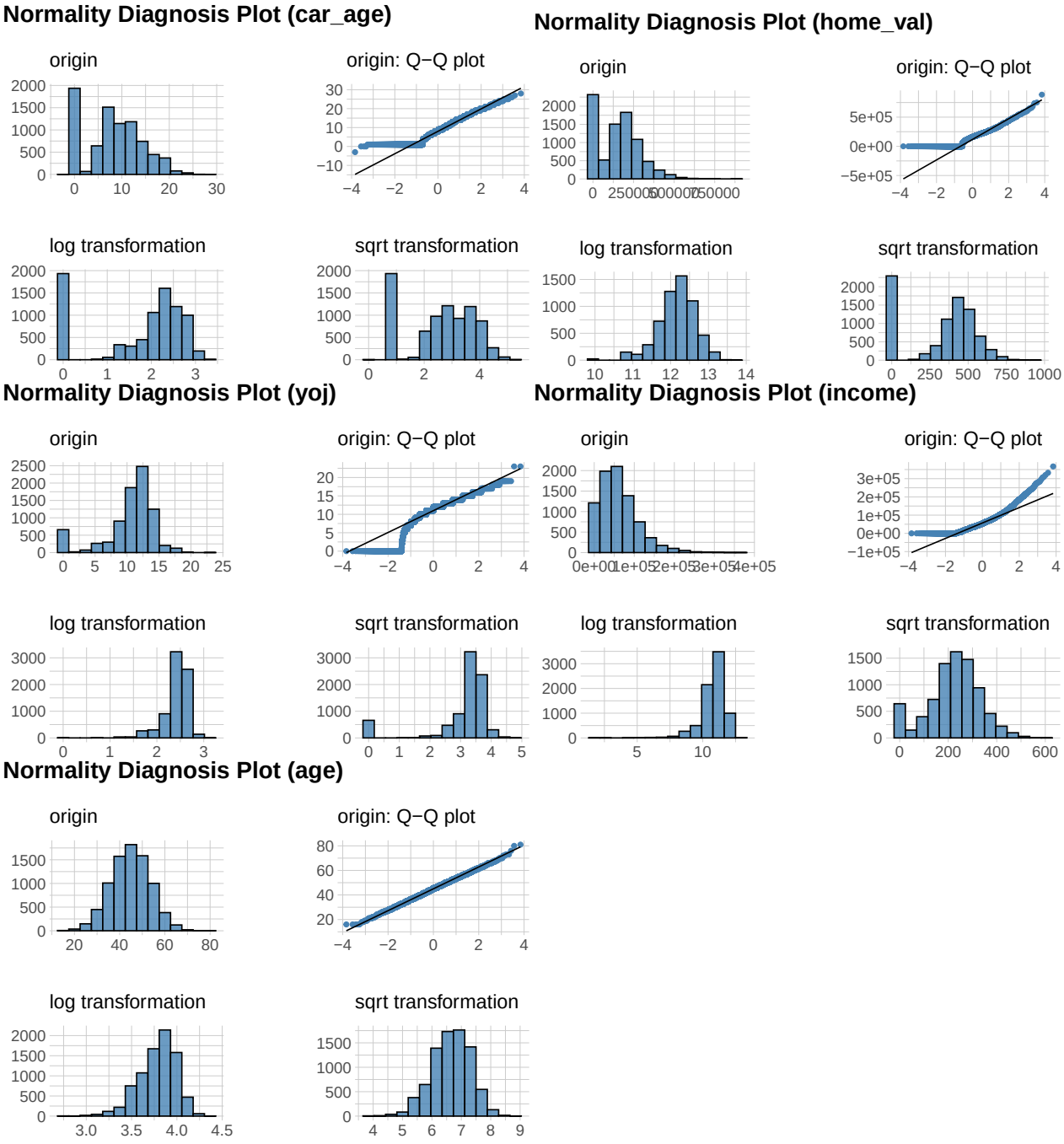



**Normality Diagnosis Plot (clm_freq)**




**Normality Diagnosis Plot (mvr_pts)**






Other diagnostics can be employed to evaluate data structure and ensure data quality.

The following table provides and overview of the levels of each categorical variable with regard to count, frequency, percentage, and rank. We can draw on this information to guide our modeling descisions downstream.

| variables | levels | N | freq | ratio | rank |
|---|---|---|---|---|---|
| target_flag | 0 | 8,161 | 6,008 | 73.618429 | 1 |
| target_flag | 1 | 8,161 | 2,153 | 26.381571 | 2 |
| parent1 | N | 8,161 | 7,084 | 86.803088 | 1 |

| variables | levels | N | freq | ratio | rank |
|---|---|---|---|---|---|
| parent1 | Y | 8,161 | 1,077 | 13.196912 | 2 |
| mstatus | Y | 8,161 | 4,894 | 59.968141 | 1 |
| mstatus | N | 8,161 | 3,267 | 40.031859 | 2 |
| sex | F | 8,161 | 4,375 | 53.608626 | 1 |
| sex | M | 8,161 | 3,786 | 46.391374 | 2 |
| education | High School | 8,161 | 3,533 | 43.291263 | 1 |
| education | Bachelors | 8,161 | 2,242 | 27.472124 | 2 |
| education | Masters | 8,161 | 1,658 | 20.316138 | 3 |
| education | PhD | 8,161 | 728 | 8.920475 | 4 |
| car_use | Private | 8,161 | 5,132 | 62.884450 | 1 |
| car_use | Commercial | 8,161 | 3,029 | 37.115550 | 2 |
| car_type | SUV | 8,161 | 2,294 | 28.109300 | 1 |
| car_type | Minivan | 8,161 | 2,145 | 26.283544 | 2 |
| car_type | Pickup | 8,161 | 1,389 | 17.019973 | 3 |
| car_type | Sports Car | 8,161 | 907 | 11.113834 | 4 |
| car_type | Van | 8,161 | 750 | 9.190050 | 5 |
| car_type | Panel Truck | 8,161 | 676 | 8.283299 | 6 |
| red_car | N | 8,161 | 5,783 | 70.861414 | 1 |
| red_car | Y | 8,161 | 2,378 | 29.138586 | 2 |
| revoked | N | 8,161 | 7,161 | 87.746600 | 1 |
| revoked | Y | 8,161 | 1,000 | 12.253400 | 2 |
| urbanicity | Urban | 8,161 | 6,492 | 79.549075 | 1 |
| urbanicity | Rural | 8,161 | 1,669 | 20.450925 | 2 |
| job | Blue Collar | 8,161 | 1,890 | 23.158927 | 1 |
| job | Clerical | 8,161 | 1,276 | 15.635339 | 2 |
| job | Manager | 8,161 | 1,236 | 15.145203 | 3 |
| job | Professional | 8,161 | 1,222 | 14.973655 | 4 |
| job | Lawyer | 8,161 | 895 | 10.966793 | 5 |
| job | Student | 8,161 | 717 | 8.785688 | 6 |
| job | Home Maker | 8,161 | 657 | 8.050484 | 7 |
| job | Doctor | 8,161 | 268 | 3.283911 | 8 |

We can view our numerical data in relation to measures of spread and central tendency as well as zero count and negative values. The latter can be used to identify erroneous entries for data that should be non-negative.

| variables | min | mean | median | max | zero | minus |
|---|---|---|---|---|---|---|
| index | 1 | 5,151.8676633 | 5,133 | 10,302 | 0 | 0 |
| target_amt | 0 | 1,504.3248376 | 0 | 107,586 | 6,008 | 0 |
| kidsdriv | 0 | 0.1710575 | 0 | 4 | 7,180 | 0 |
| homekids | 0 | 0.7212351 | 0 | 5 | 5,289 | 0 |
| travtime | 5 | 33.4857248 | 33 | 142 | 0 | 0 |
| bluebook | 1,500 | 15,709.8995221 | 14,440 | 69,740 | 0 | 0 |
| tif | 1 | 5.3513050 | 4 | 25 | 0 | 0 |
| oldclaim | 0 | 4,037.0762161 | 0 | 57,037 | 5,009 | 0 |
| clm_freq | 0 | 0.7985541 | 0 | 5 | 5,009 | 0 |
| mvr_pts | 0 | 1.6955030 | 1 | 13 | 3,712 | 0 |
| car_age | 0 | 8.3453431 | 8 | 28 | 3 | 0 |
| home_val | 0 | 154,903.4969979 | 160,333 | 885,282 | 2,294 | 0 |
| yoj | 0 | 10.5169710 | 11 | 23 | 659 | 0 |
| income | 0 | 61,501.3976228 | 53,156 | 367,030 | 615 | 0 |
| age | 16 | 44.7850754 | 45 | 81 | 0 | 0 |

To complement our review of variable distributions we can also quantify the overall count and proportion of outliers for data within each variable. The presence and overall structure (e.g., clumping) of outliers can point unique circumstances that lend insight the observations. They also provide a basis for identifying data points that may exert high leverage and influence with regard to model fit.

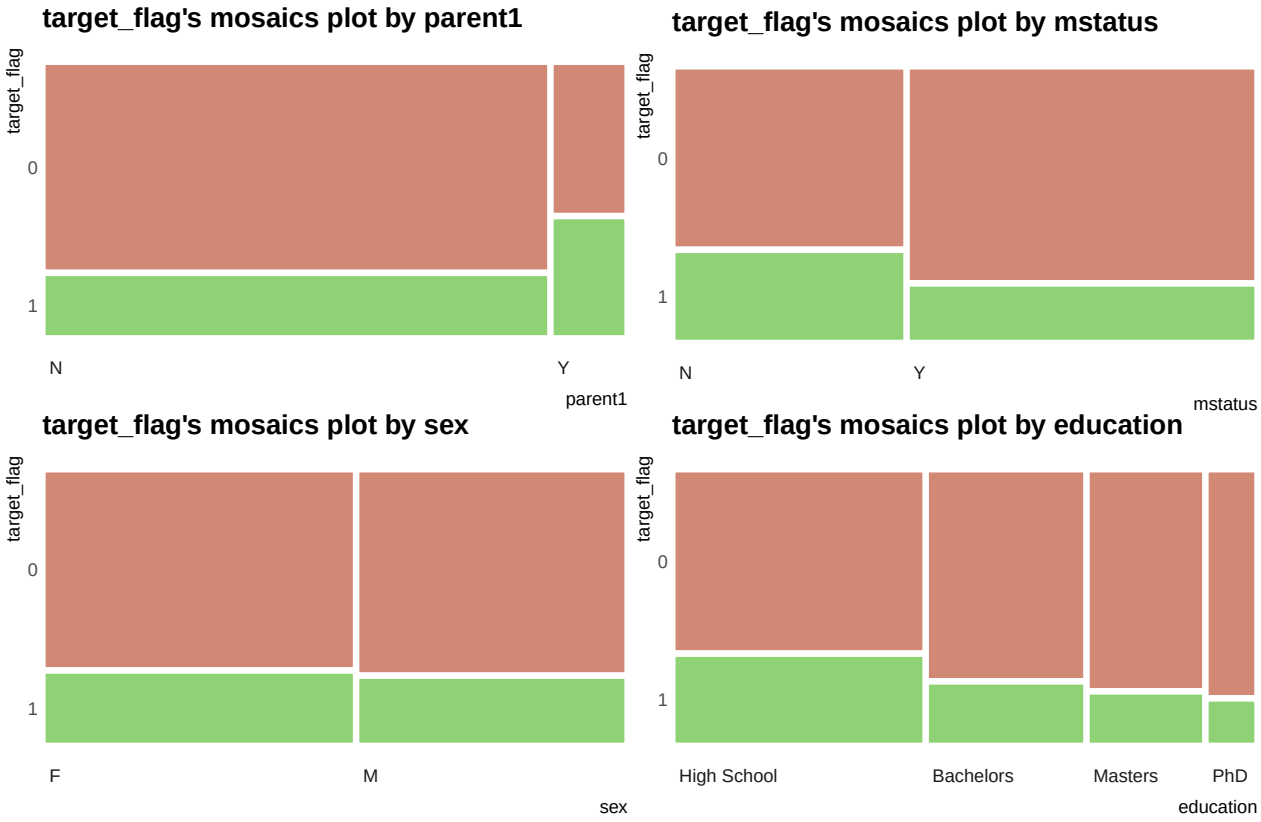
The following table lists response variable by decreasesing outlier count and related measures. It's clear that outliers comprise more than 10 percent of the data for homekids, kidsdrive, and target_amt. This is also consistent with the high count of zero values for these variables.

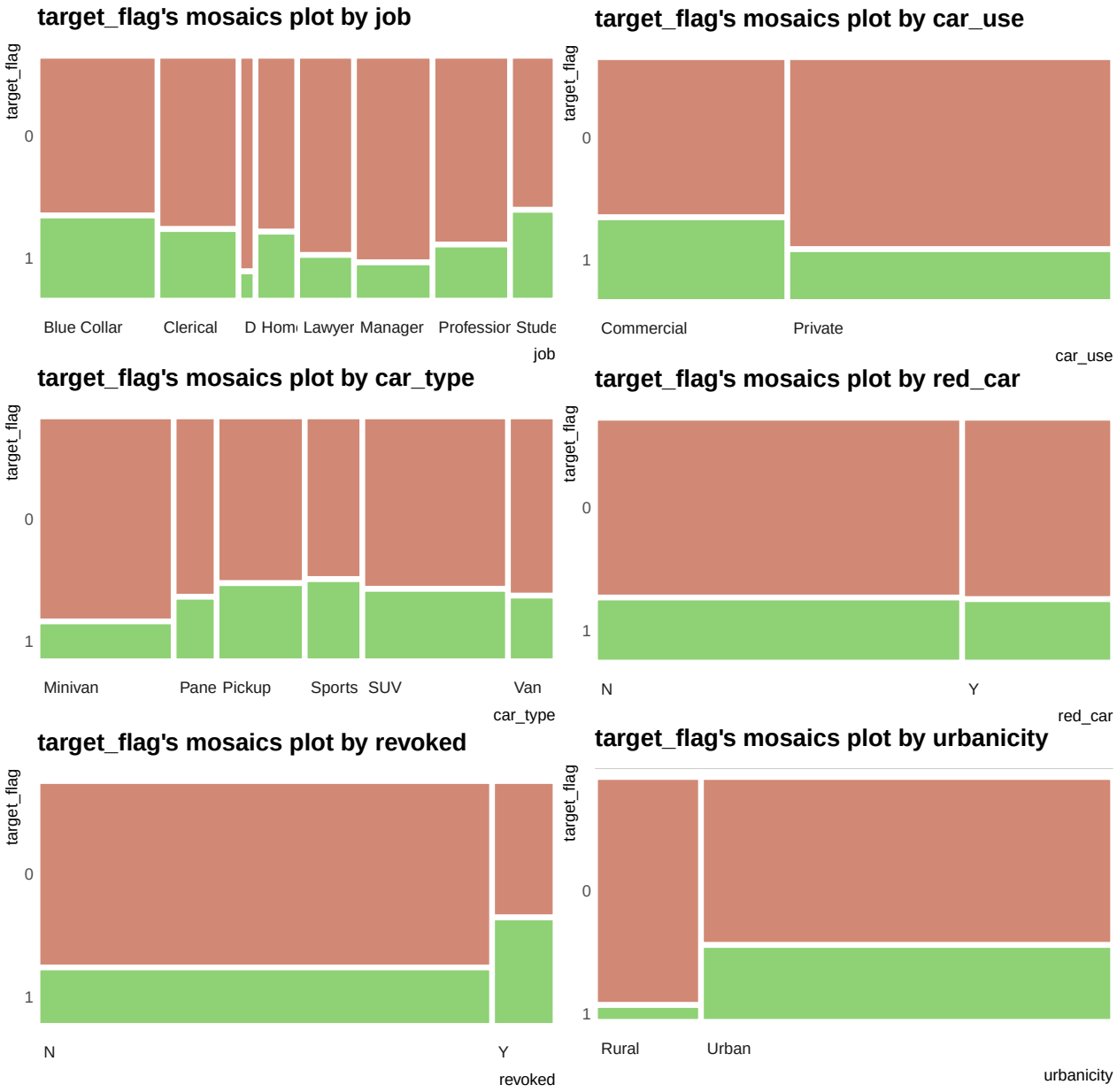
| variables | outliers_cnt | outliers_ratio | outliers_mean | with_mean | without_mean |
|---|---|---|---|---|---|
| target_amt | 1,620 | 19.851 | 7,039.976 | 1,504.325 | 133.318 |
| kidsdriv | 981 | 12.021 | 1.423 | 0.171 | 0.000 |
| homekids | 852 | 10.440 | 3.225 | 0.721 | 0.429 |
| yoj | 682 | 8.357 | 0.120 | 10.517 | 11.465 |
| oldclaim | 663 | 8.124 | 30,358.611 | 4,037.076 | 1,709.632 |
| income | 275 | 3.370 | 204,853.655 | 61,501.398 | 56,502.428 |
| tif | 160 | 1.961 | 17.869 | 5.351 | 5.101 |
| mvr_pts | 155 | 1.899 | 8.735 | 1.696 | 1.559 |
| bluebook | 104 | 1.274 | 42,806.442 | 15,709.900 | 15,360.137 |
| travtime | 63 | 0.772 | 87.492 | 33.486 | 33.066 |
| age | 32 | 0.392 | 43.688 | 44.785 | 44.789 |
| home_val | 14 | 0.172 | 663,596.571 | 154,903.497 | 154,029.347 |

| variables | outliers_cnt | outliers_ratio | outliers_mean | with_mean | without_mean |
|---|---|---|---|---|---|
| car_age | 10 | 0.123 | 25.700 | 8.345 | 8.324 |
| index | 0 | 0.000 | | 5,151.868 | 5,151.868 |
| clm_freq | 0 | 0.000 | | 0.799 | 0.799 |

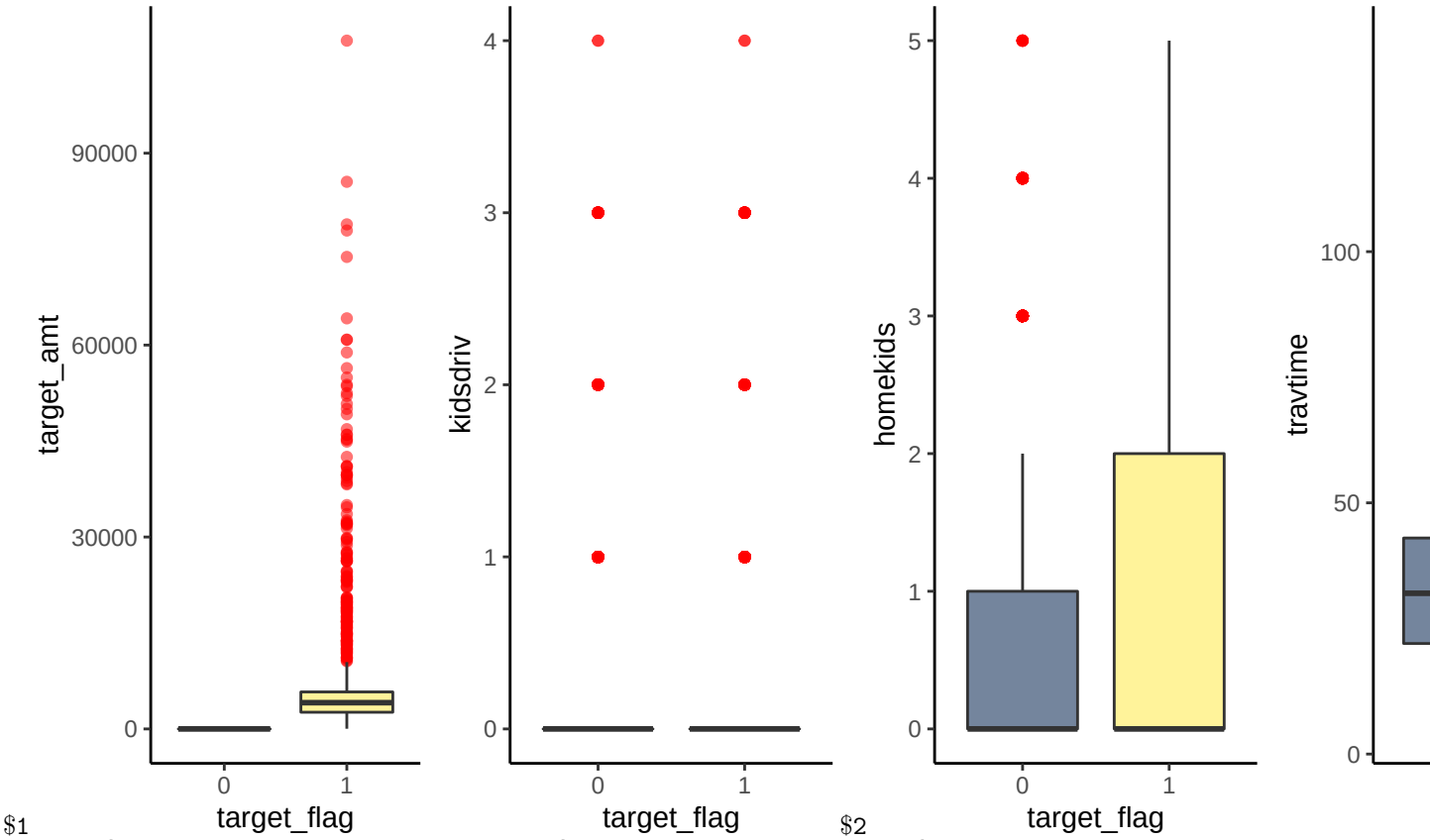
We can assess potential changes in our variable distributions in the absence of outliers as shown in the plots below. If we determine that there are legitimate reasons to remove such data due to contamination, data entry mistakes, or influence, we can remove these entries prior to modeling. However, doing so may also result in a loss of information that, in turn, reduces our ability to construct a model that generalizes well.

At this stage, we are well positioned to investigate potential relationships between our response variable (target_flag) and our predictors. For example, the following mosaic plots reveal interactions between our categorical predictors and our response, with the possible exception of two variables, 'sex' and 'red_car'. While we will still include these variables in our initial classification model, it is unlikely that they will be retained during the selection process.

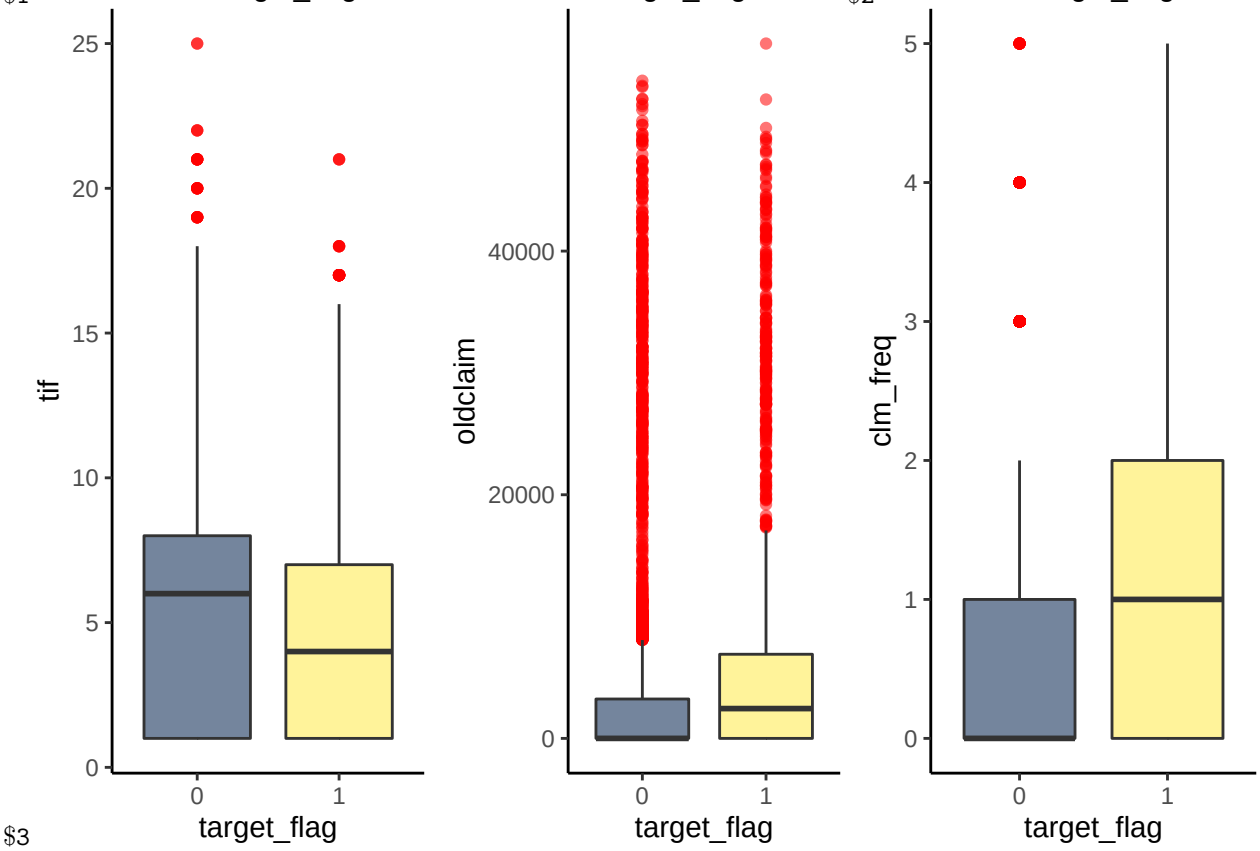




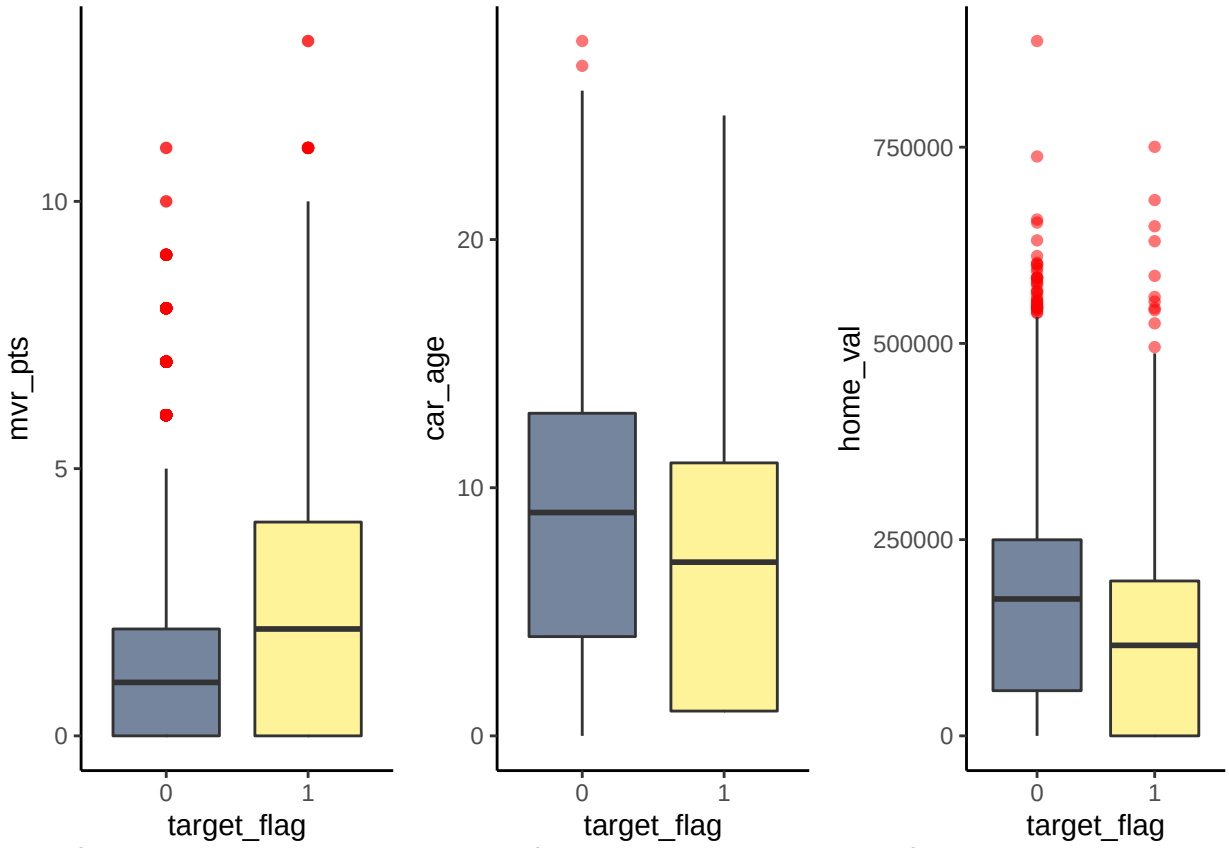

We can also evaluate interactions between numerical predictors and our response using box-plots. At first approximation, it appears that values for 'travtime', 'age', and 'yoj' are distributed similarly across the two levels of our response; suggesting that they may not have very much predictive value. However, this assessment may be conflated by the presence of outliers, particularly 'travtime'.


target_amt
90000
60000
30000
0
0
1
target_flag
kidsdriv
4
3
2
1
0
0
1
target_flag
homekids
5
4
3
2
1
0
0
1
target_flag
travtime
100
50
0

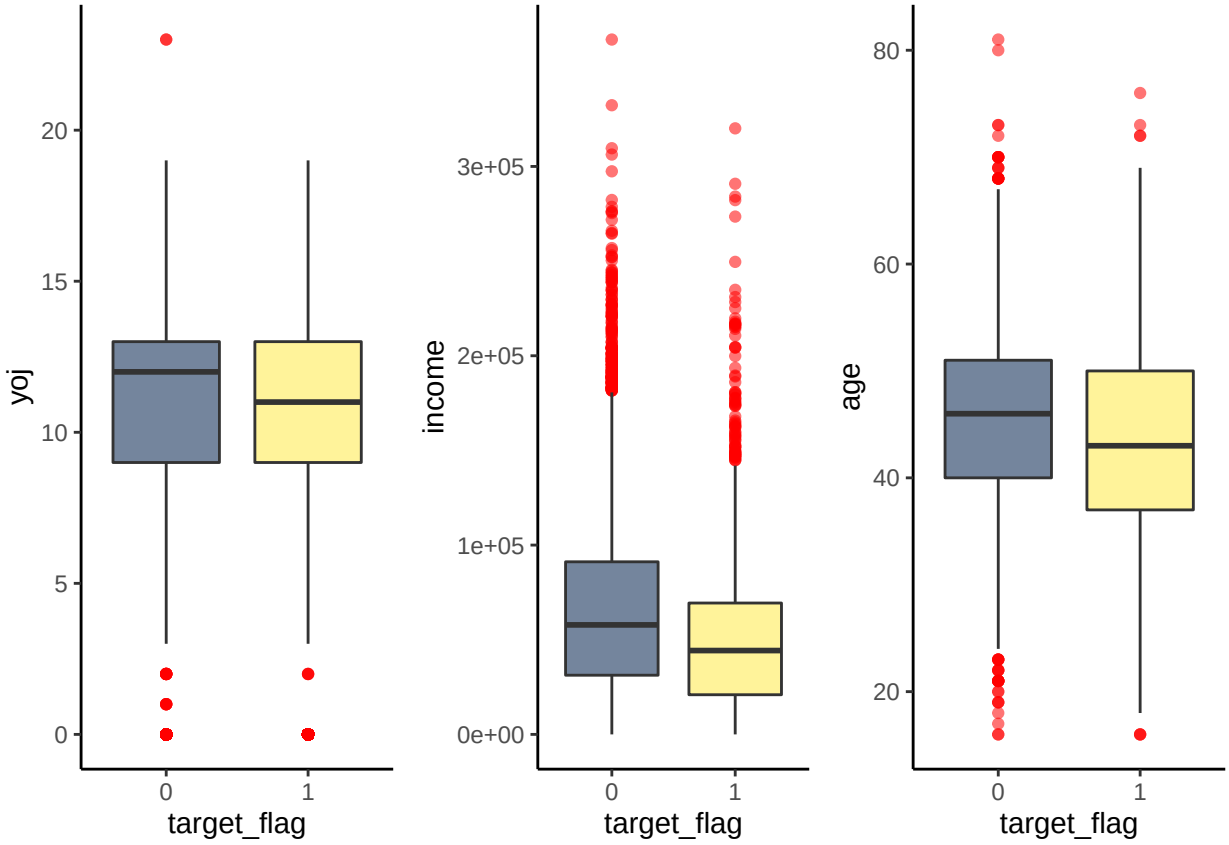
$1 $2


tif
25
20
15
10
5
0
0
1
target_flag
oldclaim
40000
20000
0
0
1
target_flag
clm_freq
5
4
3
2
1
0
0
1
target_flag

$3




$4



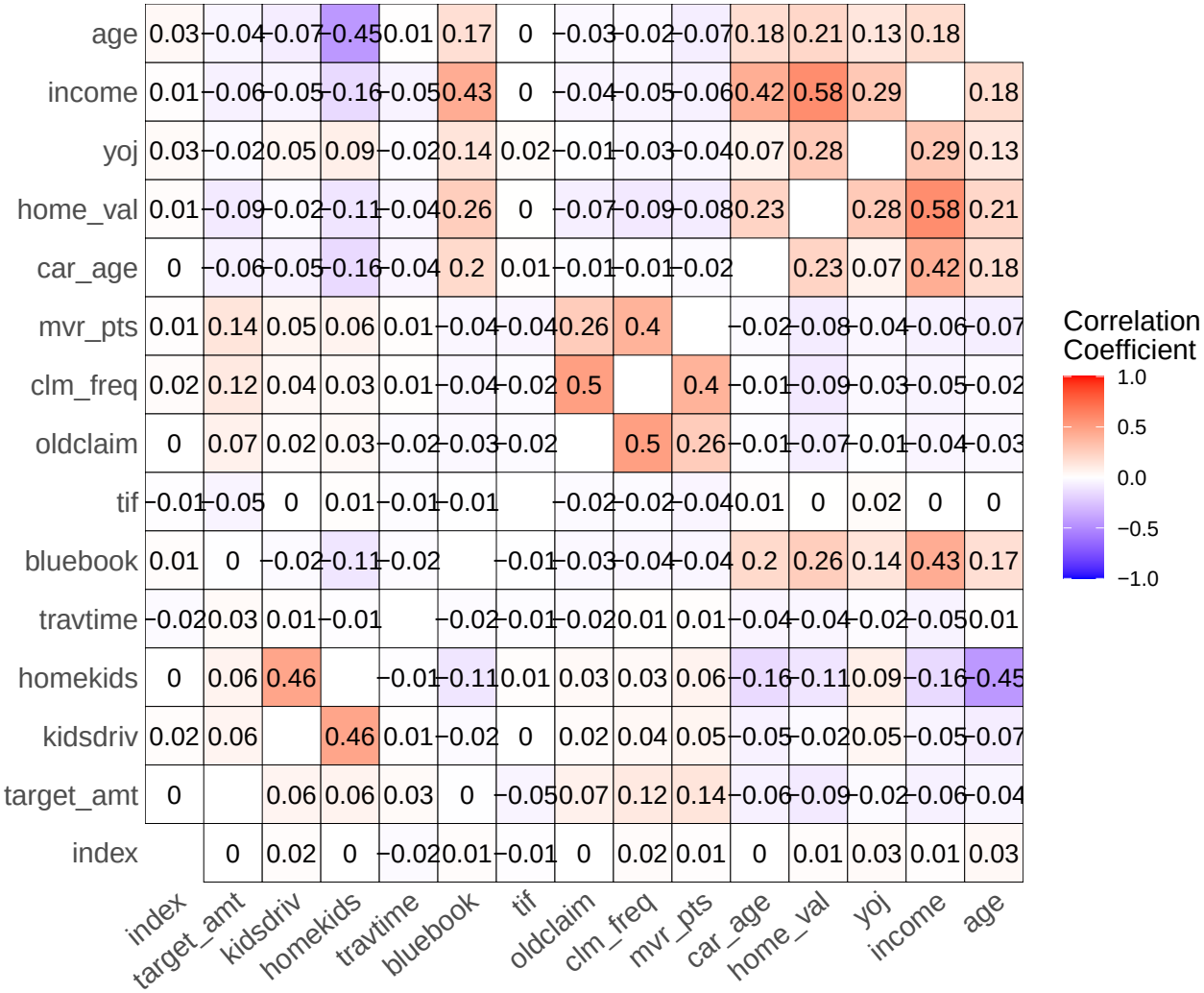

$5

attr(,“class”) [1] “list” “ggarrange” As an additional check, we can look at the relative distribution of our

variables in each response level using histograms (below). It's interesting to note the overall similarity in shape for these subsets with each variable. This pattern may be inherent to the data or, more likely, an indication that the data has been manufactured.

We should also document any substantive correlations (collinearity) between our covariates that, if unaccounted for, might lead to model misspecification. A paired correlation coefficient greater than .90 would certainly be cause for concern. We see from the correlation matrix below that 'home_val' and 'income' display the greatest correlation (~ 0.58) followed by "old_clm' and clm_freq (~0.50).

It appears that collinearity will not be a primary concern in our modeling process. Nonetheless, it may be prudent to revisit these higher correlations as we refine our model inputs - to account for interactions, etc.



## Construct Logistical Classification Model

To complete our preprocessing steps, we assessed class imbalance in our response variable, finalized modifications to our covariates, and split our data into training and test sets.

As seen in the table below, there is a 3:1 class imbalance in our response variable. However, this discrepancy is not extreme and does not require corrective steps - as might be the case with a rare event. That said, we did employ options for rebalancing the data after creating a training set. The former included under- and oversampling as well as class weights. These strategies yielded such poor results (i.e., inflated model AIC values) that we proceeded to model with unbalanced data.

Table 5: All Observations

| target_flag | n | frequency |
|---|---|---|
| 0 | 6,008 | 0.7361843 |
| 1 | 2,153 | 0.2638157 |

As noted, we did make several additional modifications to select covariates prior to building our first classification model. These included:

1. converting 'kidsdrive' to a two-level factor (Y/N) – to reduce the effect of zero inflation
2. condensing 'job' into two levels, 'blue collar' and 'professional' – to test preconceived ideas that automobile accidents and claims differentiate across these socioeconomic classes.

3. factoring and leveling the 'education' variable to reflect recognized attainment outcomes.

We then partitioned our data set into training and testing sets. Each split with a 3:1 ratio of negative (0) and positive (1) observations for our response variable.

Table 6: Training Sample

| target_flag | n | frequency |
|---|---|---|
| 0 | 4,506 | 0.7362745 |
| 1 | 1,614 | 0.2637255 |

Table 7: Test Sample

| target_flag | n | frequency |
|---|---|---|
| 0 | 1,502 | 0.7362745 |
| 1 | 538 | 0.2637255 |

### Crash Model 1: Base logistic model

Our initial classification model included the full set of predictors. And use of the Akaike information criterion (AIC) to automate variable selection.

Comparing our model's residual deviance (5499) and degrees of freedom (6094) provides one measure of model fit. On this basis, it does appear that our model performs well relative to a null expectation. This is further confirmed by our Pearson Chi Square statistic which is significant at 0.05.

After variable selection and final elimination of nonsignifiant variables, our model has the following form:

target_flag ~ kidsdriv + parent1 + mstatus + education + travtime + car_use + bluebook + tif + car_type + oldclaim + clm_freq + revoked + mvr_pts + urbanicity + home_val + income + job

| | |
|---|---|
| Observations | 6120 |
| Dependent variable | target_flag |
| Type | Generalized linear model |
| Family | binomial |
| Link | logit |

| | |
|---|---|
| $\chi^2(25)$ | 1562.15 |
| Pseudo-R² (Cragg-Uhler) | 0.33 |
| Pseudo-R² (McFadden) | 0.22 |
| AIC | 5551.32 |
| BIC | 5726.03 |

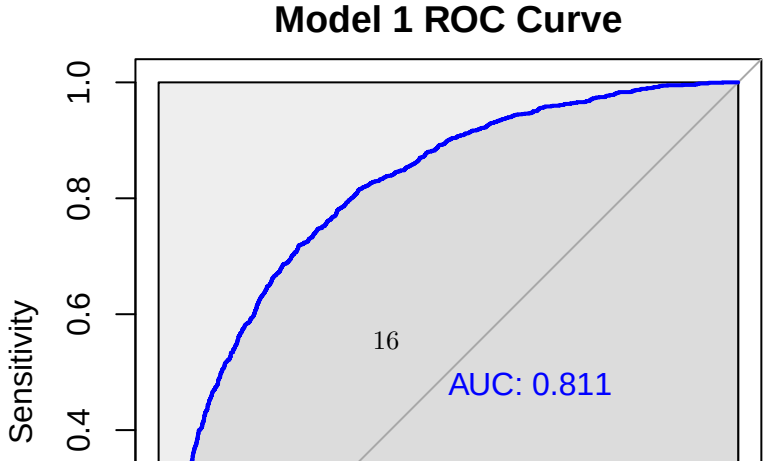
**Crash Model 1 Evaluation**

Review of our model diagnostics indicates that the model performs well (AUC = ~ .81) in discriminating between our negative (1=crash) and positive (o= no crash) class outcomes using an evaluation threshold of 0.50. This is consistent with our model sensitivity (i.e., true positive rate = 0.92) and to a lesser extent, specificity (true negative rate = 0.41) measures.

***Diagnostics***

| | Est. | S.E. | z val. | p |
|---|---|---|---|---|
| (Intercept) | -2.41 | 0.22 | -11.14 | 0.00 |
| kidsdrivY | 0.65 | 0.11 | 5.88 | 0.00 |
| homekids | 0.07 | 0.04 | 1.92 | 0.05 |
| parent1Y | 0.30 | 0.13 | 2.36 | 0.02 |
| mstatusY | -0.49 | 0.10 | -4.97 | 0.00 |
| educationBachelors | -0.54 | 0.09 | -6.15 | 0.00 |
| educationMasters | -0.39 | 0.11 | -3.53 | 0.00 |
| educationPhD | -0.50 | 0.16 | -3.06 | 0.00 |
| travtime | 0.02 | 0.00 | 7.27 | 0.00 |
| car_usePrivate | -0.75 | 0.09 | -8.00 | 0.00 |
| bluebook | -0.00 | 0.00 | -5.16 | 0.00 |
| tif | -0.05 | 0.01 | -6.48 | 0.00 |
| car_typePanel Truck | 0.69 | 0.17 | 4.15 | 0.00 |
| car_typePickup | 0.59 | 0.11 | 5.12 | 0.00 |
| car_typeSports Car | 1.03 | 0.12 | 8.39 | 0.00 |
| car_typeSUV | 0.68 | 0.10 | 6.95 | 0.00 |
| car_typeVan | 0.72 | 0.14 | 5.19 | 0.00 |
| oldclaim | -0.00 | 0.00 | -2.99 | 0.00 |
| clm_freq | 0.20 | 0.03 | 6.00 | 0.00 |
| revokedY | 0.90 | 0.10 | 8.65 | 0.00 |
| mvr_pts | 0.12 | 0.02 | 7.61 | 0.00 |
| urbanicityUrban | 2.32 | 0.13 | 17.87 | 0.00 |
| home_val | -0.00 | 0.00 | -3.60 | 0.00 |
| yoj | -0.01 | 0.01 | -1.61 | 0.11 |
| income | -0.00 | 0.00 | -3.12 | 0.00 |
| jobProfessional | -0.14 | 0.09 | -1.46 | 0.14 |

Standard errors: MLE



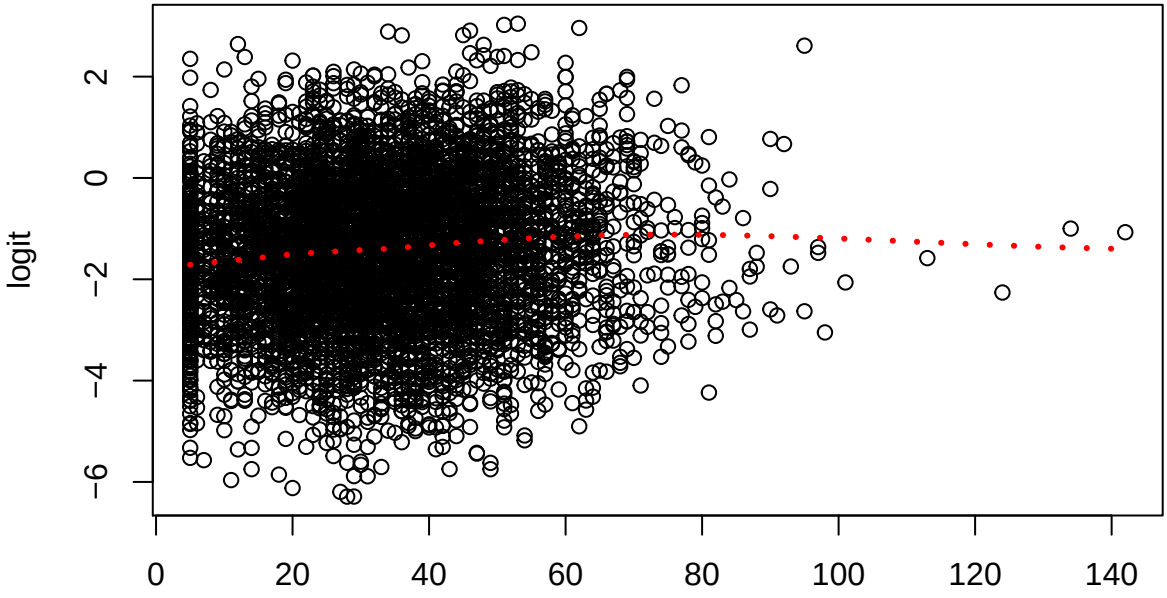
***Dispersion***

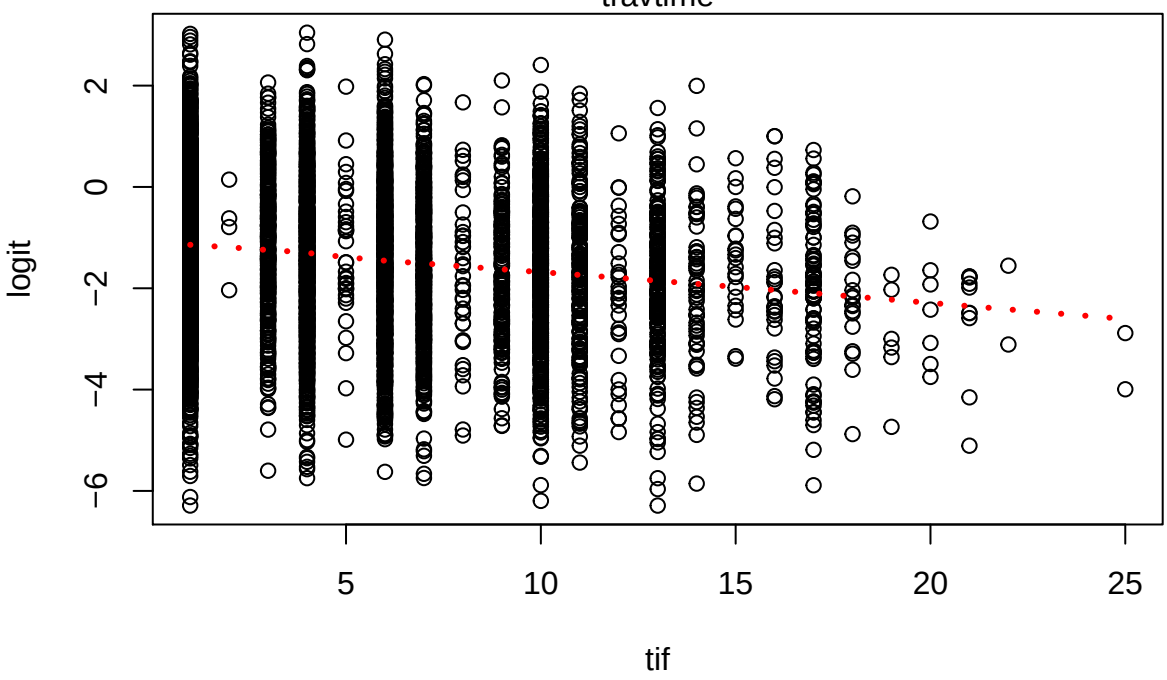
We did not detect overdispersion in the model1 based on our assessment of the residual deviance and chi-square test.

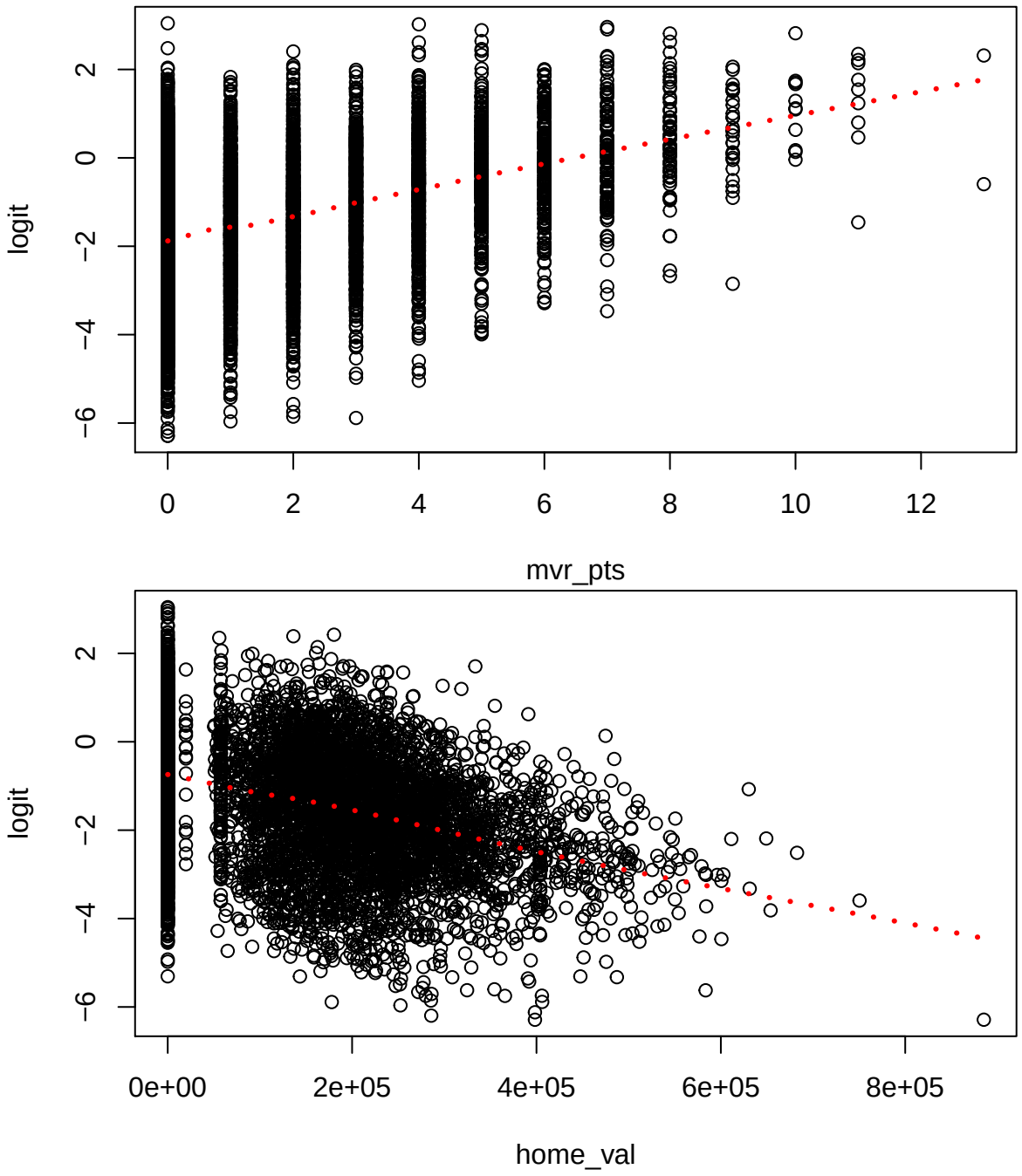
[1] "We divide the deviance by the residuals to obtain the value 0.9024. There is no overt concern since the values is not greater than 1" [1] "Next we obtain a Chi-Squared test statistic of 0.3133 This communicates that the null hypothesis is not rejected and their are no problems with dispersion."

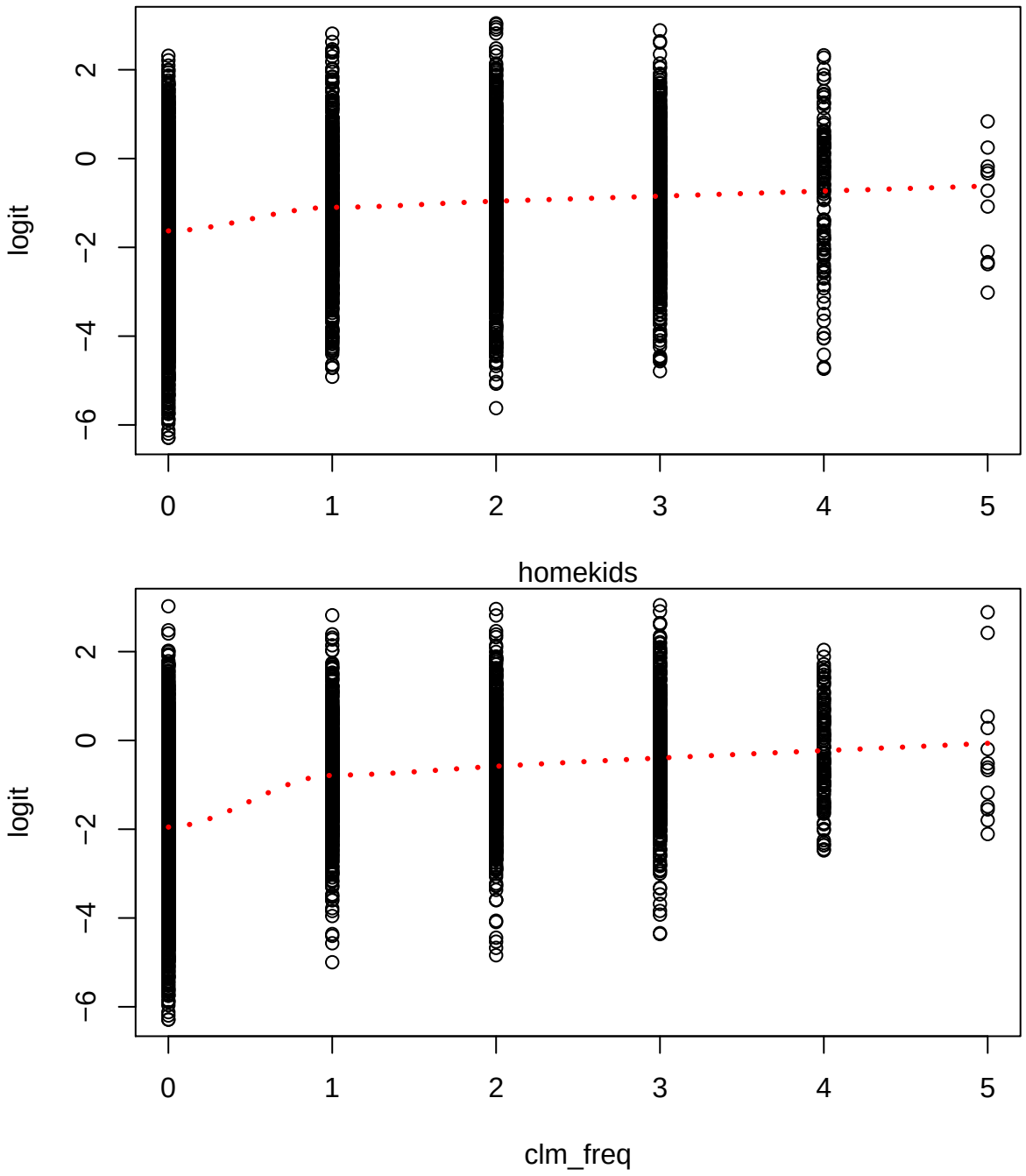
***Assumption of Linearity***

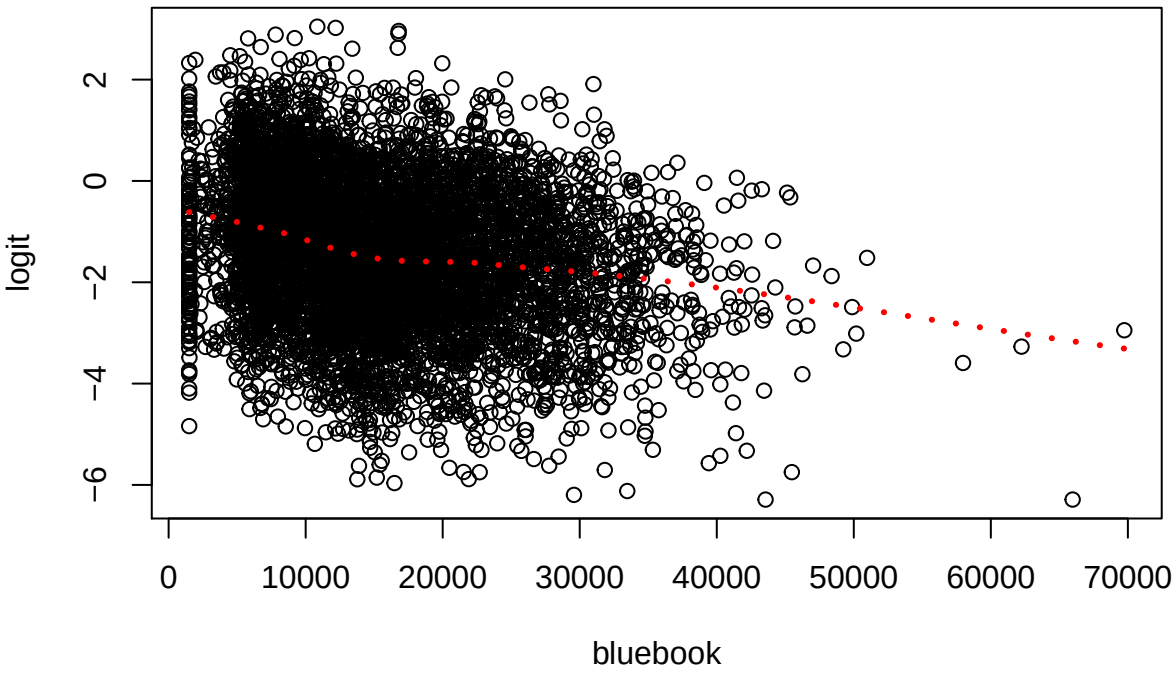
A key assumption in a logistic model is that the logit and covariates are linearly related. Our tests for linearity confirm the validity of our model with the possible exception of 'home-kids'. We note, however, that evidence for nonlinearity is not clear in this case.




logit
2
0
−2
−4
−6
0
2
4
6
8
10
12
mvr_pts
logit
2
0
−2
−4
−6
0e+00
2e+05
4e+05
6e+05
8e+05
home_val

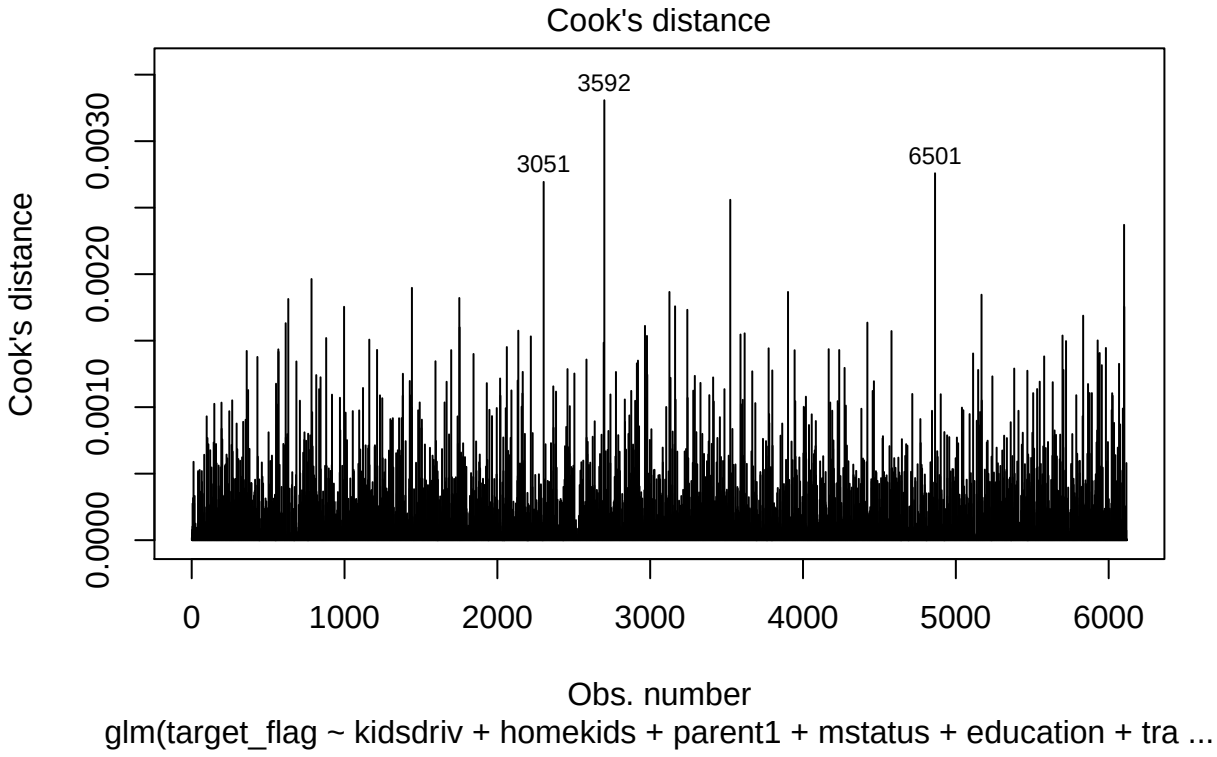

logit
2
0
−2
−4
−6
0
1
2
3
4
5
homekids
logit
2
0
−2
−4
−6
0
1
2
3
4
5
clm_freq

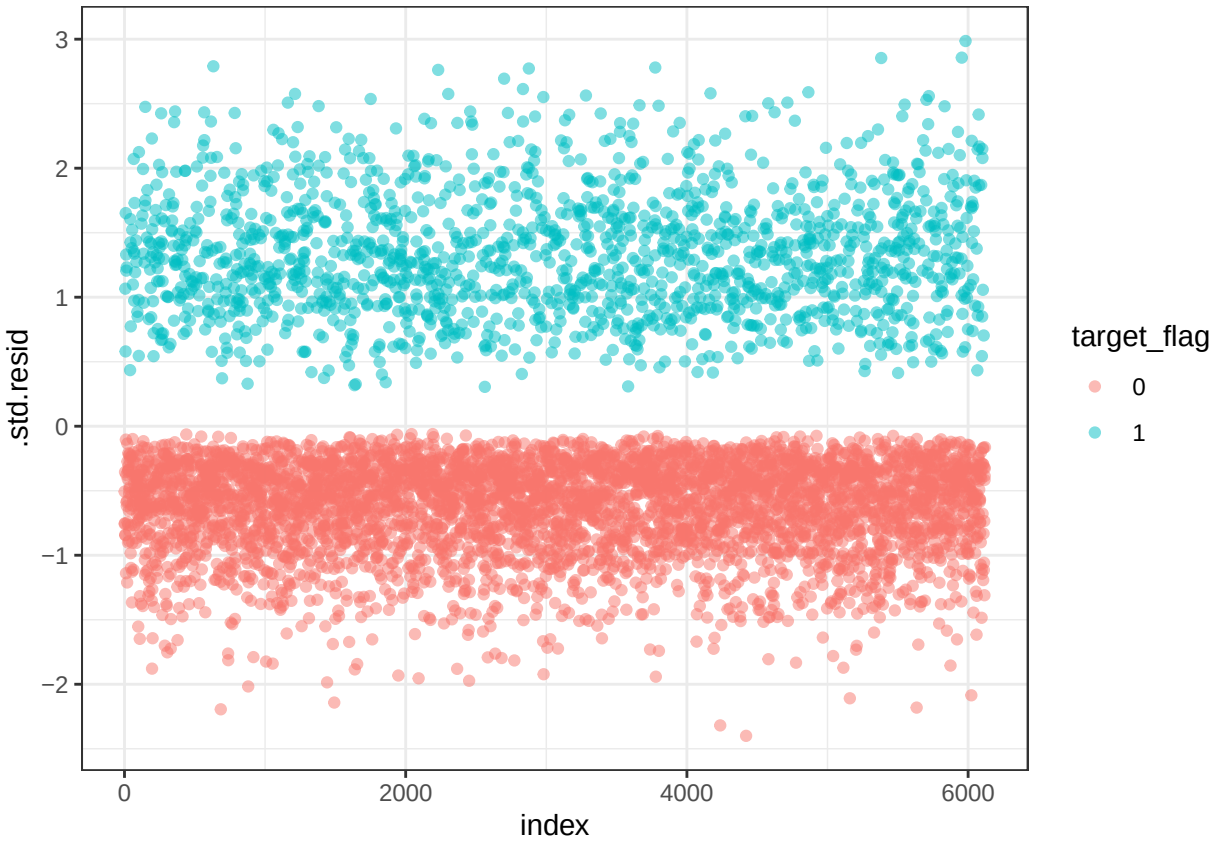



***Outliers & Influential Points***

To assess any impact of influential observations on our model results we examined both the standardized residuals (.std.resid) and the Cook's distance (.cooksd). While the latter distinguished several notable outliers (3722, 3592, 6501), these data were not influential (D >1). This is consistent with our plot of standardized residuals (below) - with no observations exeeding 3 standard deviations.

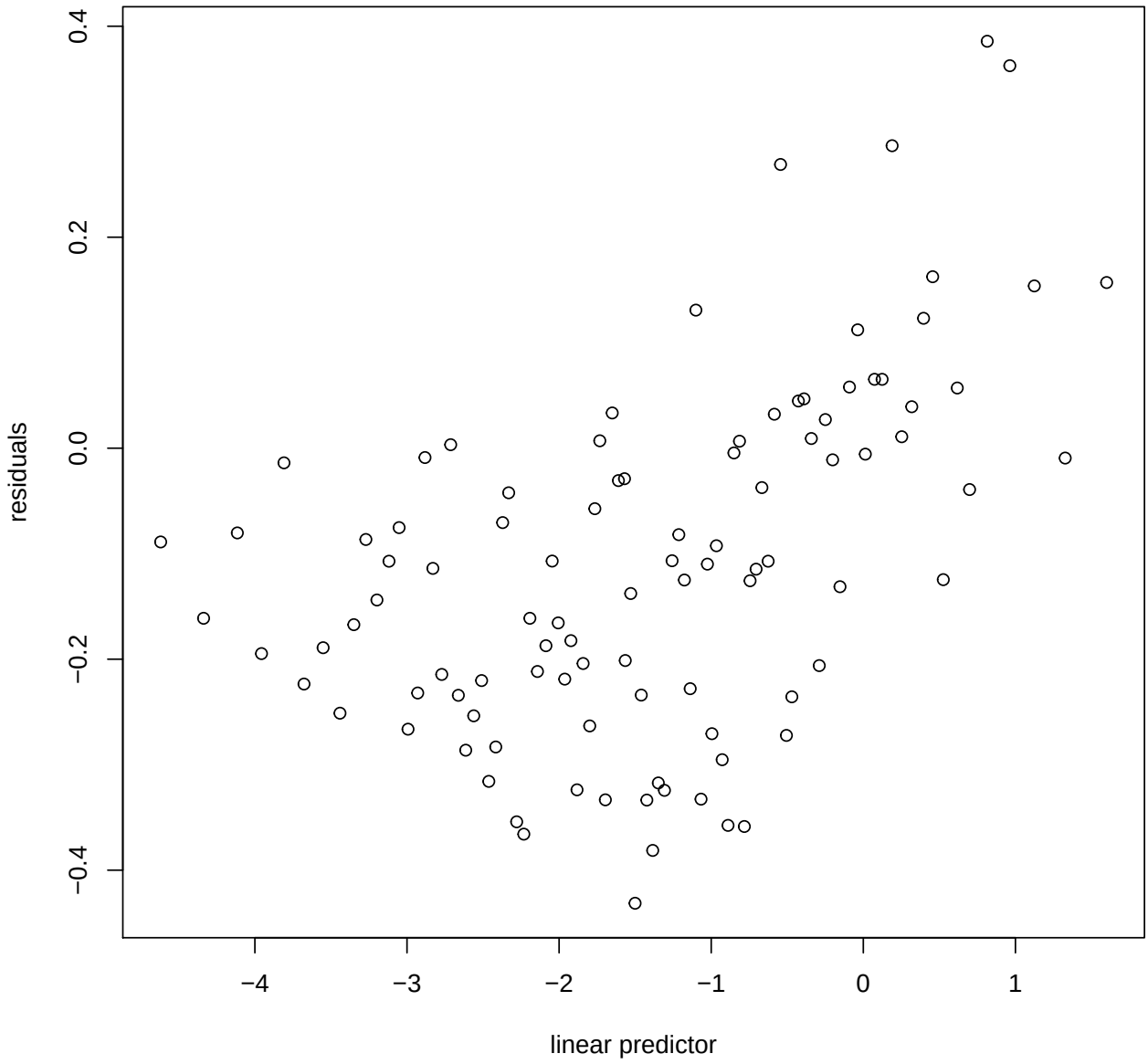
**http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/





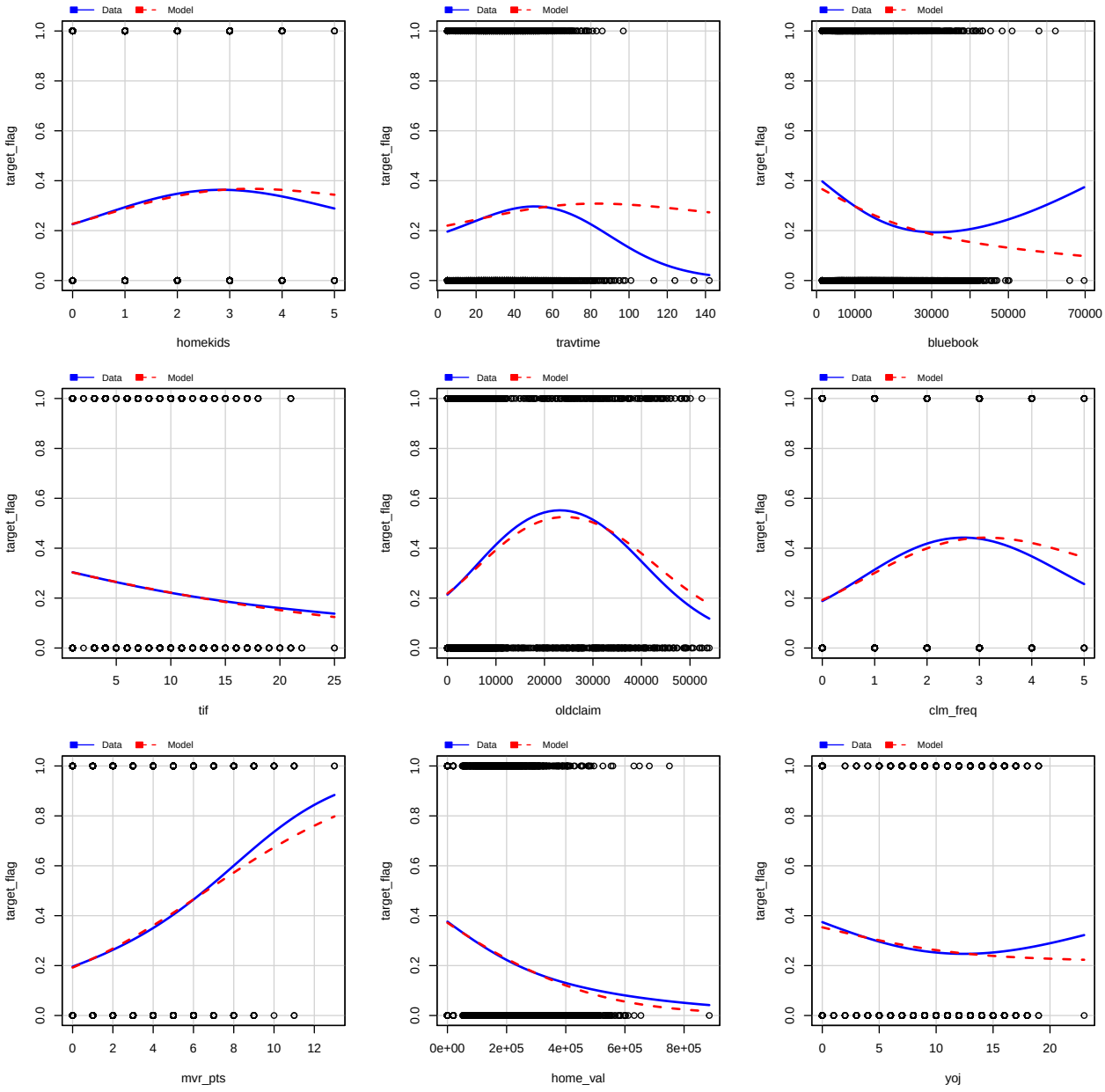

***Check for Independence***

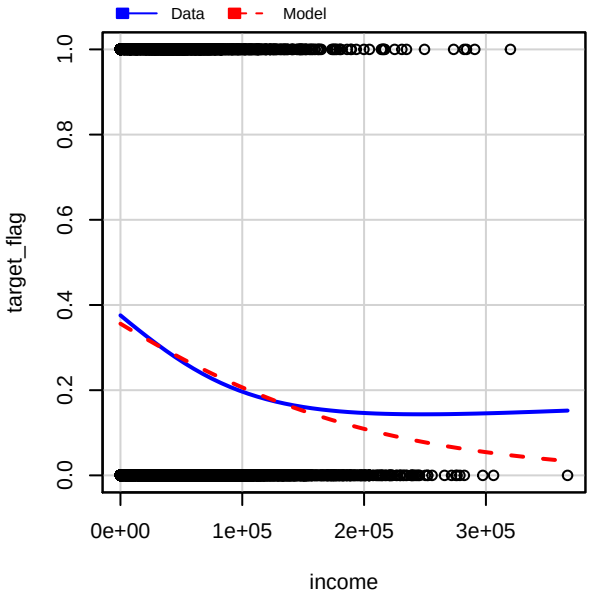
An additional assumption in a logistic regression model is independence among model residuals. This can be assessed by plotting binned residuals against the linear predictor (logit). On this basis, there does appear to be a pattern of dependence among our residuals (below) which points to a model misspecification. We speculate that this may derive from grouping effects among our categorical variables, given that we are not using time series data. Despite extensive experimentation, we were not able to remove this effect from this model or others that follow.

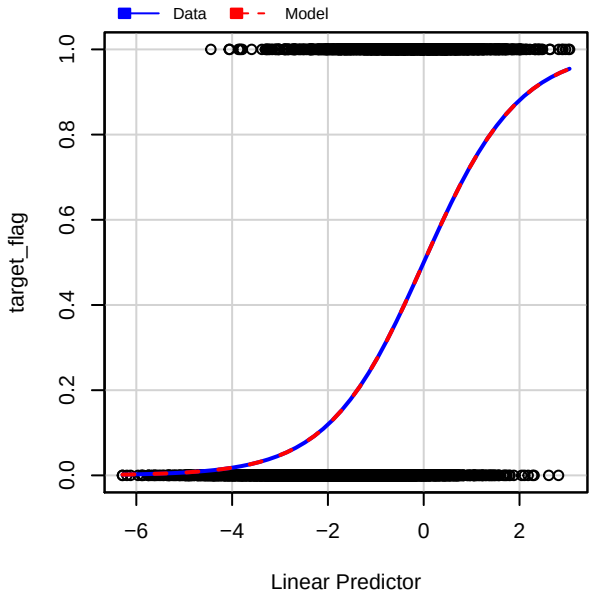



***Goodness of Fit - marginal plots***

As a final check on our fitted model, we plotted the marginal effects for each of our variables (below). Our results indicated that variable transformations may help improve the fit for travtime, bluebook, income, and possible, clm_freq.


Data
Model
target_flag
homekids
travtime
bluebook
tif
oldclaim
clm_freq
mvr_pts
home_val
yoj


Data
Model
target_flag
0.0
0.2
0.4
0.6
0.8
1.0
0e+00
1e+05
2e+05
3e+05
income


Data
Model
target_flag
0.0
0.2
0.4
0.6
0.8
1.0
−6
−4
−2
0
2
Linear Predictor

## Crash Model 2: Apply Predictor Transformations

Our second classification model, model2, included transformations on 'income' (square root), 'bluebook' (log), and travtime (quadratic). We decided on these transformations on the basis of their distributions as well as iterative experimentation with the model. Similar to model1, we employed AIC to assist variable selection. Model2 has the following form after selection:

target_flag = kidsdriv + parent1 + mstatus + education + car_use + tif + car_type + oldclaim + revoked + urbanicity + home_val + job + travtime + I(travtimeˆ2) + mvr_pts + clm_freq + log_bluebook + sqrt_income

Our model AIC (5530) decreased slightly relative to model1 (5551) indicating a slightly better fit to the data. This is also indicated by the slight increase in our model's AUC (.831). Other measure of performance were similar to model1 (sensitivity = .92, specificity = .42).

The transformations applied to 'income', 'bluebook', and 'travtime' did improve the marginal effects of our model, as inferred from our marginal plots (below).

| | |
|---|---|
| Observations | 6120 |
| Dependent variable | target_flag |
| Type | Generalized linear model |
| Family | binomial |
| Link | logit |

| | |
|---|---|
| $\chi^2(24)$ | 1581.45 |
| Pseudo-R² (Cragg-Uhler) | 0.33 |
| Pseudo-R² (McFadden) | 0.22 |
| AIC | 5530.02 |
| BIC | 5698.00 |

| | Est. | S.E. | z val. | p |
|---|---|---|---|---|
| (Intercept) | 0.44 | 0.62 | 0.72 | 0.47 |
| kidsdrivY | 0.73 | 0.10 | 7.09 | 0.00 |
| parent1Y | 0.40 | 0.11 | 3.66 | 0.00 |
| mstatusY | -0.49 | 0.09 | -5.15 | 0.00 |
| educationBachelors | -0.52 | 0.09 | -5.93 | 0.00 |
| educationMasters | -0.37 | 0.11 | -3.37 | 0.00 |
| educationPhD | -0.53 | 0.15 | -3.43 | 0.00 |
| car_usePrivate | -0.75 | 0.09 | -7.97 | 0.00 |
| tif | -0.05 | 0.01 | -6.35 | 0.00 |
| car_typePanel Truck | 0.60 | 0.16 | 3.73 | 0.00 |
| car_typePickup | 0.60 | 0.11 | 5.24 | 0.00 |
| car_typeSports Car | 1.02 | 0.12 | 8.26 | 0.00 |
| car_typeSUV | 0.70 | 0.10 | 7.11 | 0.00 |
| car_typeVan | 0.74 | 0.14 | 5.35 | 0.00 |
| oldclaim | -0.00 | 0.00 | -3.01 | 0.00 |
| revokedY | 0.91 | 0.10 | 8.74 | 0.00 |
| urbanicityUrban | 2.31 | 0.13 | 17.79 | 0.00 |
| home_val | -0.00 | 0.00 | -3.38 | 0.00 |
| jobProfessional | -0.19 | 0.10 | -1.93 | 0.05 |
| travtime | 0.04 | 0.01 | 4.91 | 0.00 |
| I(travtime^2) | -0.00 | 0.00 | -2.92 | 0.00 |
| mvr_pts | 0.12 | 0.02 | 7.70 | 0.00 |
| clm_freq | 0.20 | 0.03 | 6.02 | 0.00 |
| log_bluebook | -0.37 | 0.06 | -5.79 | 0.00 |
| sqrt_income | -0.00 | 0.00 | -4.55 | 0.00 |

Standard errors: MLE

**Crash Model 2 Evaluation**

***Diagnostics***

Confusion Matrix and Statistics

```
      Reference
```

Prediction 0 1 0 4163 928 1 343 686

```
           Accuracy : 0.7923
```

```
            95% CI : (0.7819, 0.8024)
No Information Rate : 0.7363
P-Value [Acc > NIR] : < 2.2e-16

              Kappa : 0.3948
```

Mcnemar's Test P-Value : < 2.2e-16

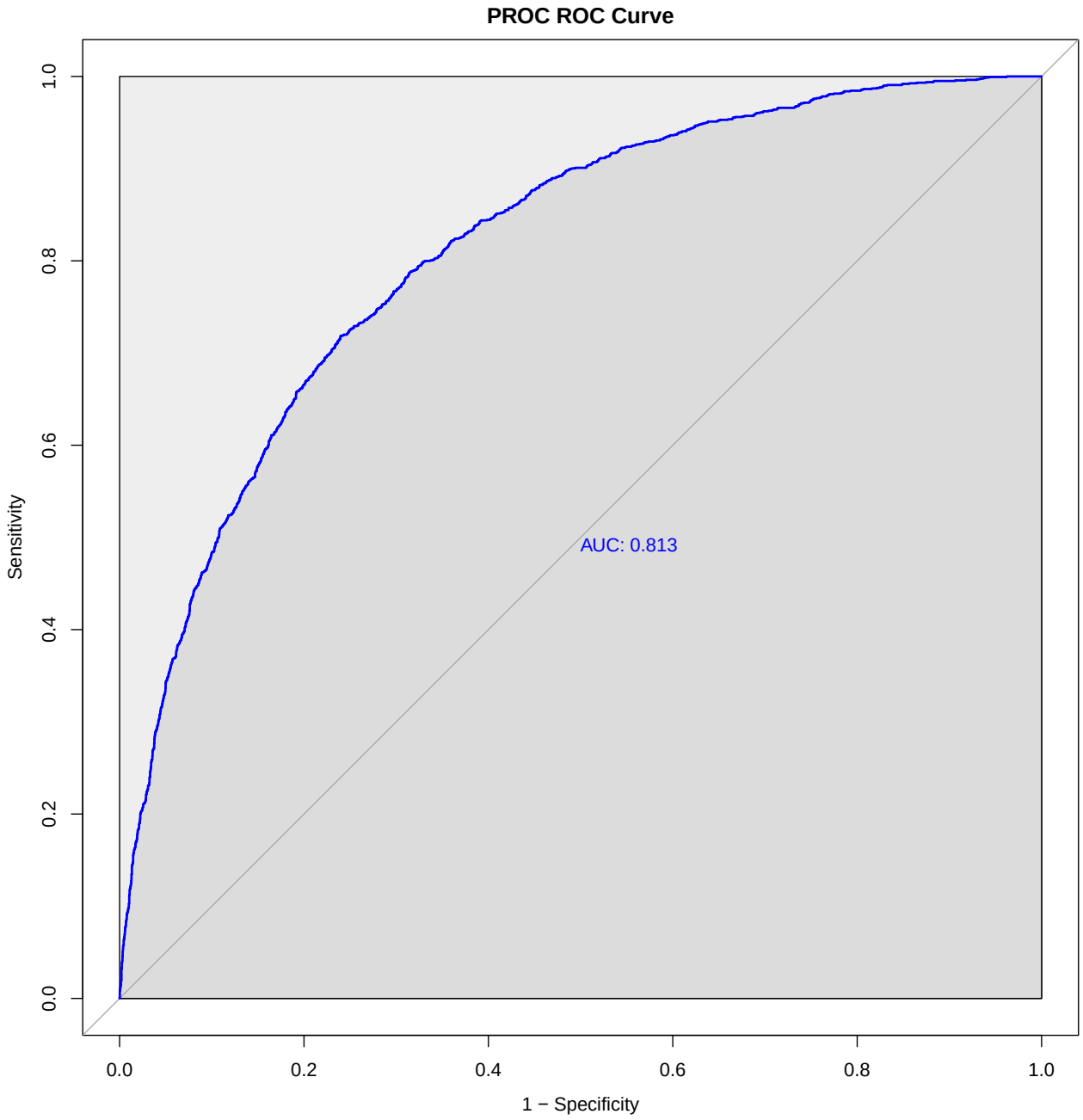
```
        Sensitivity : 0.9239
        Specificity : 0.4250
     Pos Pred Value : 0.8177
     Neg Pred Value : 0.6667
         Prevalence : 0.7363
     Detection Rate : 0.6802
```

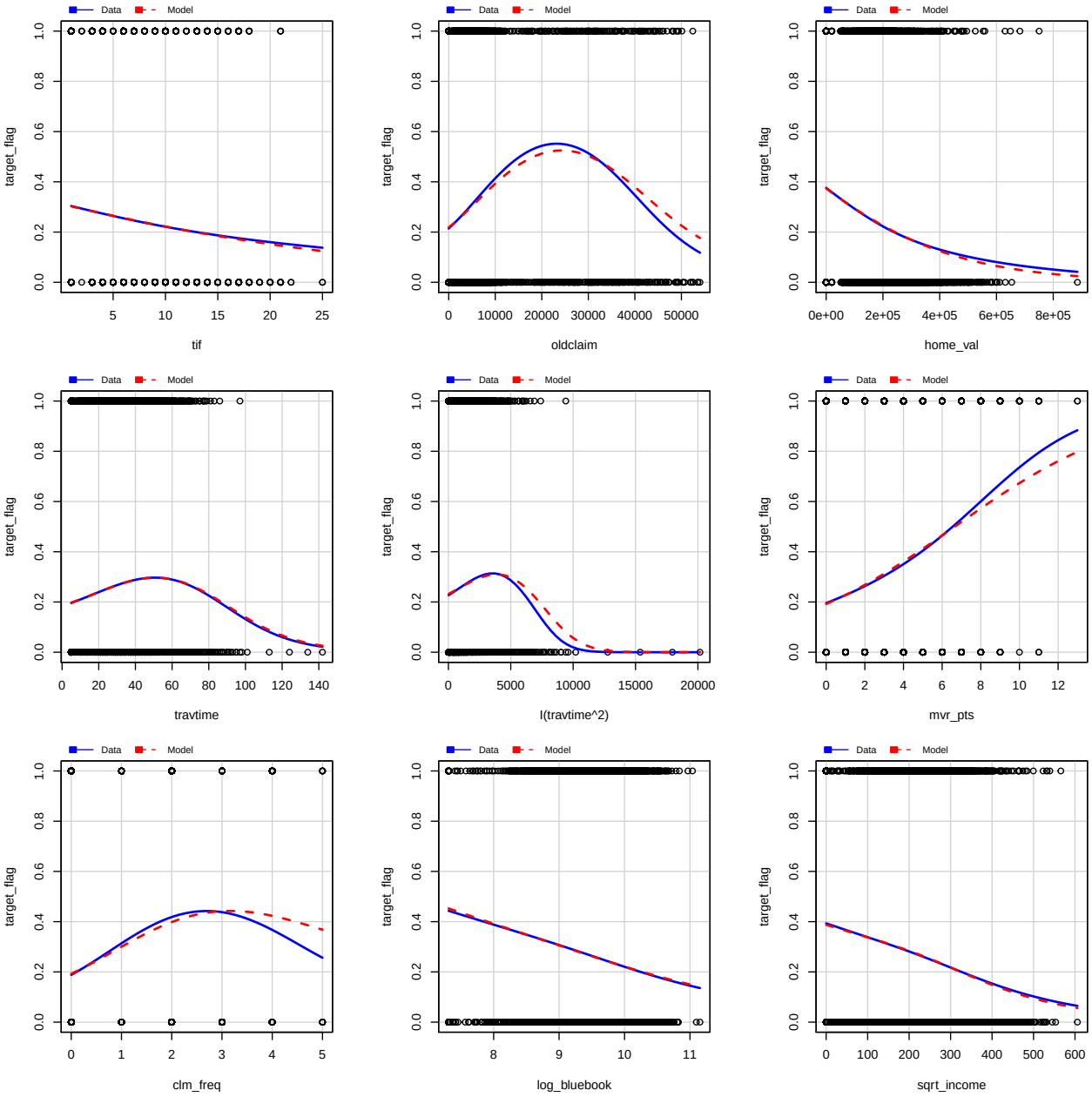
Detection Prevalence : 0.8319
Balanced Accuracy : 0.6745

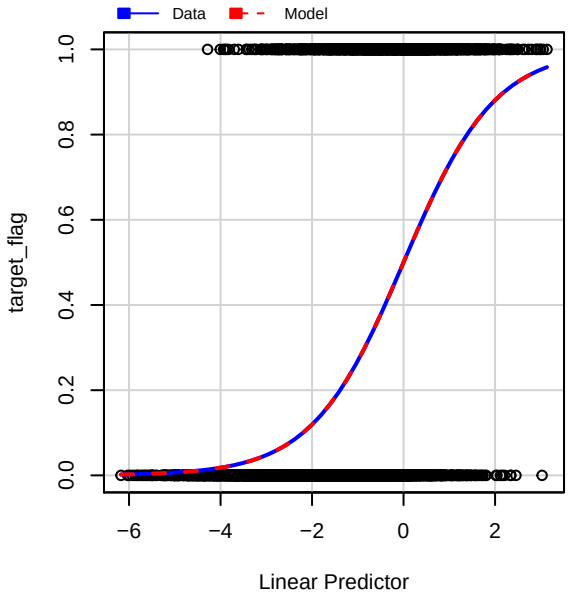
```
   'Positive' Class : 0
```

## A tibble: 1 x 12

model predictors sensitivity specificity pos_rate neg_rate precision recall 1 tran~ 24 0.924 0.425 0.818 0.667 0.818 0.924 # ... with 4 more variables: f1 , auc , AIC , BIC

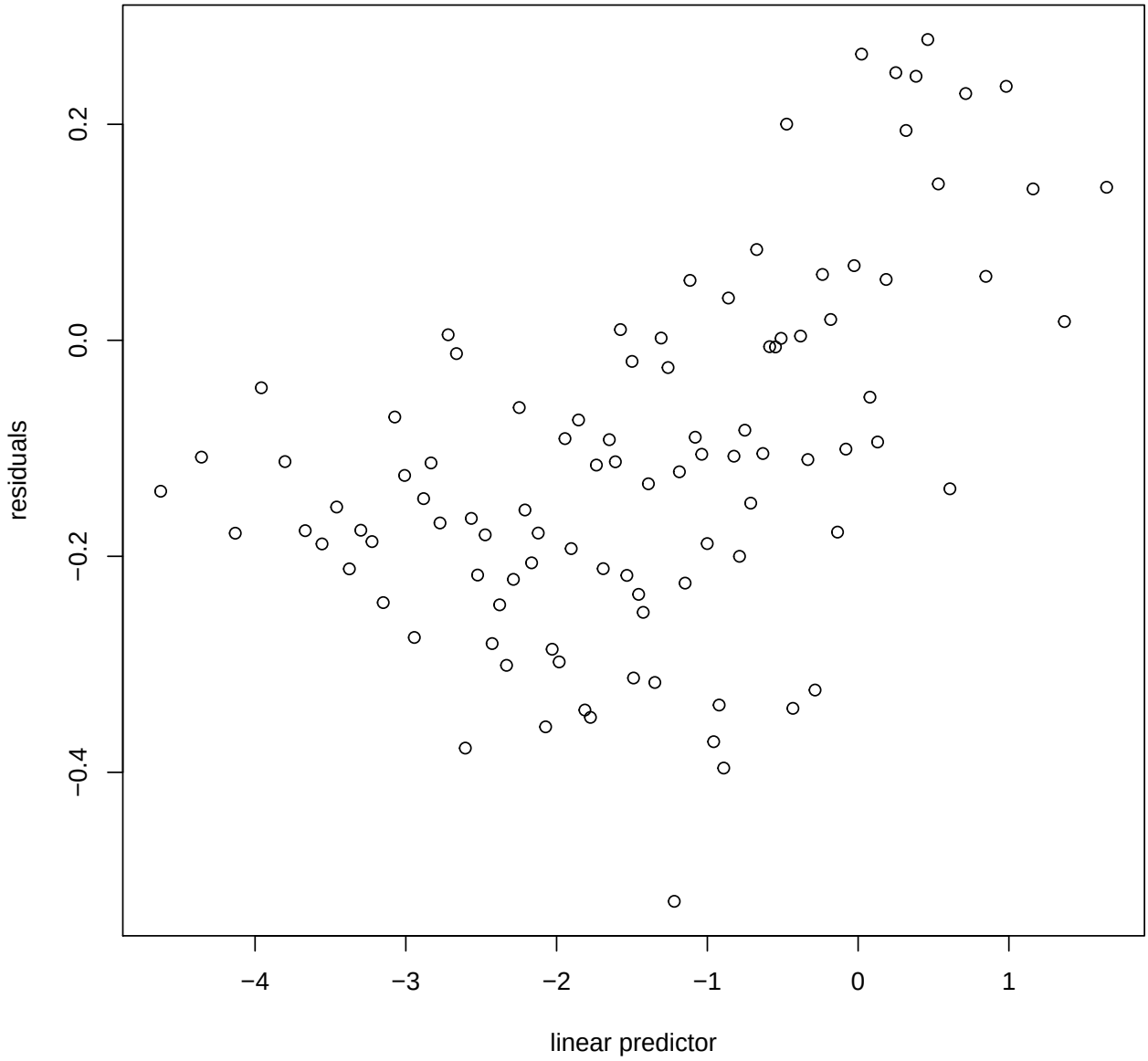

PROC ROC Curve
Sensitivity
1 − Specificity
0.0
0.2
0.4
0.6
0.8
1.0
AUC: 0.813


Data
Model
target_flag
tif
oldclaim
home_val
travtime
I(travtime^2)
mvr_pts
clm_freq
log_bluebook
sqrt_income


Data
Model
target_flag
0.0
0.2
0.4
0.6
0.8
1.0
−6
−4
−2
0
2
Linear Predictor

***Independence***

The persistence of a pattern means that we cannot yet rule out the possibility that the model is misspecified.



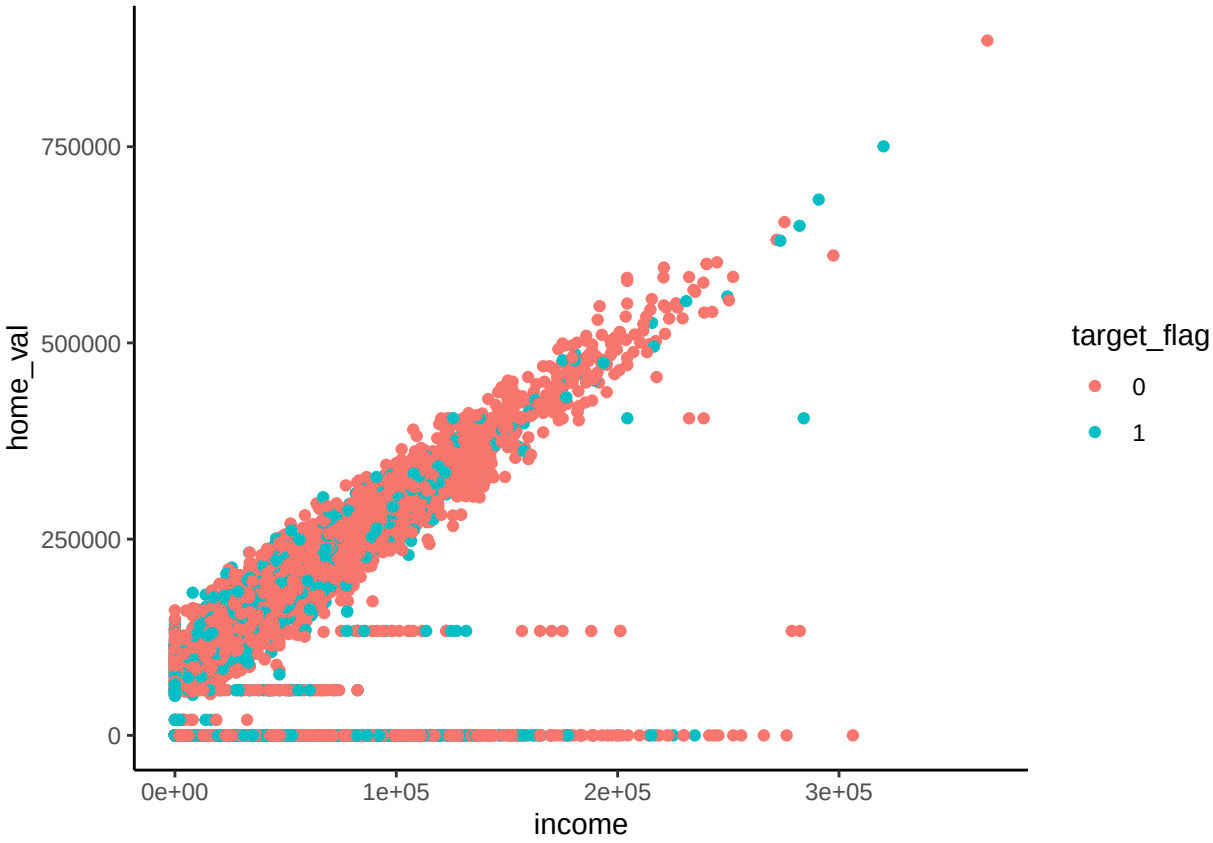
## Crash Model 3 - Feature engineering and Interactions among predictor variables

For our final model, model3, we established a new variable called 'liquidity' which comprised the a ratio of home value to income - with 1 added to numerator and denomenator to offset the effect of zero values. We established this variable to account for correlation between these variables (see below) and to address possible contamination in the data. For example the excess of zero values for 'home_val' and 'income' could owe to the presence of renters, under-age dependents, and/or unemployed drivers.

To better capture grouping effects (see histograms) and to simplify our model we constructed the following factor types from selected numerical variables.

`-mvr_pts` = factor("none", "low", "high") `-investment` = factor("low", "high") `-tif` factor=("low", "moderate", "high") `-clm_freq` = factor("none","moderate","high")

We also tested for possible interactions between categorical variables (CramersV) and continuous and categorical variables (logistic regression). Our results indicated possible interactions between 'car_use' and 'car_type', 'urbanicity' and 'travtime', 'mvr_pts' and 'revoked', and 'car_type' and 'clm_freq'. We included these interactions in model3.




Model3 had the following form after variable selection with AiC

target_flag = kidsdriv + parent1 + mstatus + education + travtime + car_use + log(bluebook) + tif + car_type + oldclaim + clm_freq + revoked + mvr_pts + urbanicity + liquidity + car_use:car_type + travtime:urbanicity

Our model AIC (5586) incrased relative to model2 and model1 indicating a slightly poorer fit to the data. This is also indicated by the slight decrease in our model3's AUC (.81). Other measure of performance were similar to model1 (sensitivity = .92, specificity = .40).

Other model diagnostics (e.g., linearity, dispersion, influential obs, residual independence) yielded resuts similar to model1 and model2.

| | |
|---|---|
| Observations | 6120 |
| Dependent variable | target_flag |
| Type | Generalized linear model |
| Family | binomial |
| Link | logit |

| | |
|---|---|
| $\chi^2(29)$ | 1535.15 |
| Pseudo-R² (Cragg-Uhler) | 0.32 |
| Pseudo-R² (McFadden) | 0.22 |
| AIC | 5586.33 |
| BIC | 5787.90 |

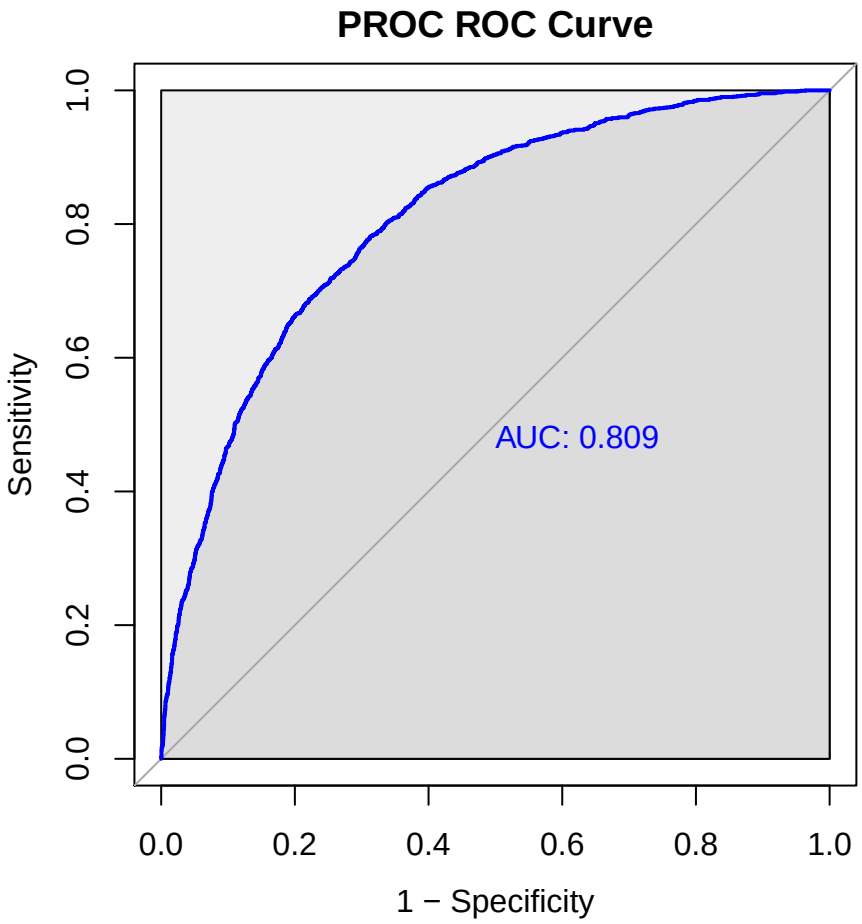
***Diagnostics***

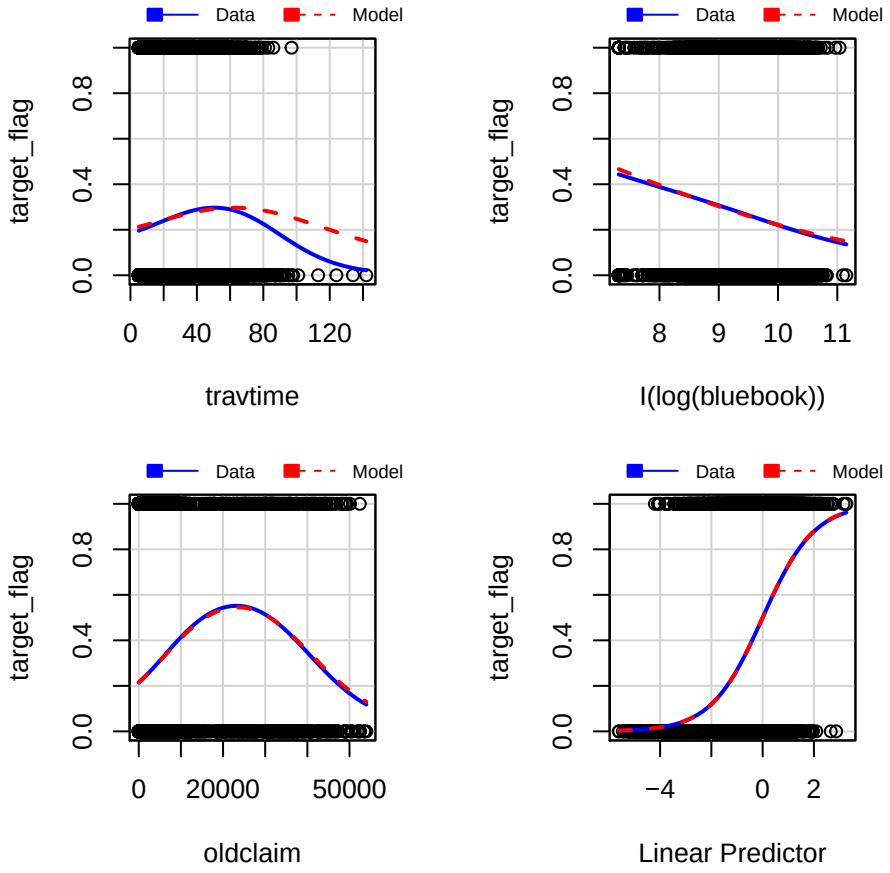
## A tibble: 1 x 12

model predictors sensitivity specificity pos_rate neg_rate precision recall 1 Feat~ 30 0.924 0.398 0.811 0.653 0.811 0.924 # ... with 4 more variables: f1 , auc , AIC , BIC

| | Est. | S.E. | z val. | p |
|---|---|---|---|---|
| (Intercept) | 1.40 | 0.65 | 2.15 | 0.03 |
| kidsdrivY | 0.69 | 0.10 | 6.78 | 0.00 |
| parent1Y | 0.42 | 0.11 | 3.87 | 0.00 |
| mstatusY | -0.43 | 0.09 | -4.56 | 0.00 |
| educationBachelors | -0.71 | 0.08 | -8.55 | 0.00 |
| educationMasters | -0.70 | 0.10 | -7.29 | 0.00 |
| educationPhD | -1.03 | 0.14 | -7.53 | 0.00 |
| travtime | 0.00 | 0.01 | 0.46 | 0.64 |
| car_usePrivate | -0.56 | 0.17 | -3.29 | 0.00 |
| I(log(bluebook)) | -0.45 | 0.06 | -7.43 | 0.00 |
| tifmoderate | -0.36 | 0.07 | -5.02 | 0.00 |
| tifhigh | -0.43 | 0.12 | -3.66 | 0.00 |
| car_typePanel Truck | 0.67 | 0.19 | 3.54 | 0.00 |
| car_typePickup | 0.80 | 0.17 | 4.58 | 0.00 |
| car_typeSports Car | 0.72 | 0.27 | 2.65 | 0.01 |
| car_typeSUV | 0.90 | 0.19 | 4.64 | 0.00 |
| car_typeVan | 0.95 | 0.19 | 4.90 | 0.00 |
| oldclaim | -0.00 | 0.00 | -4.28 | 0.00 |
| clm_freqmoderate | 0.71 | 0.09 | 7.89 | 0.00 |
| clm_freqhigh | 0.98 | 0.20 | 4.95 | 0.00 |
| revokedY | 0.97 | 0.11 | 9.21 | 0.00 |
| mvr_ptslow | 0.25 | 0.08 | 3.25 | 0.00 |
| mvr_ptshigh | 0.47 | 0.09 | 5.07 | 0.00 |
| urbanicityUrban | 1.64 | 0.29 | 5.75 | 0.00 |
| liquidityhigh | -0.36 | 0.09 | -4.08 | 0.00 |
| car_usePrivate:car_typePanel Truck | | | | |
| car_usePrivate:car_typePickup | -0.40 | 0.24 | -1.68 | 0.09 |
| car_usePrivate:car_typeSports Car | 0.37 | 0.30 | 1.22 | 0.22 |
| car_usePrivate:car_typeSUV | -0.23 | 0.22 | -1.03 | 0.30 |
| car_usePrivate:car_typeVan | -0.62 | 0.29 | -2.11 | 0.03 |
| travtime:urbanicityUrban | 0.01 | 0.01 | 2.24 | 0.03 |

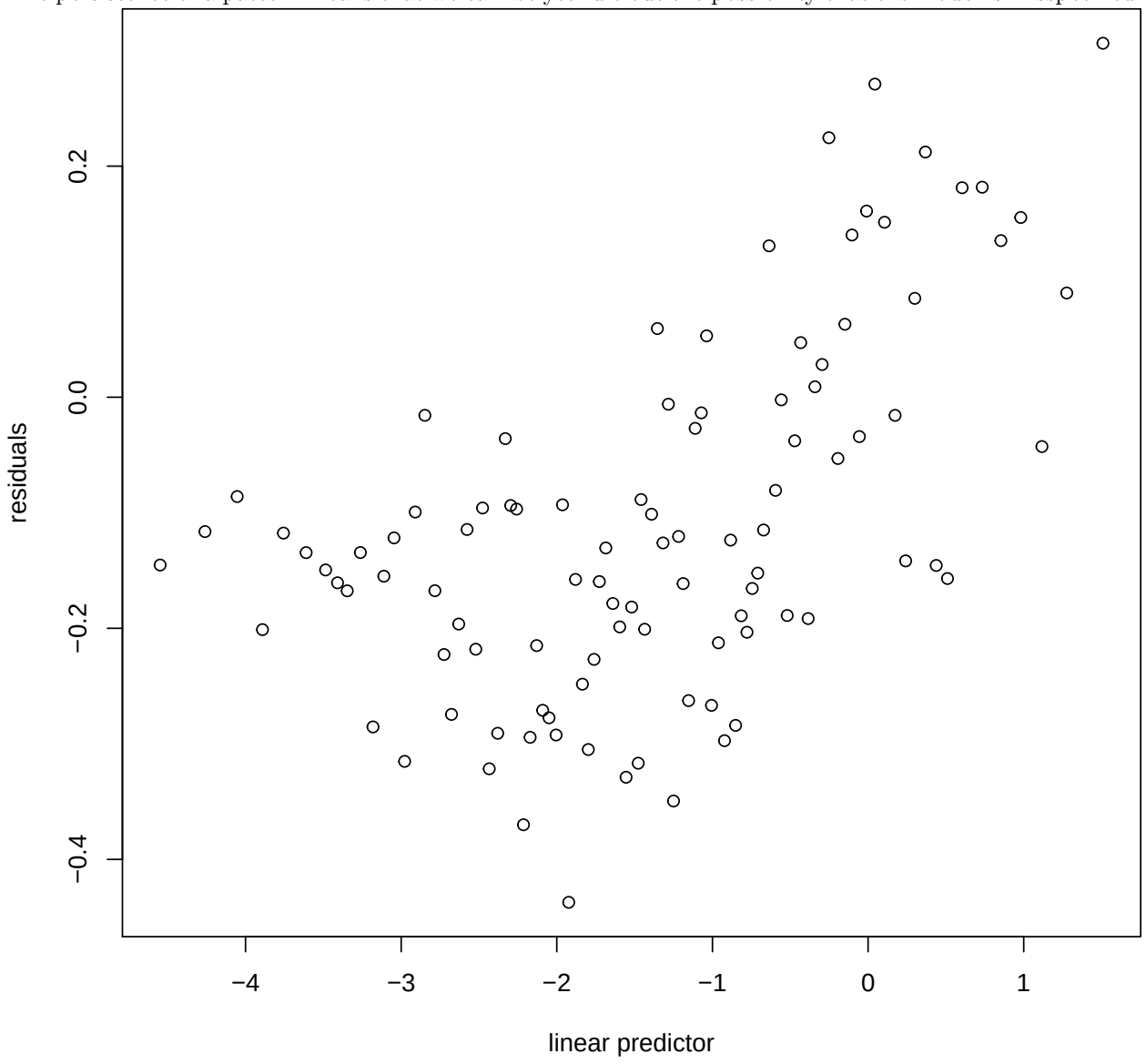
Standard errors: MLE


PROC ROC Curve
Sensitivity
1 − Specificity
0.0
0.2
0.4
0.6
0.8
1.0
AUC: 0.809




We did not detect overdispersion in the model1 based on our assessment of the residual deviance and chi-square test. [1] "We divide the deviance by the residuals to obtain the value 0.907442589140304. There is no overt concern since the values is not greater than 1" [1] "Next we obtain a Chi-Squared test statistic of 0.8729 This communicates that the null hypothesis is not rejected and their are no problems with dispersion."

The persistence of a pattern means that we cannot yet rule out the possibility that the model is misspecified.



## Crash Logistic Classification Model Selection

Model performance is similar across all cases. Model1 had the highest accuracy. Model2 has the lowest AIC.

The models are able to discriminate response classes effectively given a 0.5 evaluation threshold. While the discriminatory performance was similar across our models, we selected model2 for test validation since it had the lowest AIC and highest AUC values

| model | predictors | sensitivity | specificity | pos_rate | neg_rate | precision |
|---|---|---|---|---|---|---|
| Base Model: base variables | 25 | 0.9227696 | 0.4144981 | 0.8148148 | 0.6578171 | 0.814814 |
| transformation Model: reduced variables | 24 | 0.9238793 | 0.4250310 | 0.8177175 | 0.6666667 | 0.817717 |
| Feature_Eng+Transform Model: reduced variables | 30 | 0.9243231 | 0.3983891 | 0.8109424 | 0.6534553 | 0.810942 |

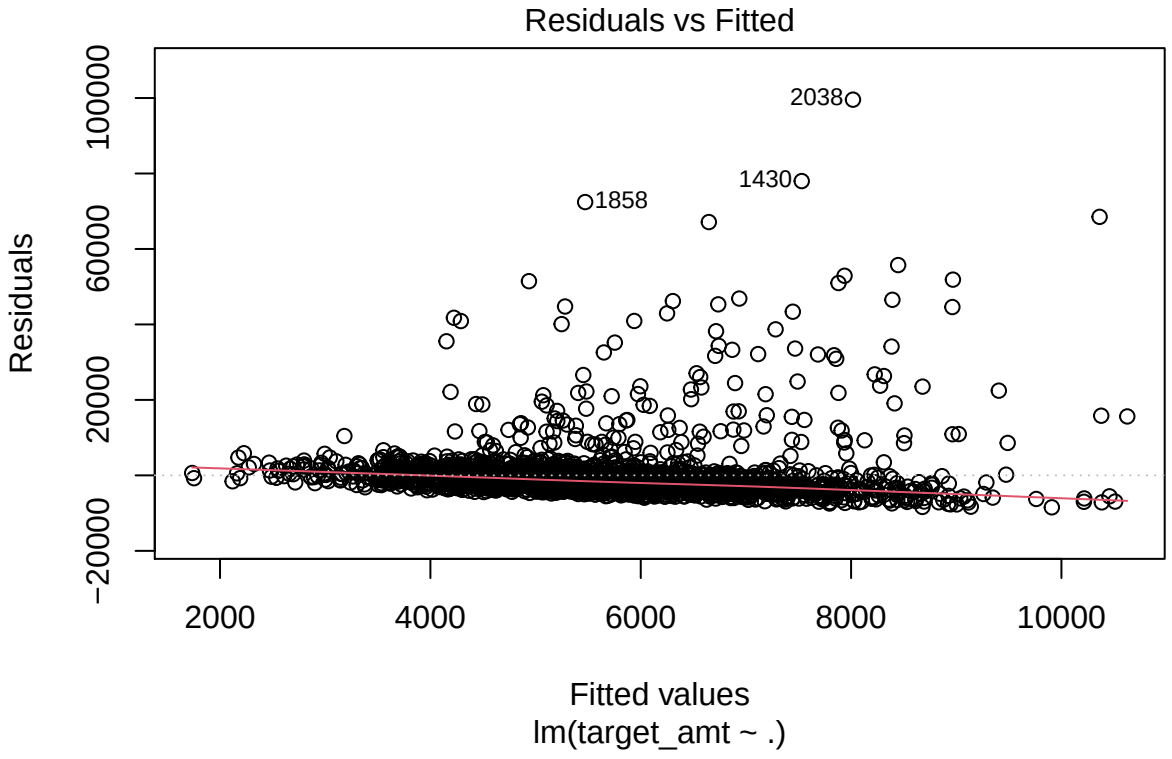
## Construct Linear Regression Model

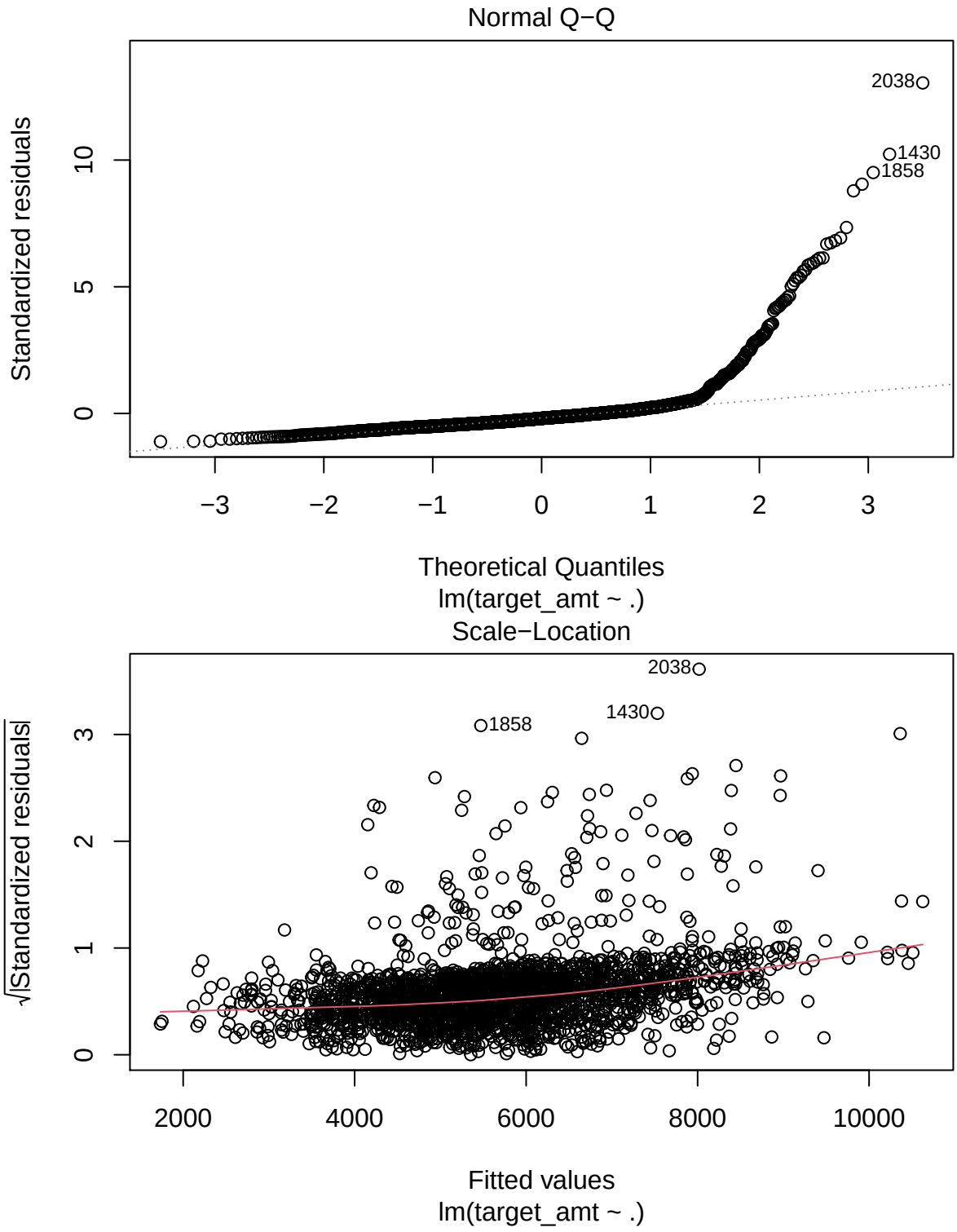
Firstly, we will like to how the saturated model performs under the standard Gaussian assumptions. We find there are only four variables with significant p-values; and the r-squared is very low. Also the residual plots fail the required assumptions regarding the normal distribution and constant variance. We will experiment with the variable selection, but we also need to either transform the response variable or change the link function.

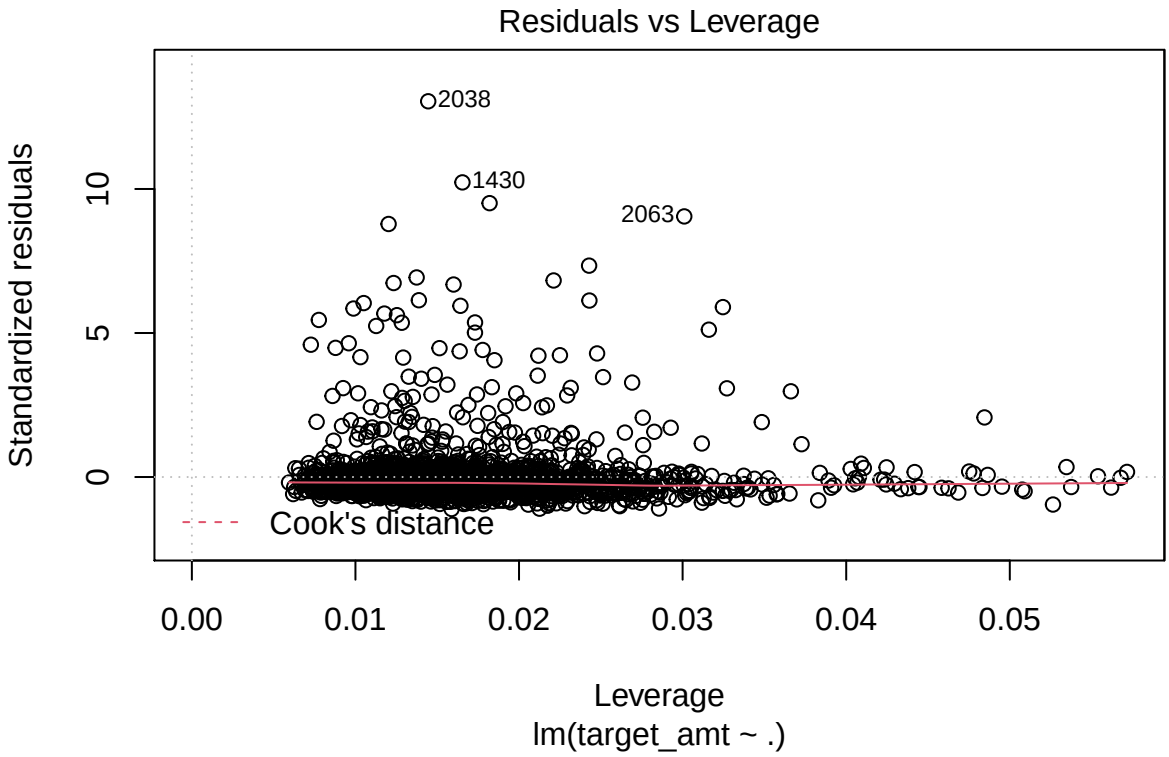
### Cost Model 1: Base Model

| | |
|---|---|
| Observations | 2152 |
| Dependent variable | target_amt |
| Type | OLS linear regression |

| | |
|---|---|
| F(35,2116) | 1.89 |
| $R^2$ | 0.03 |
| Adj. $R^2$ | 0.01 |



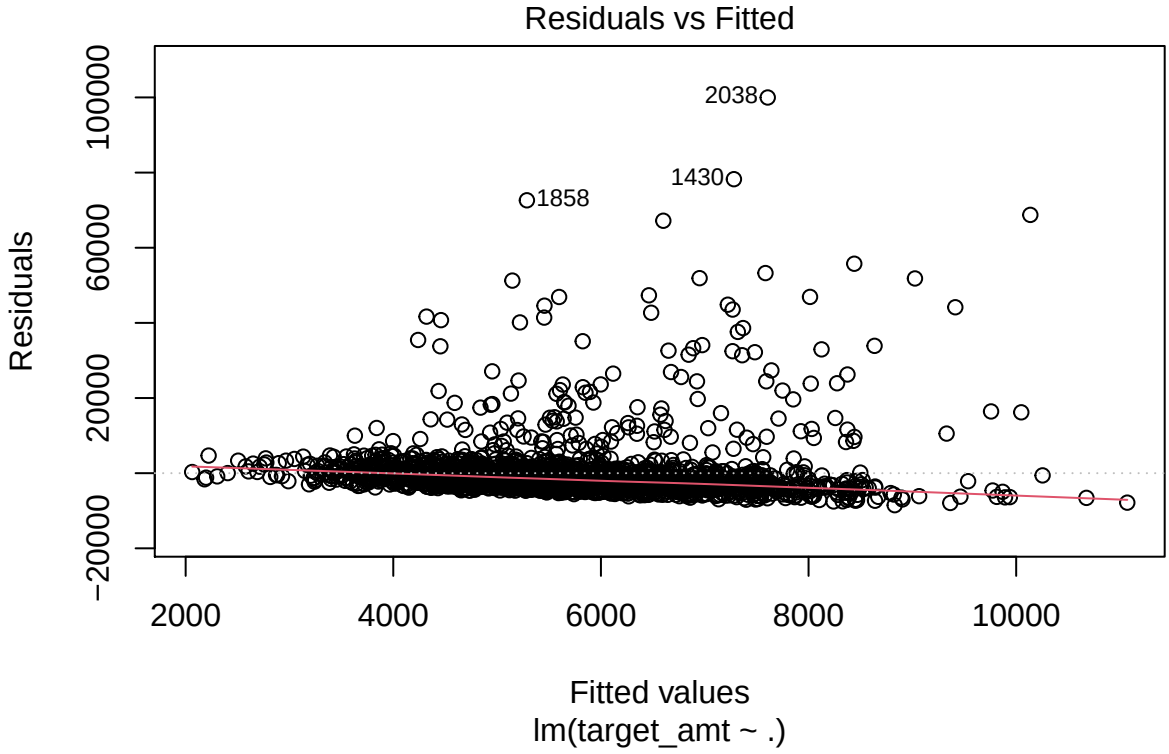

Normal Q−Q
Standardized residuals
2038
1430
1858
Theoretical Quantiles
lm(target_amt ~ .)
Scale−Location
√|Standardized residuals|
2038
1430
1858
Fitted values
lm(target_amt ~ .)

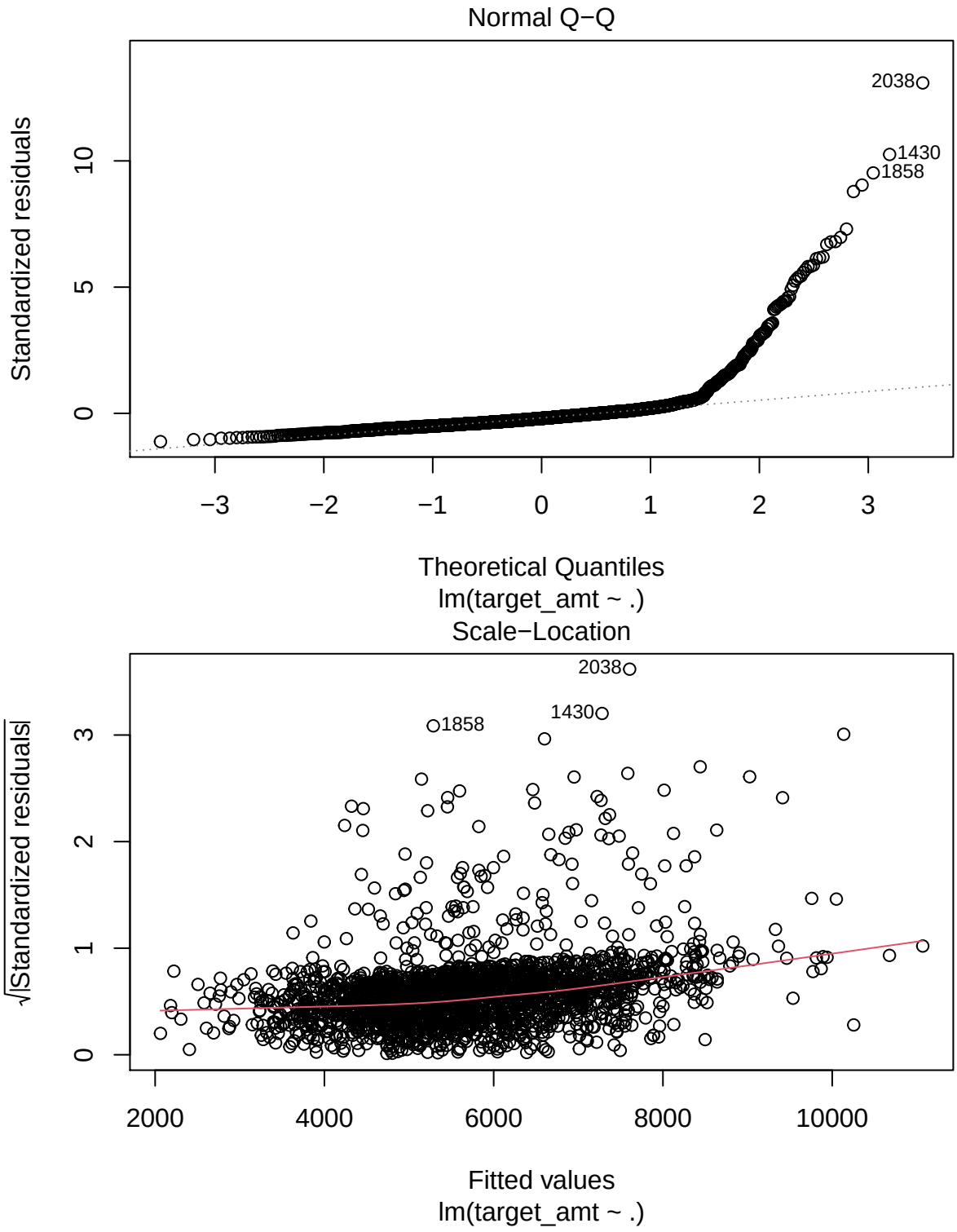

Residuals vs Leverage
2038
1430
2063
Standardized residuals
10
5
0
Cook's distance
0.00
0.01
0.02
0.03
0.04
0.05
Leverage
lm(target_amt ~ .)

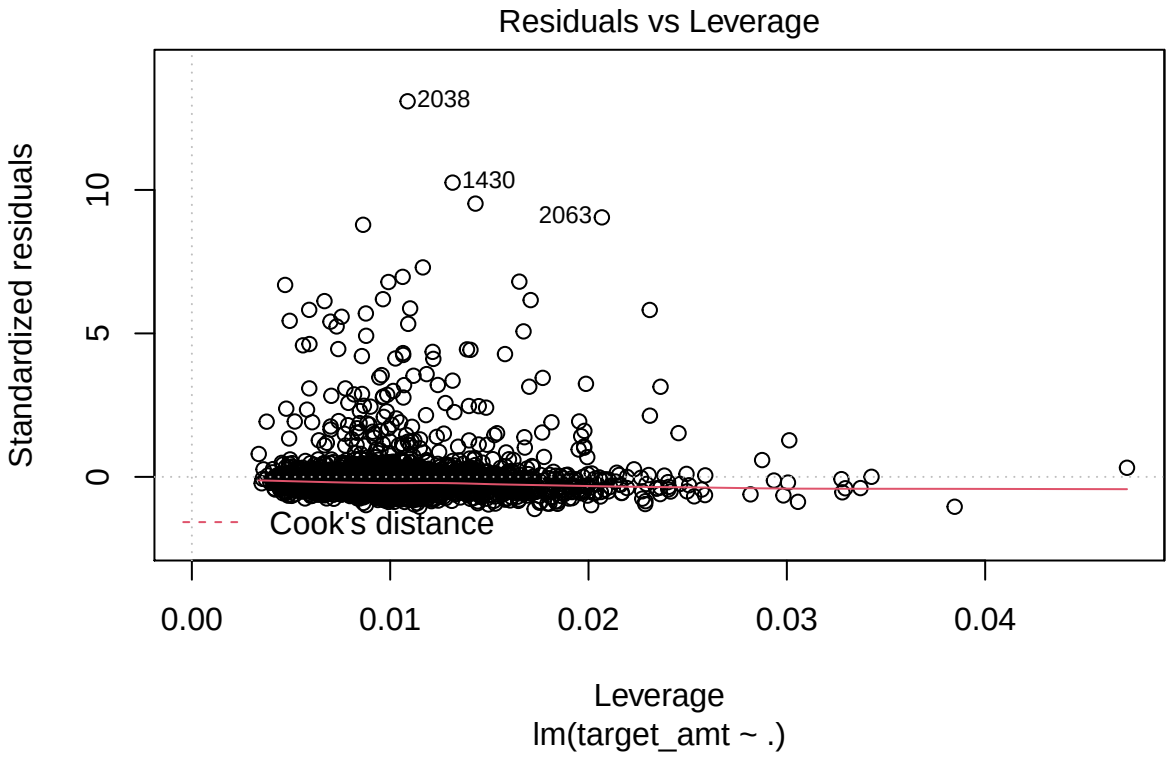
### Cost Model 2: Feature Reduction

Removing the variables below reduces Residual Standard Error:

```
-parent1
-age
-homekids
-kidsdriv
-red_car
-urbanicity
-job
```



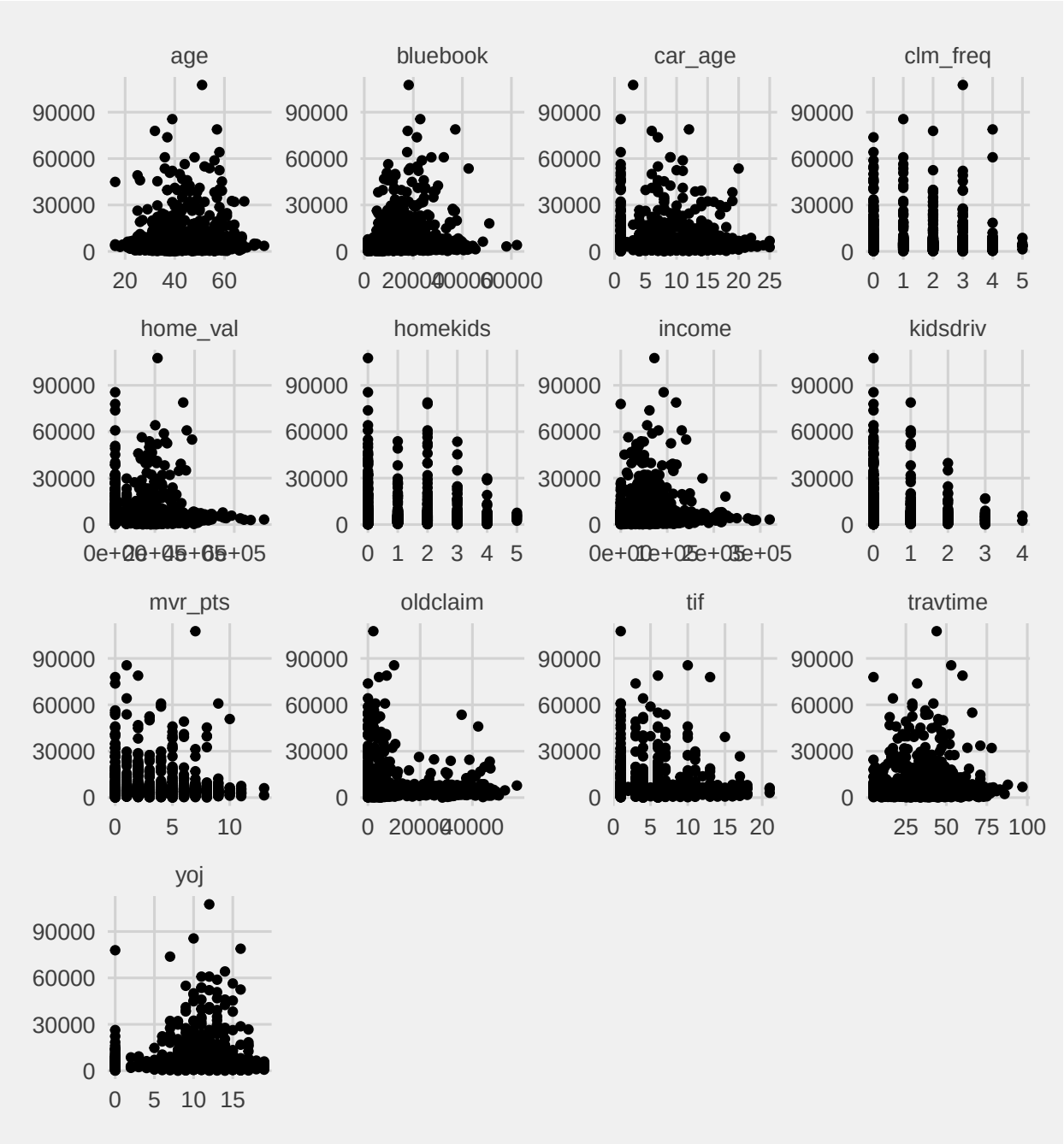

Normal Q−Q
Standardized residuals
2038
1430
1858
Theoretical Quantiles
lm(target_amt ~ .)
Scale−Location
√|Standardized residuals|
2038
1430
1858
Fitted values
lm(target_amt ~ .)


Residuals vs Leverage
2038
1430
2063
Standardized residuals
10
5
0
Cook's distance
0.00
0.01
0.02
0.03
0.04
Leverage
lm(target_amt ~ .)

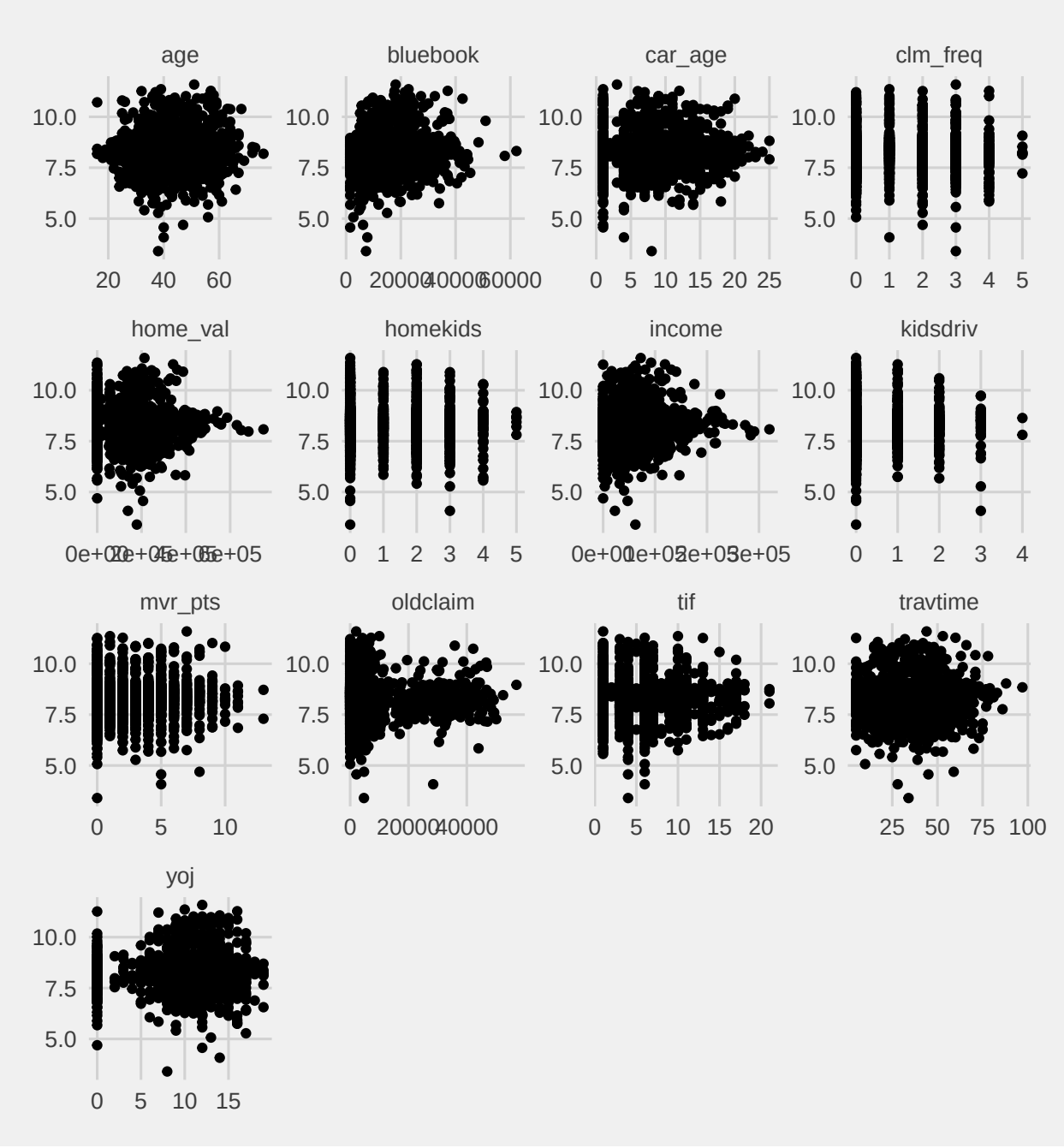
### Cost Model 3: Transformation & Weights

We will attempt to correct the heteroskedasticity of the residual plots through transformations.

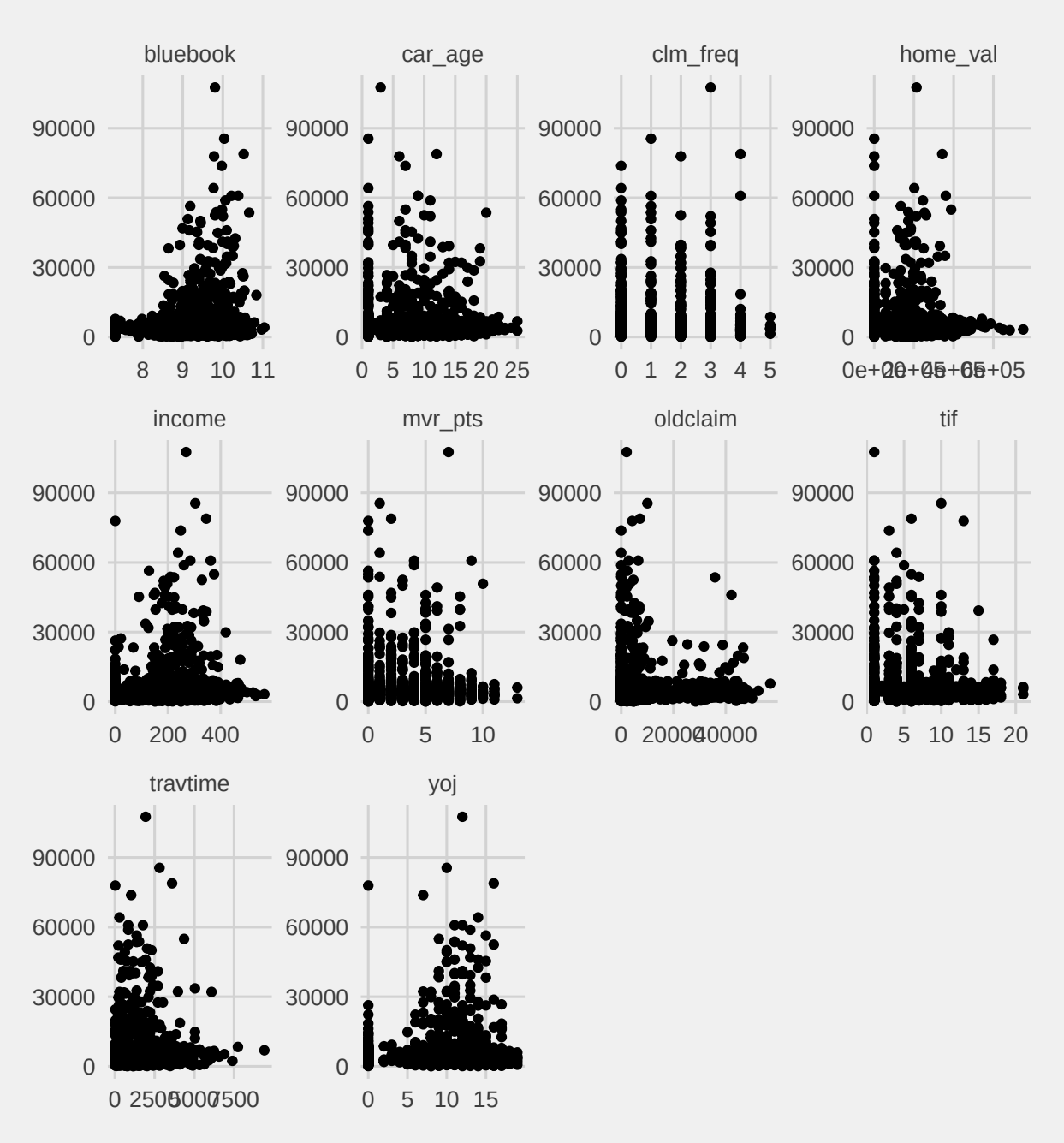
Lets review the linearity as they currently stand from the plots below.



Now with a log transformation on the response. We can see that much of the linearity we noticed above is non-existant.

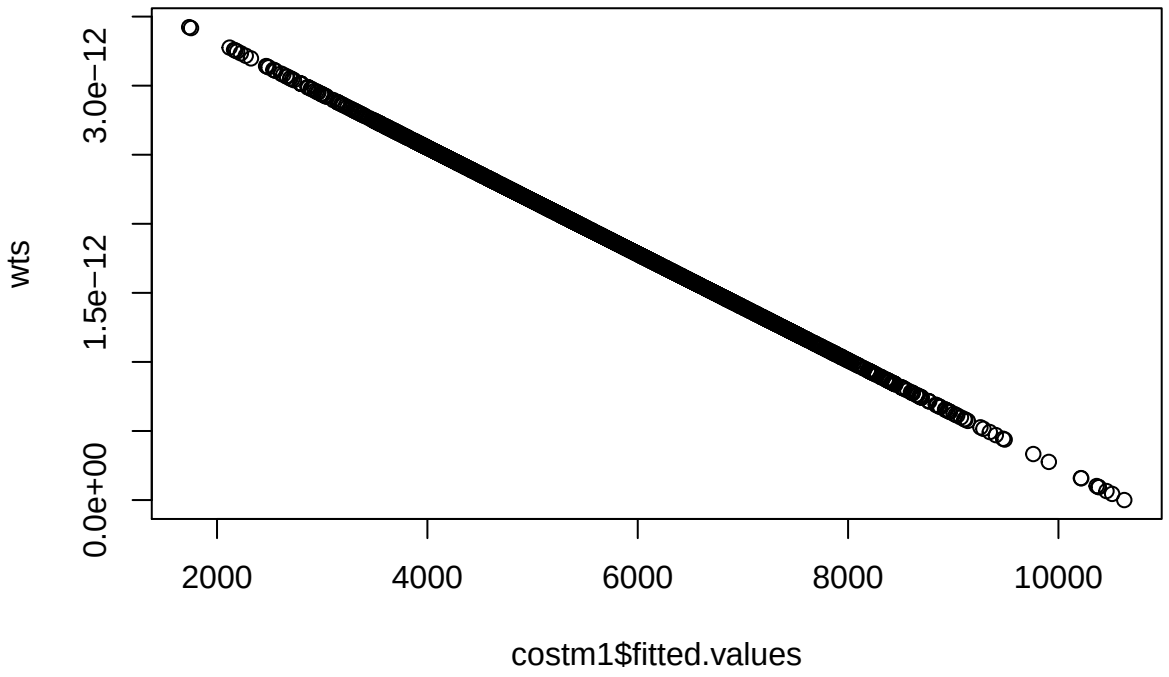


Lastly, lets review with the below predictor transformations instead. `-sqrt(income) -log(bluebook) -travtime^2`

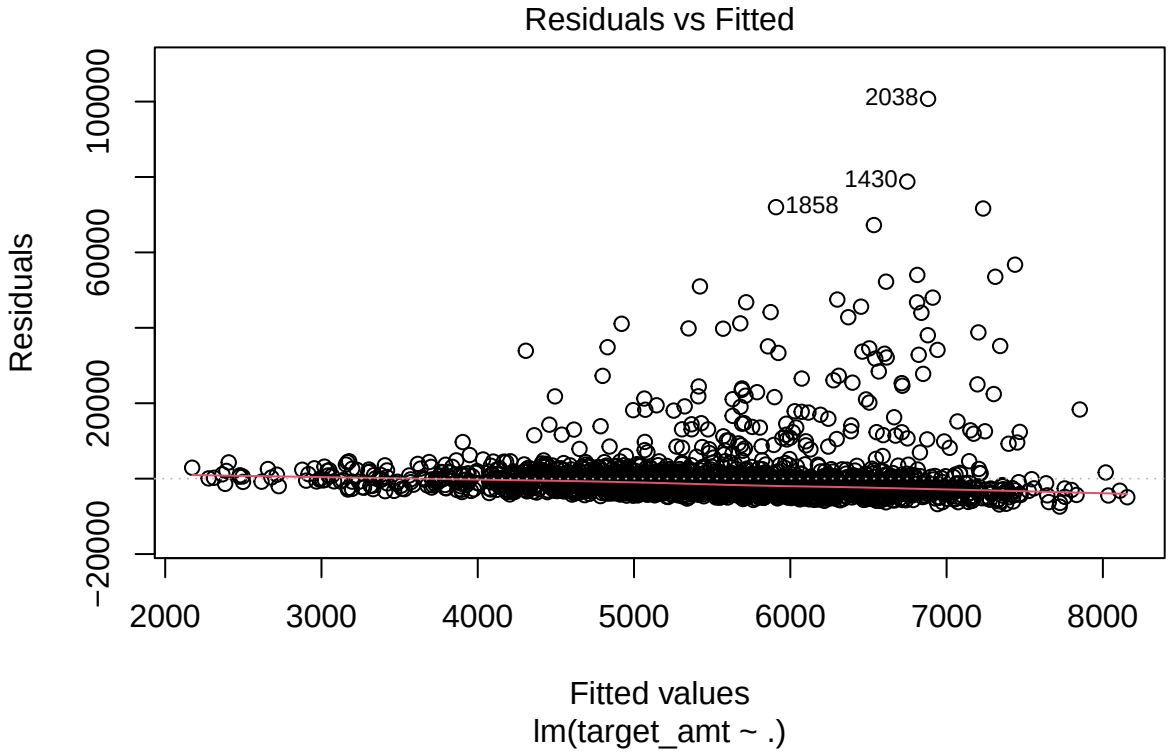


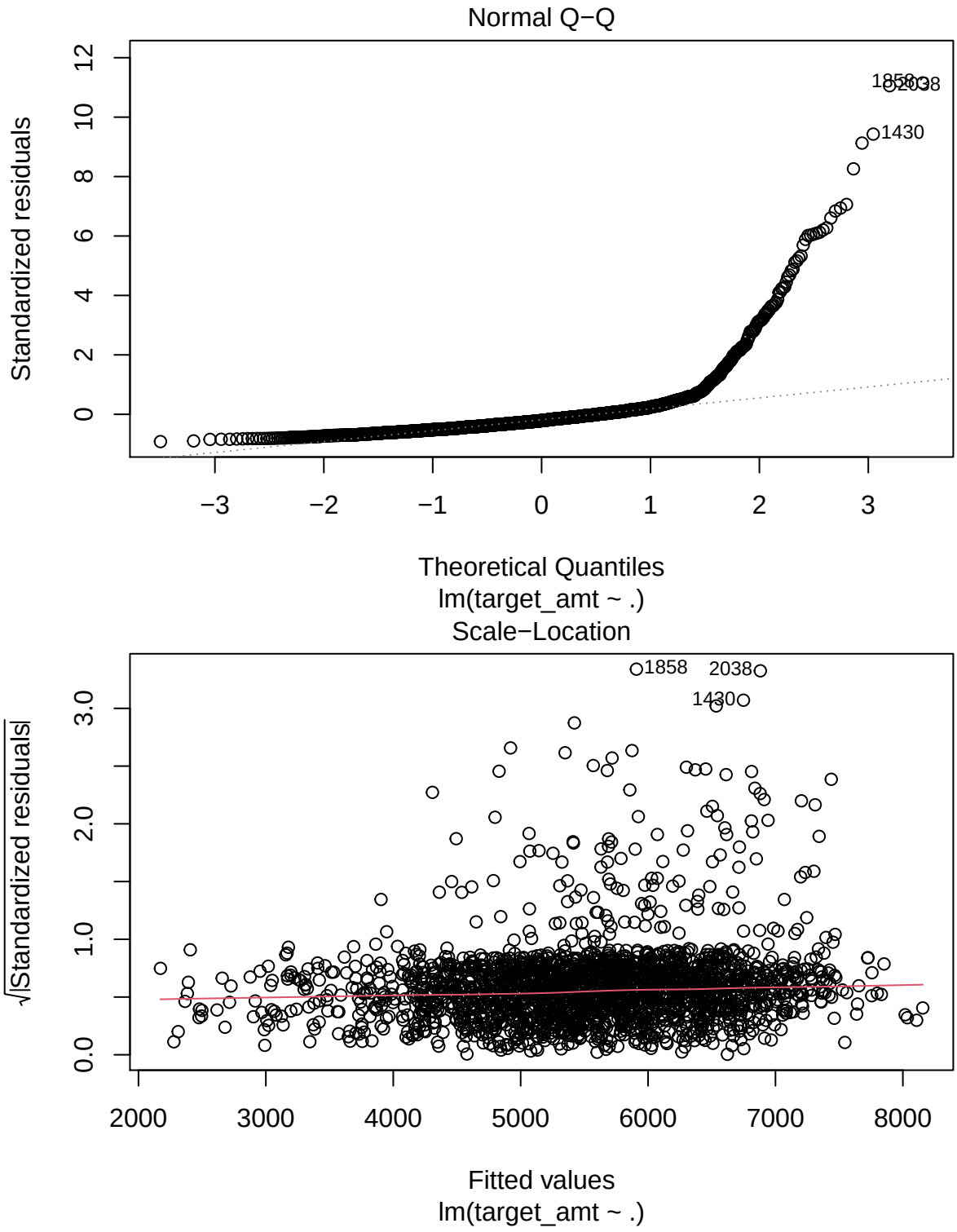
Now that we have an impression of the linearity; we can move onto selecting the weights. Our strategy is to use the results from our base model, regressing the residuals against its fitted values. We end up with a distribution of values which loosely represent the variance.

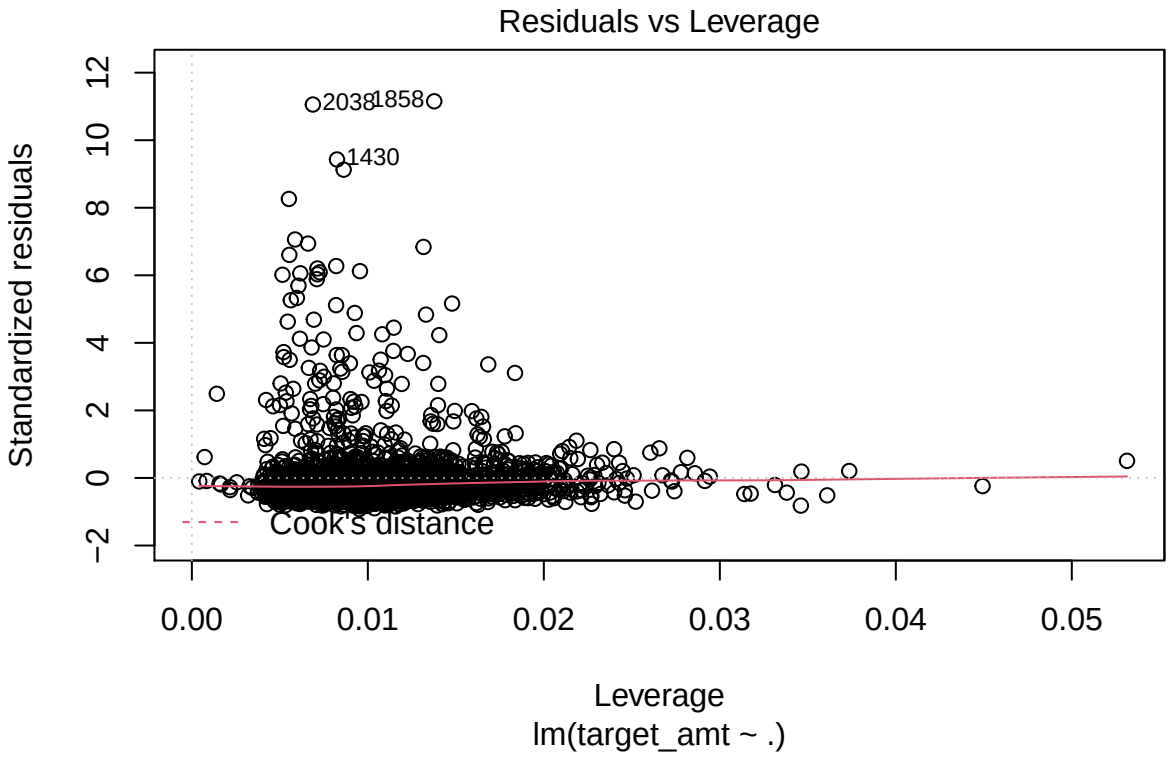
Below is a plot of how the weights will be applied along the scale of the response variable.




The below weighted model has the lowest residual standard error but the $R^2$ is still very low.




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(target_amt ~ .)
Scale−Location
√|Standardized residuals|
Fitted values
lm(target_amt ~ .)
1858
2038
1430


Residuals vs Leverage
Standardized residuals
Leverage
lm(target_amt ~ .)
Cook's distance
2038
1858
1430
−2
0
2
4
6
8
10
12
0.00
0.01
0.02
0.03
0.04
0.05

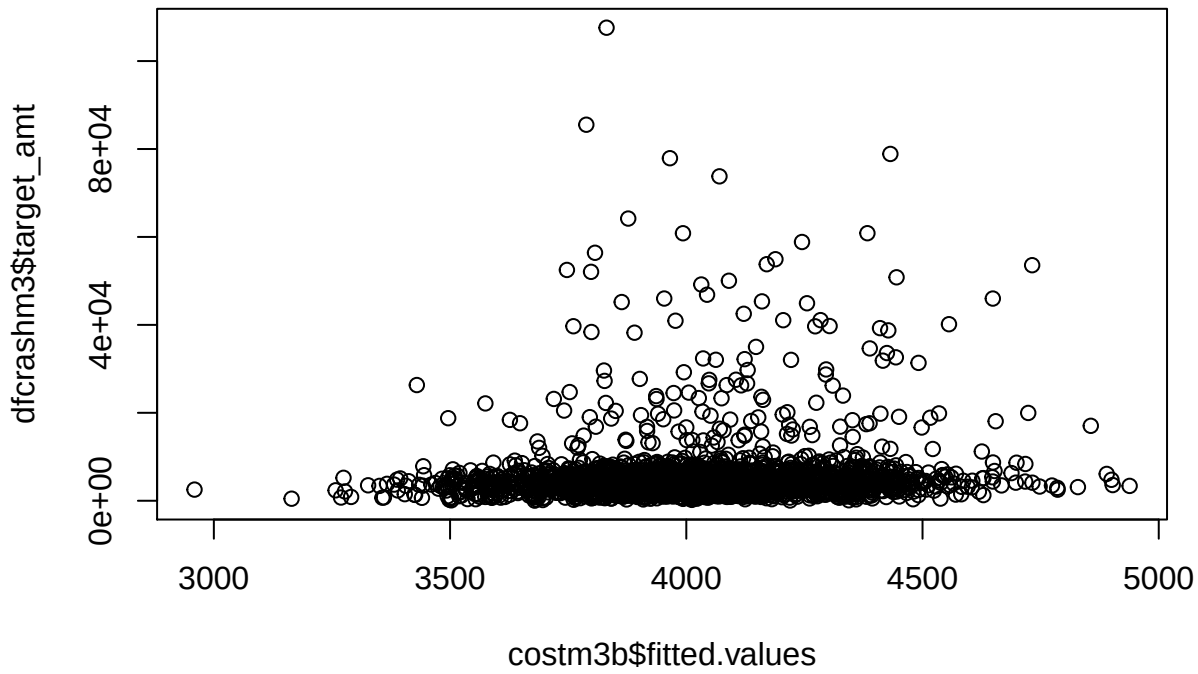
## Cost Model 4 Robust Linear Regression

As an alternative approach to weighting the model we implement a Robust Linear Regression. The function uses Iteratively Reweighted Least Squares (IRLS) to obtain a maximum likelihood estimate of the parameters. The residual standard error is reduced again.

Call: rlm(formula = target_amt ~ ., data = dfcrashm3, weights = wts, method = "MM") Residuals: Min 1Q Median -0.0064315 -0.0018675 0.0001238 3Q Max 0.0024650 0.1041708

Coefficients: Value
(Intercept) 1964.5401 mstatusY -135.7274 sexM -3.1481 educationBachelors -280.4810 educationMasters -22.4394 educationPhD 65.2545 travtime -0.0002 car_usePrivate 10.1784 bluebook 220.7171 tif 5.6431 car_typePanel Truck 207.4707 car_typePickup 115.0655 car_typeSports Car 38.5263 car_typeSUV 24.3042 car_typeVan 28.9382 oldclaim 0.0040 clm_freq -79.1315 revokedY 32.5710 mvr_pts 54.7649 car_age 8.1444 home_val 0.0006 yoj -0.3504 income -0.7760 Std. Error (Intercept) 888.4787 mstatusY 123.9422 sexM 166.0591 educationBachelors 141.0430 educationMasters 215.4034 educationPhD 295.3290 travtime 0.0428 car_usePrivate 118.5274 bluebook 90.9390 tif 12.2343 car_typePanel Truck 274.8497 car_typePickup 167.0316 car_typeSports Car 210.4564 car_typeSUV 183.2774 car_typeVan 236.2534 oldclaim 0.0064 clm_freq 45.5167 revokedY 141.7032 mvr_pts 20.3858 car_age 13.5440 home_val 0.0006 yoj 14.1460 income 0.7924 t value
(Intercept) 2.2111 mstatusY -1.0951 sexM -0.0190 educationBachelors -1.9886 educationMasters -0.1042 educationPhD 0.2210 travtime -0.0040 car_usePrivate 0.0859 bluebook 2.4271 tif 0.4613 car_typePanel Truck 0.7549 car_typePickup 0.6889 car_typeSports Car 0.1831 car_typeSUV 0.1326 car_typeVan 0.1225 oldclaim 0.6211 clm_freq -1.7385 revokedY 0.2299 mvr_pts 2.6864 car_age 0.6013 home_val 0.9309 yoj -0.0248 income -0.9793

Residual standard error: 0.003125 on 2129 degrees of freedom

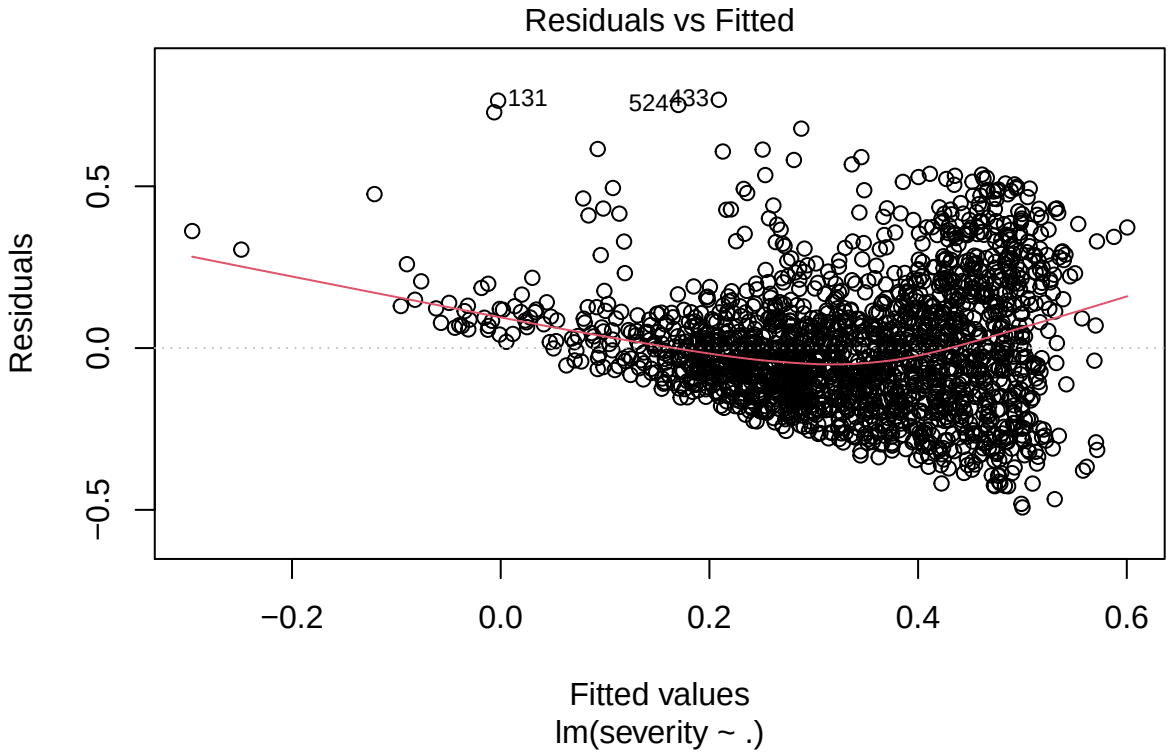


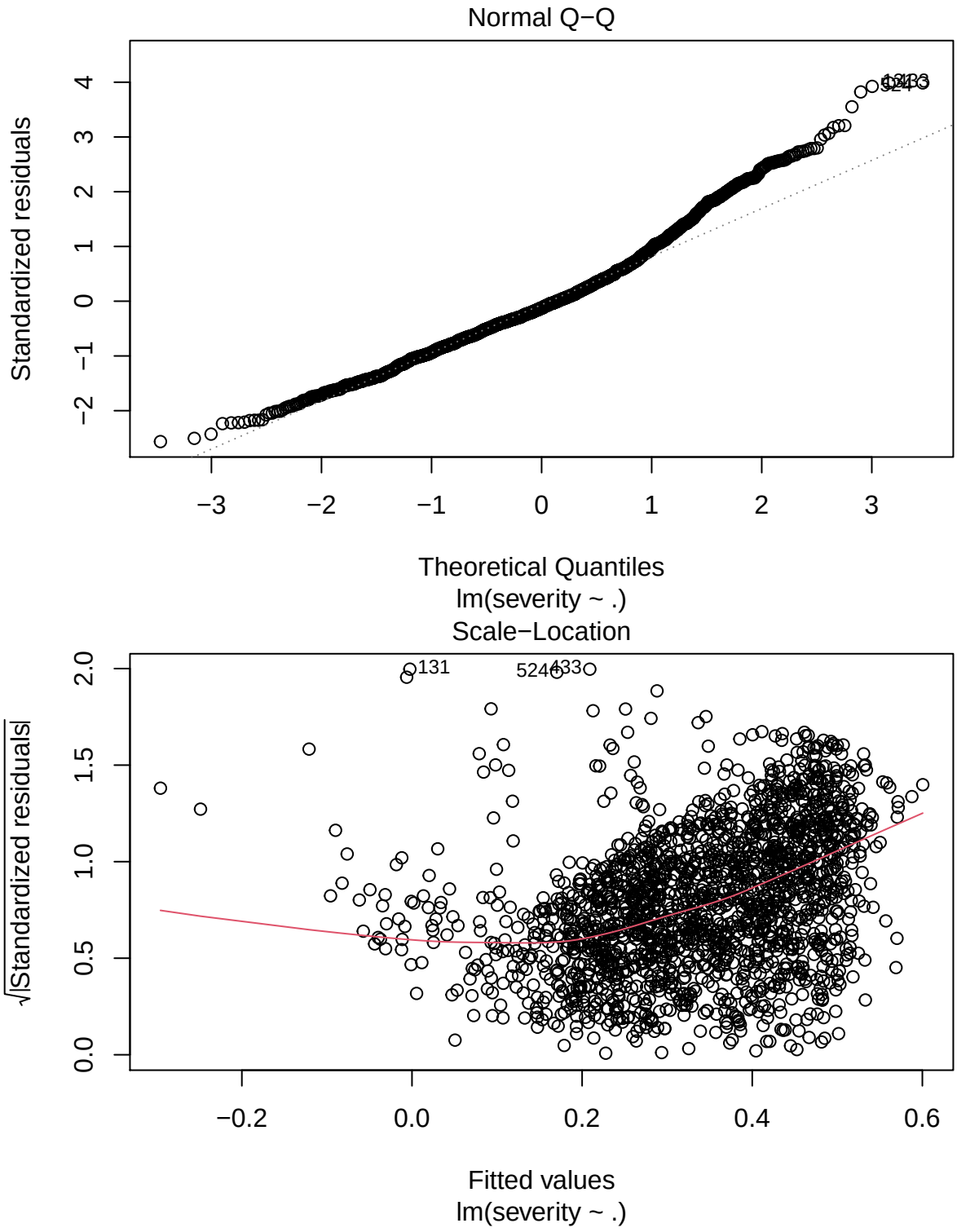
### Cost Model 5 Target Interaction Term

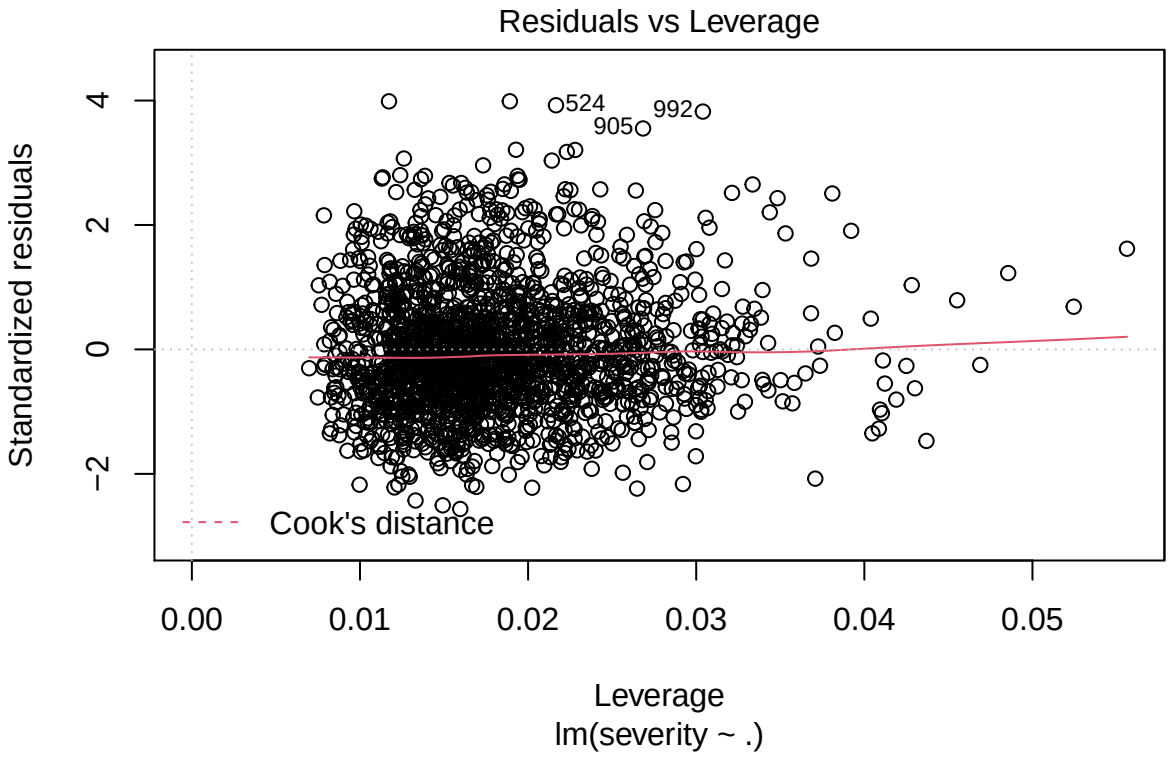
Although there has been some improvement across the above models, $R^2$ is still very low. We now move to rethink the target variable. It stands to reason that the cost of a crash is mostly a function of the value of the car, and the p-values from the above models support this hypothesis. Rather than regressing on the cost, which renders most predictors useless, we model the severity of the accident. We can represent severity as `cost/bluebook`.

We will drop the ratios that are >1 assuming these involved incidental bodily harm.

Now lets take another look at the relationships.




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(severity ~ .)
Scale−Location
√|Standardized residuals|
131
524
433
Fitted values
lm(severity ~ .)




To apply the beta distribution we use a general additive model(GAM); which gives us the option of adding non-linear variables. After some find tuning; we find that adding the 'log(bluebook)' and income as non-linear predictors improves our models. We applied an anova test to confirm.

Family: Beta regression(4.951) Link function: logit

Formula: severity ~ mstatus + sex + mstatus + sex + education + car_use + s(log(bluebook)) + tif + car_type + oldclaim + clm_freq + revoked + mvr_pts + home_val + yoj + s(income) + travtime + income + car_age + oldclaim + liquidity + bluebook

Estimated degrees of freedom: 4.41 3.31 total = 33.72

REML score: -560.4059 rank: 44/45

Family: Beta regression(4.951) Link function: logit

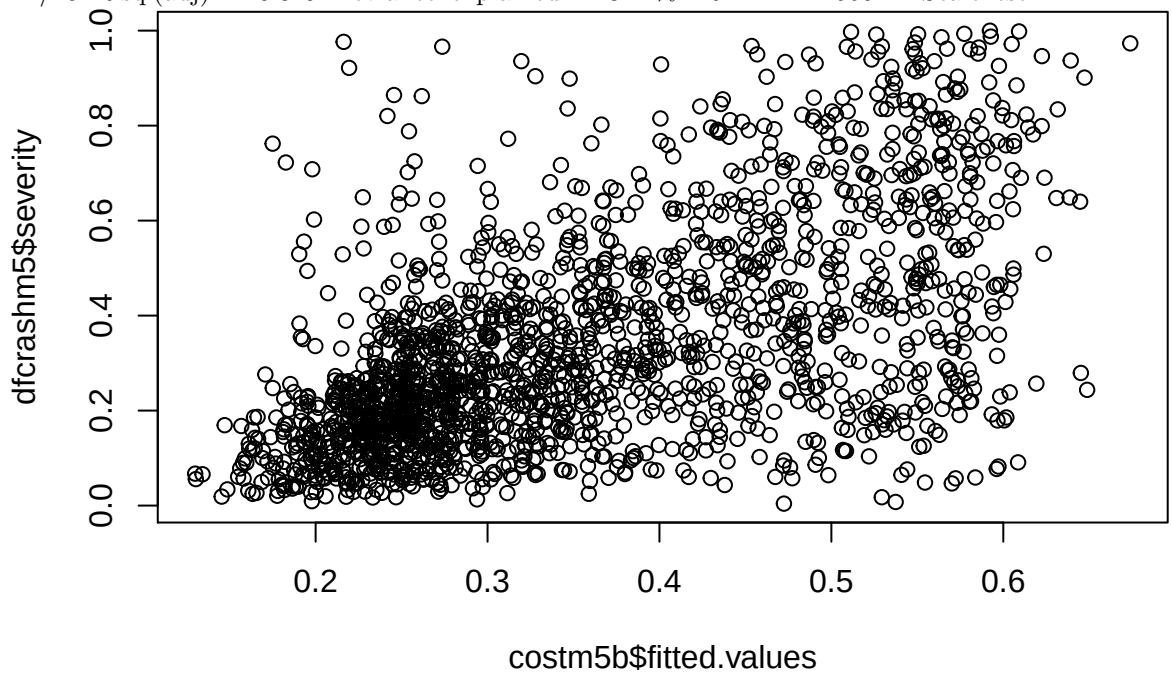
Formula: severity ~ mstatus + sex + mstatus + sex + education + car_use + s(log(bluebook)) + tif + car_type + oldclaim + clm_freq + revoked + mvr_pts + home_val + yoj + s(income) + travtime + income + car_age + oldclaim + liquidity + bluebook

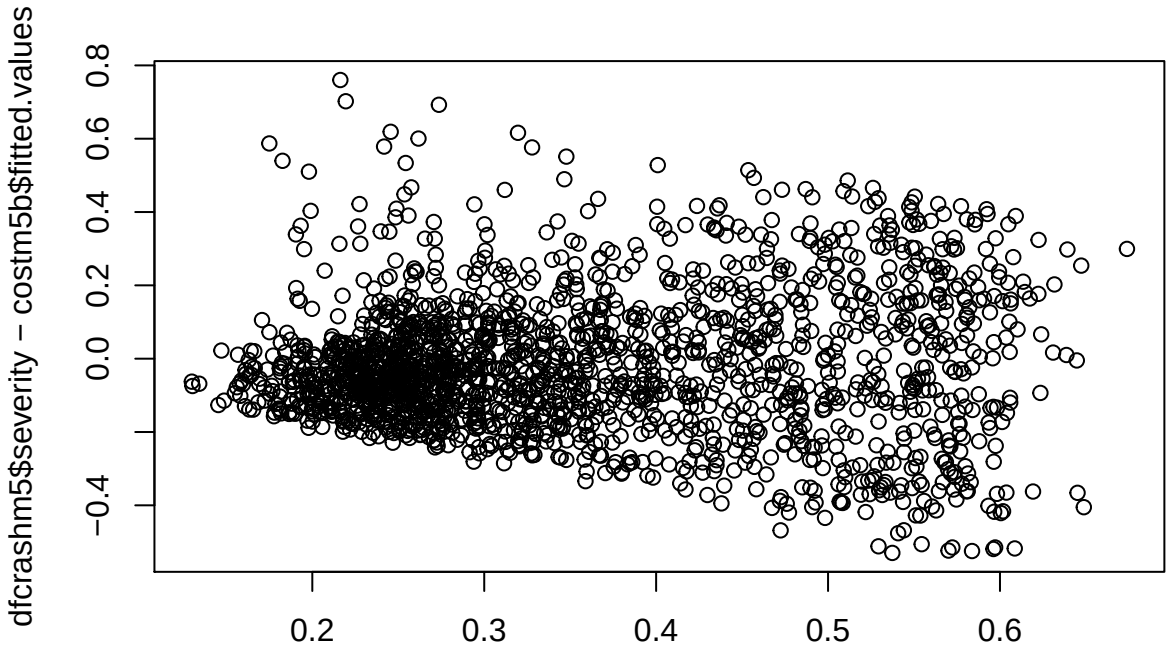
Parametric coefficients: Estimate (Intercept) 0.000e+00 mstatusY -8.350e-02 sexM 3.344e-02 educationBachelors -2.642e-02 educationMasters 4.952e-02 educationPhD 1.590e-01 car_usePrivate 5.404e-02 tifmoderate -1.005e-03 tifhigh 4.776e-02 car_typePanel Truck 1.392e-01 car_typePickup 3.250e-02 car_typeSports Car 4.992e-02 car_typeSUV 6.188e-02 car_typeVan 1.355e-02 oldclaim 3.676e-06 clm_freqmoderate -5.273e-02 clm_freqhigh -2.403e-01 revokedY -3.633e-02 mvr_ptslow 5.923e-02 mvr_ptshigh 1.068e-01 home_val -5.717e-07 yoj 5.216e-04 travtime -1.694e-06 income -3.851e-03 car_age 2.698e-03 liquidityhigh 1.547e-01 bluebook 5.379e-07 Std. Error z value (Intercept) 0.000e+00 NA mstatusY 5.242e-02 -1.593 sexM 6.899e-02 0.485 educationBachelors 5.805e-02 -0.455 educationMasters 8.610e-02 0.575 educationPhD 1.199e-01 1.325 car_usePrivate 4.853e-02 1.114 tifmoderate 4.290e-02 -0.023 tifhigh 7.119e-02 0.671 car_typePanel Truck 1.173e-01 1.187 car_typePickup 7.198e-02 0.452 car_typeSports Car 8.845e-02 0.564 car_typeSUV 7.802e-02 0.793 car_typeVan 9.163e-02 0.148 oldclaim 2.888e-06 1.273 clm_freqmoderate 5.260e-02 -1.002 clm_freqhigh 1.076e-01 -2.234 revokedY 6.144e-02 -0.591 mvr_ptslow 4.738e-02 1.250 mvr_ptshigh 5.287e-02 2.021 home_val 4.082e-07 -1.401 yoj 6.930e-03 0.075 travtime 1.720e-05 -0.099 income 2.973e-03 -1.295 car_age 5.341e-03 0.505 liquidityhigh 8.779e-02 1.762 bluebook 3.907e-05 0.014 Pr(>|z|)

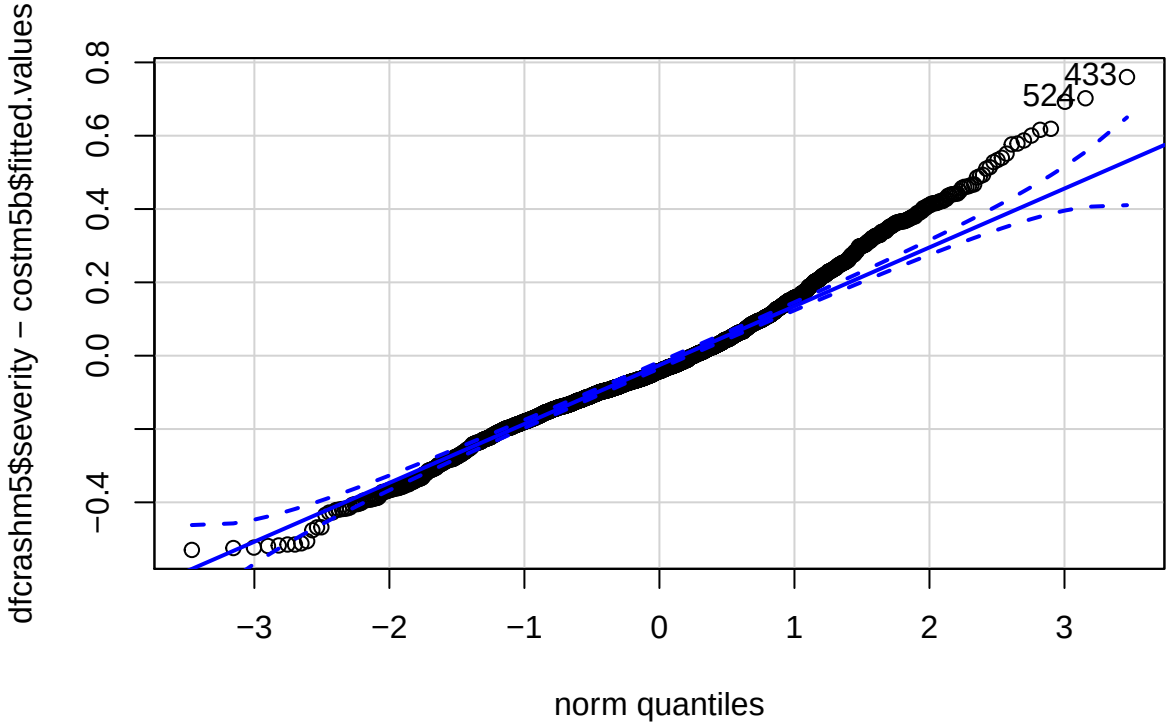
(Intercept) NA
mstatusY 0.1112
sexM 0.6279
educationBachelors 0.6490
educationMasters 0.5652
educationPhD 0.1851
car_usePrivate 0.2655
tifmoderate 0.9813
tifhigh 0.5023
car_typePanel Truck 0.2353
car_typePickup 0.6516
car_typeSports Car 0.5725
car_typeSUV 0.4277
car_typeVan 0.8825
oldclaim 0.2030
clm_freqmoderate 0.3161
clm_freqhigh 0.0255 *revokedY 0.5543*
*mvr_ptslow 0.2112*
*mvr_ptshigh 0.0433* home_val 0.1613
yoj 0.9400
travtime 0.9215
income 0.1953
car_age 0.6134
liquidityhigh 0.0781 . bluebook 0.9890
— Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms: edf Ref.df Chi.sq s(log(bluebook)) 4.411 5.537 86.945 s(income) 3.309 4.201 8.095 p-value
s(log(bluebook)) <2e-16 *** s(income) 0.0978 .
— Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Rank: 44/45 R-sq.(adj) = 0.329 Deviance explained = 34.4% -REML = -560.41 Scale est. = 1 n = 1873






[1]

433 524

| Resid. Df | Resid. Dev | Df | Deviance | Pr(>Chi) |
|---|---|---|---|---|
| 1,847.000 | -1,247.423 | | | |
| 1,835.864 | -1,348.361 | 11.13582 | 100.938 | 0.00000000000000001363278 |

# Conclusion

The above analysis uses data on a collection of drivers to predict who will be involved in a crash, and the cost implications given said crash.
We approached the former using logistic regression to assign a probability of crash; and letting a designated threshold bucket the observations into crash vs no-crash. In reviewing the distribtion of the variables and their relationship to the target, we opted to transform a selection of the variables into factors based on the grouping behavior. Through additional transformations and feature engineering, we landed on a model that minimized AUC and AIC. After building the model to bucket the observations into the binary groups, we moved onto constructing second model to predict the cost implications. We were able to make slight improvements through feature selections and tranformations, but the low $R^2$ warrented additional experimenting. Prompted by the distribution of the residuals; we applied WLS, Robust Regression, and IWLS. The jump in $R^2$ was encouraging. However, it appeared that 'bluebook' variable was carrying all of the models. The decision was made to rethink the target variable as a ratio 'bluebook' and regression on a [0,1] severity using the beta function.

| | Est. | S.E. | t val. | p |
|---|---|---|---|---|
| (Intercept) | 3033.35 | 1578.49 | 1.92 | 0.05 |
| kidsdriv | -166.24 | 315.92 | -0.53 | 0.60 |
| homekids | 210.31 | 207.33 | 1.01 | 0.31 |
| parent1Y | 250.54 | 587.57 | 0.43 | 0.67 |
| mstatusY | -866.52 | 506.86 | -1.71 | 0.09 |
| sexM | 1386.05 | 656.58 | 2.11 | 0.03 |
| educationBachelors | 624.64 | 503.43 | 1.24 | 0.21 |
| educationMasters | 1229.57 | 884.69 | 1.39 | 0.16 |
| educationPhD | 2712.55 | 1148.39 | 2.36 | 0.02 |
| travtime | 0.11 | 11.07 | 0.01 | 0.99 |
| car_usePrivate | -451.13 | 490.36 | -0.92 | 0.36 |
| bluebook | 0.13 | 0.03 | 4.11 | 0.00 |
| tif | -15.62 | 42.51 | -0.37 | 0.71 |
| car_typePanel Truck | -479.65 | 947.40 | -0.51 | 0.61 |
| car_typePickup | -33.04 | 593.34 | -0.06 | 0.96 |
| car_typeSports Car | 1027.20 | 749.34 | 1.37 | 0.17 |
| car_typeSUV | 886.21 | 666.45 | 1.33 | 0.18 |
| car_typeVan | 116.83 | 764.04 | 0.15 | 0.88 |
| red_carY | -169.69 | 496.72 | -0.34 | 0.73 |
| oldclaim | 0.03 | 0.02 | 1.13 | 0.26 |
| clm_freq | -112.68 | 158.03 | -0.71 | 0.48 |
| revokedY | -1138.56 | 516.34 | -2.21 | 0.03 |
| mvr_pts | 110.59 | 68.43 | 1.62 | 0.11 |
| urbanicityUrban | 103.64 | 755.98 | 0.14 | 0.89 |
| car_age | -97.94 | 45.32 | -2.16 | 0.03 |
| home_val | 0.00 | 0.00 | 1.07 | 0.28 |
| yoj | 30.80 | 49.13 | 0.63 | 0.53 |
| income | -0.01 | 0.01 | -1.87 | 0.06 |
| age | 17.31 | 21.24 | 0.81 | 0.42 |
| jobClerical | -215.69 | 581.04 | -0.37 | 0.71 |
| jobDoctor | -1724.93 | 1728.43 | -1.00 | 0.32 |
| jobHome Maker | -560.48 | 865.76 | -0.65 | 0.52 |
| jobLawyer | 453.40 | 1020.92 | 0.44 | 0.66 |
| jobManager | -931.80 | 799.38 | -1.17 | 0.24 |
| jobProfessional | 552.92 | 644.27 | 0.86 | 0.39 |
| jobStudent | -472.80 | 715.09 | -0.66 | 0.51 |

Standard errors: OLS

| | |
|---|---|
| Observations | 2152 |
| Dependent variable | target_amt |
| Type | OLS linear regression |

| | |
|---|---|
| F(22,2129) | 2.64 |
| $R^2$ | 0.03 |
| Adj. $R^2$ | 0.02 |

| | Est. | S.E. | t val. | p |
|---|---|---|---|---|
| (Intercept) | 3571.48 | 1038.08 | 3.44 | 0.00 |
| mstatusY | -938.87 | 417.27 | -2.25 | 0.02 |
| sexM | 1261.87 | 581.68 | 2.17 | 0.03 |
| educationBachelors | 713.39 | 478.44 | 1.49 | 0.14 |
| educationMasters | 1310.45 | 714.76 | 1.83 | 0.07 |
| educationPhD | 2060.06 | 990.70 | 2.08 | 0.04 |
| travtime | 0.86 | 10.99 | 0.08 | 0.94 |
| car_usePrivate | -447.45 | 410.59 | -1.09 | 0.28 |
| bluebook | 0.13 | 0.03 | 4.27 | 0.00 |
| tif | -12.42 | 42.34 | -0.29 | 0.77 |
| car_typePanel Truck | -574.10 | 914.66 | -0.63 | 0.53 |
| car_typePickup | -99.23 | 583.77 | -0.17 | 0.87 |
| car_typeSports Car | 1003.69 | 741.29 | 1.35 | 0.18 |
| car_typeSUV | 858.01 | 657.72 | 1.30 | 0.19 |
| car_typeVan | 55.54 | 753.32 | 0.07 | 0.94 |
| oldclaim | 0.02 | 0.02 | 1.00 | 0.32 |
| clm_freq | -115.53 | 156.64 | -0.74 | 0.46 |
| revokedY | -1019.23 | 511.52 | -1.99 | 0.05 |
| mvr_pts | 122.81 | 67.93 | 1.81 | 0.07 |
| car_age | -97.69 | 45.18 | -2.16 | 0.03 |
| home_val | 0.00 | 0.00 | 1.18 | 0.24 |
| yoj | 57.88 | 42.24 | 1.37 | 0.17 |
| income | -0.01 | 0.01 | -1.91 | 0.06 |

Standard errors: OLS

| | |
|---|---|
| Observations | 2152 |
| Dependent variable | target_amt |
| Type | OLS linear regression |

| | |
|---|---|
| F(22,2128) | 2.00 |
| $R^2$ | 0.02 |
| Adj. $R^2$ | 0.01 |

| | Est. | S.E. | t val. | p |
|---|---|---|---|---|
| (Intercept) | -6027.79 | 2644.87 | -2.28 | 0.02 |
| mstatusY | -376.96 | 368.96 | -1.02 | 0.31 |
| sexM | 797.65 | 494.33 | 1.61 | 0.11 |
| educationBachelors | 172.57 | 419.86 | 0.41 | 0.68 |
| educationMasters | 521.31 | 641.22 | 0.81 | 0.42 |
| educationPhD | 850.51 | 879.15 | 0.97 | 0.33 |
| travtime | -0.07 | 0.13 | -0.52 | 0.61 |
| car_usePrivate | -239.51 | 352.84 | -0.68 | 0.50 |
| bluebook | 1249.46 | 270.71 | 4.62 | 0.00 |
| tif | -4.85 | 36.42 | -0.13 | 0.89 |
| car_typePanel Truck | -331.96 | 818.19 | -0.41 | 0.68 |
| car_typePickup | -191.73 | 497.23 | -0.39 | 0.70 |
| car_typeSports Car | 675.67 | 626.50 | 1.08 | 0.28 |
| car_typeSUV | 422.65 | 545.59 | 0.77 | 0.44 |
| car_typeVan | 71.65 | 703.29 | 0.10 | 0.92 |
| oldclaim | 0.02 | 0.02 | 1.24 | 0.21 |
| clm_freq | -181.05 | 135.50 | -1.34 | 0.18 |
| revokedY | -935.46 | 421.83 | -2.22 | 0.03 |
| mvr_pts | 105.58 | 60.69 | 1.74 | 0.08 |
| car_age | -63.15 | 40.32 | -1.57 | 0.12 |
| home_val | 0.00 | 0.00 | 0.38 | 0.70 |
| yoj | 42.32 | 42.11 | 1.00 | 0.32 |
| income | -2.18 | 2.36 | -0.92 | 0.36 |

Standard errors: OLS

| | |
|---|---|
| Observations | 1873 |
| Dependent variable | severity |
| Type | OLS linear regression |

| | |
|---|---|
| F(33,1839) | 23.04 |
| $R^2$ | 0.29 |
| Adj. $R^2$ | 0.28 |

| | Est. | S.E. | t val. | p |
|---|---|---|---|---|
| (Intercept) | 0.55 | 0.04 | 13.07 | 0.00 |
| kidsdrivY | 0.00 | 0.01 | 0.19 | 0.85 |
| homekids | 0.00 | 0.01 | 0.50 | 0.61 |
| parent1Y | -0.01 | 0.02 | -0.32 | 0.75 |
| mstatusY | -0.02 | 0.01 | -1.74 | 0.08 |
| sexM | -0.01 | 0.02 | -0.36 | 0.72 |
| educationBachelors | -0.01 | 0.01 | -0.91 | 0.36 |
| educationMasters | -0.00 | 0.02 | -0.18 | 0.86 |
| educationPhD | 0.05 | 0.03 | 1.80 | 0.07 |
| travtime | 0.00 | 0.00 | 0.35 | 0.73 |
| car_usePrivate | 0.01 | 0.01 | 0.55 | 0.58 |
| bluebook | -0.00 | 0.00 | -17.88 | 0.00 |
| tifmoderate | -0.00 | 0.01 | -0.08 | 0.93 |
| tifhigh | -0.00 | 0.02 | -0.18 | 0.85 |
| car_typePanel Truck | 0.10 | 0.02 | 3.97 | 0.00 |
| car_typePickup | 0.04 | 0.02 | 2.55 | 0.01 |
| car_typeSports Car | 0.02 | 0.02 | 1.12 | 0.26 |
| car_typeSUV | 0.02 | 0.02 | 0.99 | 0.32 |
| car_typeVan | -0.01 | 0.02 | -0.38 | 0.71 |
| red_carY | 0.01 | 0.01 | 0.57 | 0.57 |
| oldclaim | 0.00 | 0.00 | 1.64 | 0.10 |
| clm_freqmoderate | -0.02 | 0.01 | -1.82 | 0.07 |
| clm_freqhigh | -0.06 | 0.02 | -2.38 | 0.02 |
| revokedY | -0.01 | 0.01 | -0.47 | 0.64 |
| mvr_ptslow | 0.01 | 0.01 | 1.19 | 0.23 |
| mvr_ptshigh | 0.02 | 0.01 | 1.93 | 0.05 |
| urbanicityUrban | 0.01 | 0.02 | 0.70 | 0.48 |
| car_age | -0.00 | 0.00 | -0.07 | 0.94 |
| home_val | -0.00 | 0.00 | -0.36 | 0.72 |
| yoj | 0.00 | 0.00 | 1.51 | 0.13 |
| income | -0.00 | 0.00 | -1.36 | 0.17 |
| age | -0.00 | 0.00 | -1.33 | 0.18 |
| jobProfessional | 0.01 | 0.01 | 1.08 | 0.28 |
| liquidityhigh | 0.02 | 0.02 | 1.14 | 0.25 |

Standard errors: OLS