Crash & Cost Prediction

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

10/15/2021

Introduction

The below analysis centers around predicting the probability of a car crash; and the cost implications of said crash, based on a collection of observations. Naturally we will begin with an exploration of the data to build an initial impression on the relationships; which will guide our variable transformations and/or variable selections. This will lead into the construction of two models: a logistic regression for the binary target variable of Crash vs No Crash; and a linear model for the target dollar cost variable. Ultimately, we will integrate both results to provide a summary from the context of an insurance provider.

In this report we will:

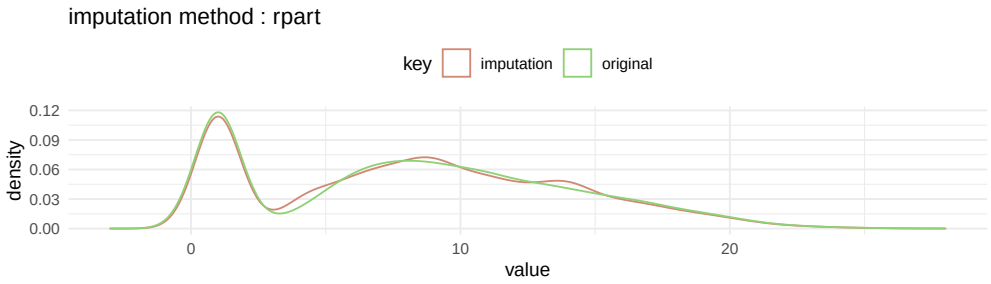
- Explore the data
- Transform data to address multicollenearity and meet variable distribution needs
- Compare different models and select the most accurate model
- Test our model on the evaluation dataset

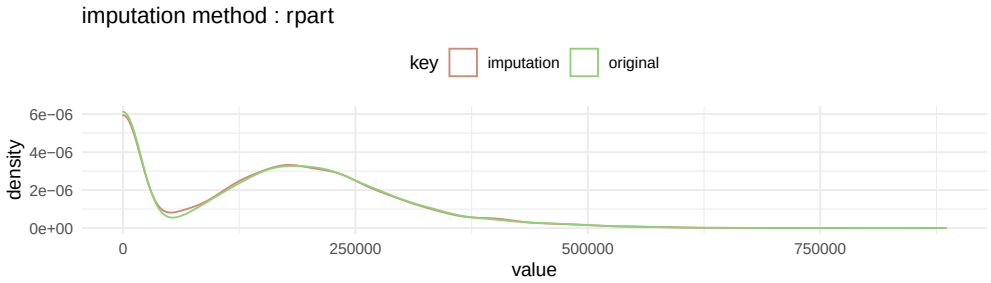
Data cleaning

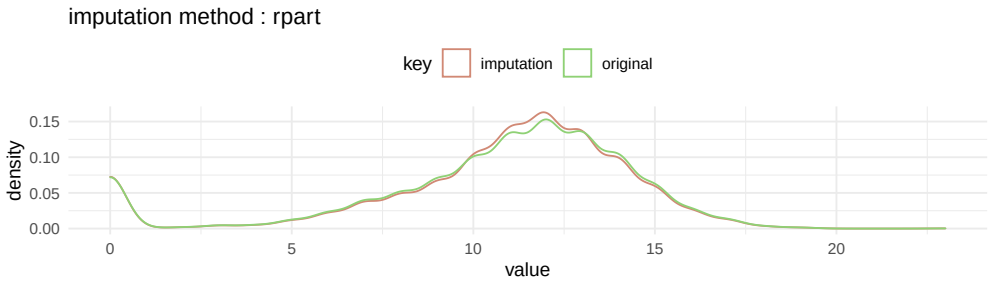
variables	types	missing_count	missing_percent
job	factor	526	6.4452886
car_age	numeric	510	6.2492342
home_val	numeric	464	5.6855777
yoj	numeric	454	5.5630437
income	numeric	445	5.4527631
age	numeric	6	0.0735204

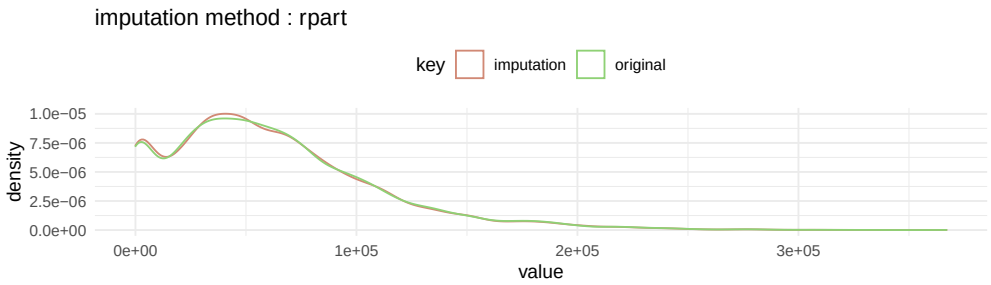
We move to impute the missing data:
Recursive Partitioning and Regression Trees is used to impute the numerical variable.
Multivariate Imputation by Chained Equations is used to impute the categorical variable.

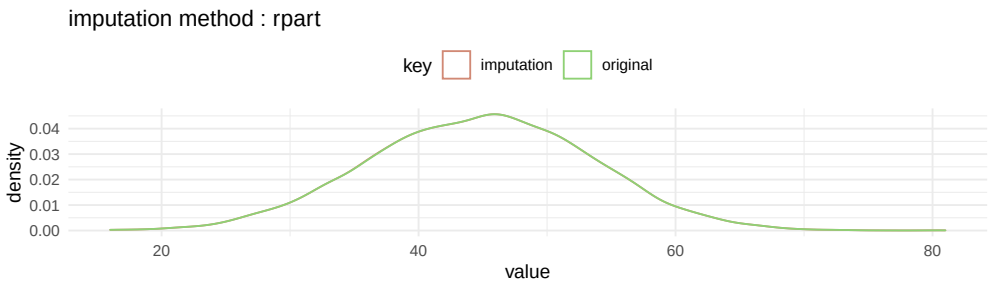
The following plots confirm the imputation follows the nature of the existing data, so we are confident the results our analysis are not affected.

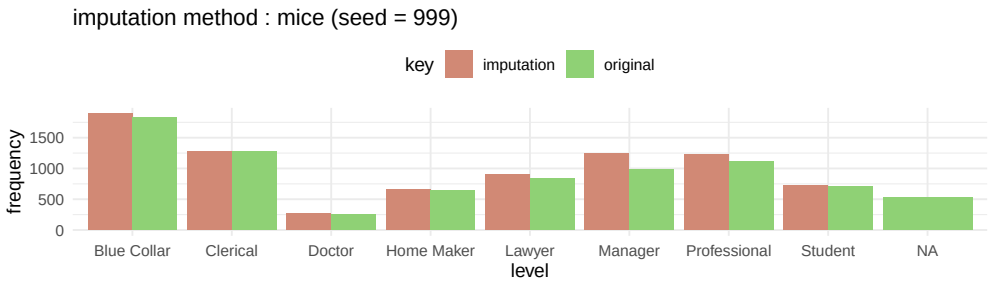

imputation method : rpart
key
imputation
original
density
0.12
0.09
0.06
0.03
0.00
0
10
20
value


imputation method : rpart
key
imputation
original
density
6e−06
4e−06
2e−06
0e+00
0
250000
500000
750000
value


imputation method : rpart
key
imputation
original
density
0.15
0.10
0.05
0.00
0
5
10
15
20
value


imputation method : rpart
key
imputation
original
density
1.0e−05
7.5e−06
5.0e−06
2.5e−06
0.0e+00
0e+00
1e+05
2e+05
3e+05
value

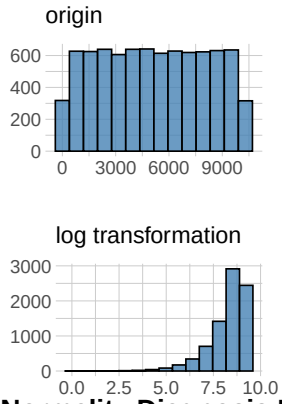

imputation method : rpart
key
imputation
original
density
0.04
0.03
0.02
0.01
0.00
20
40
60
80
value


imputation method : mice (seed = 999)
key
imputation
original
frequency
1500
1000
500
0
Blue Collar
Clerical
Doctor
Home Maker
Lawyer
Manager
Professional
Student
NA
level

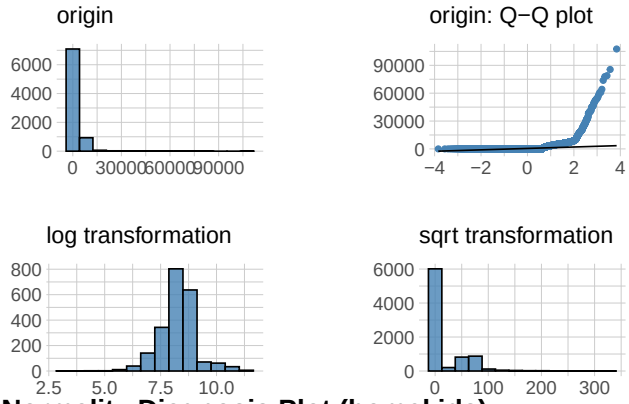
Distributions

There are some important findings from examining the histograms of the variables.Response variables: Both of our target variables are very skewed with a long right tail. 'target_amt' appears to respond well to a log transformation. However 'target_flag' is categorical; so we will plan on implementing a zero inflation strategy. Predictors: 'car_age' and 'home_val' show a bi-modal distribution, with centers around zero and more normal appearing right tail. This is to be expected with 'home_val' as those who do not have a home would return a zero value. The same is not obvious for why 'car_age' would have so many clustered closed to zero. We cannot say more without further context, but it should be noted in case there are issues down the line.

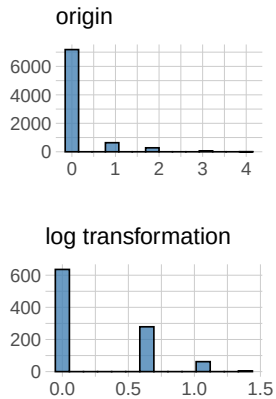
Normality Diagnosis Plot (index)



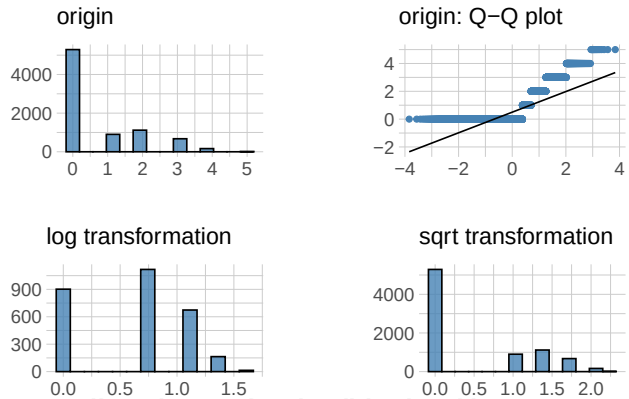
Normality Diagnosis Plot (target_amt)



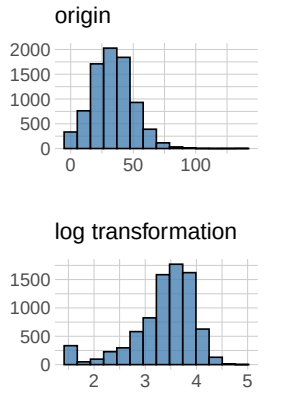
Normality Diagnosis Plot (kidsdriv)



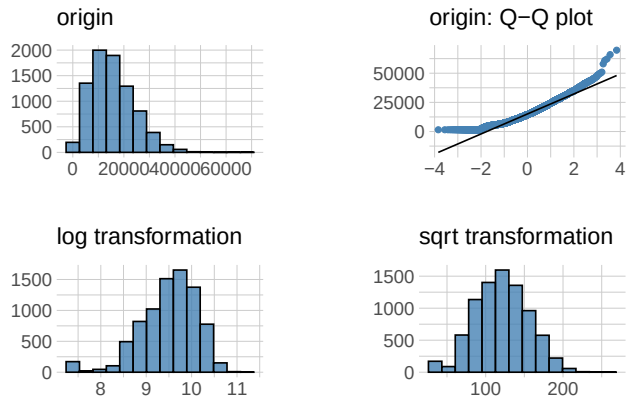
Normality Diagnosis Plot (homekids)



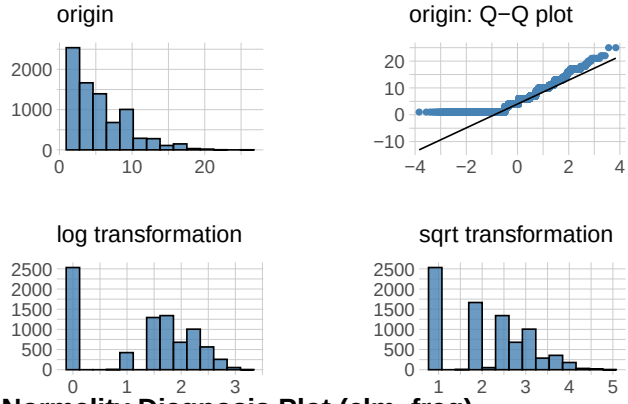
Normality Diagnosis Plot (travtime)



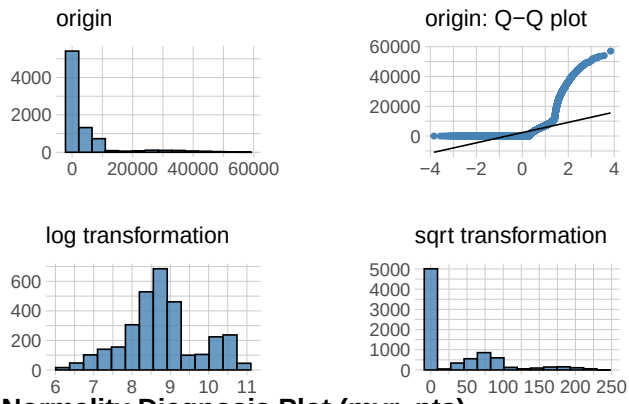
Normality Diagnosis Plot (bluebook)



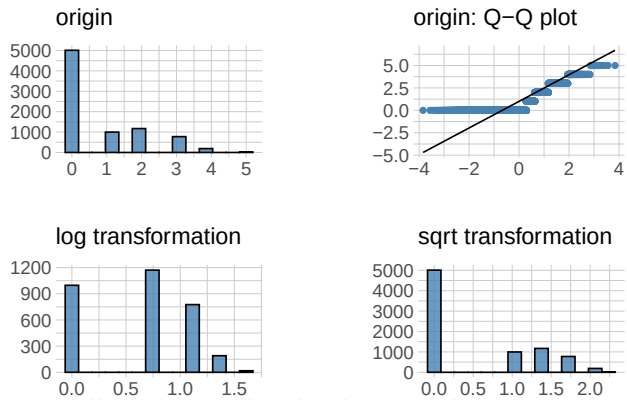
Normality Diagnosis Plot (tif)



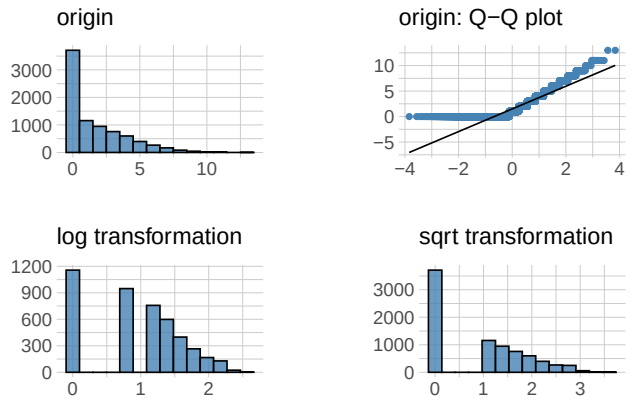
Normality Diagnosis Plot (oldclaim)



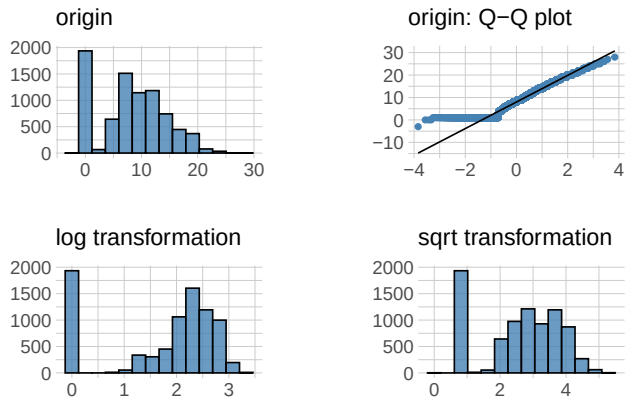
Normality Diagnosis Plot (clm_freq)



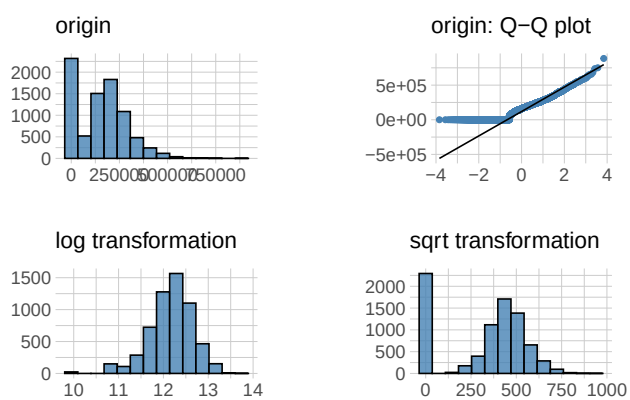
Normality Diagnosis Plot (mvr_pts)

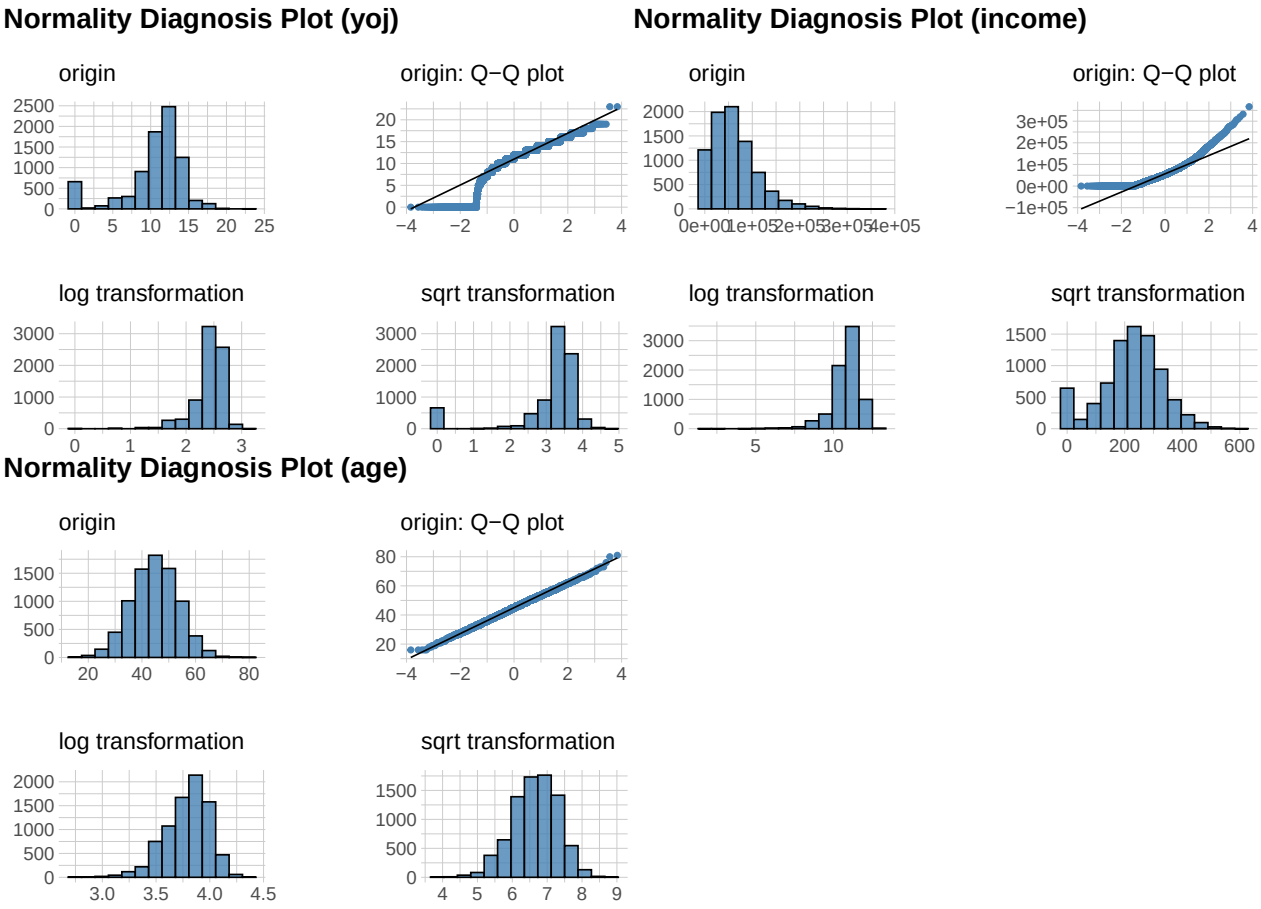


Normality Diagnosis Plot (car_age)



Normality Diagnosis Plot (home_val)





Outliers

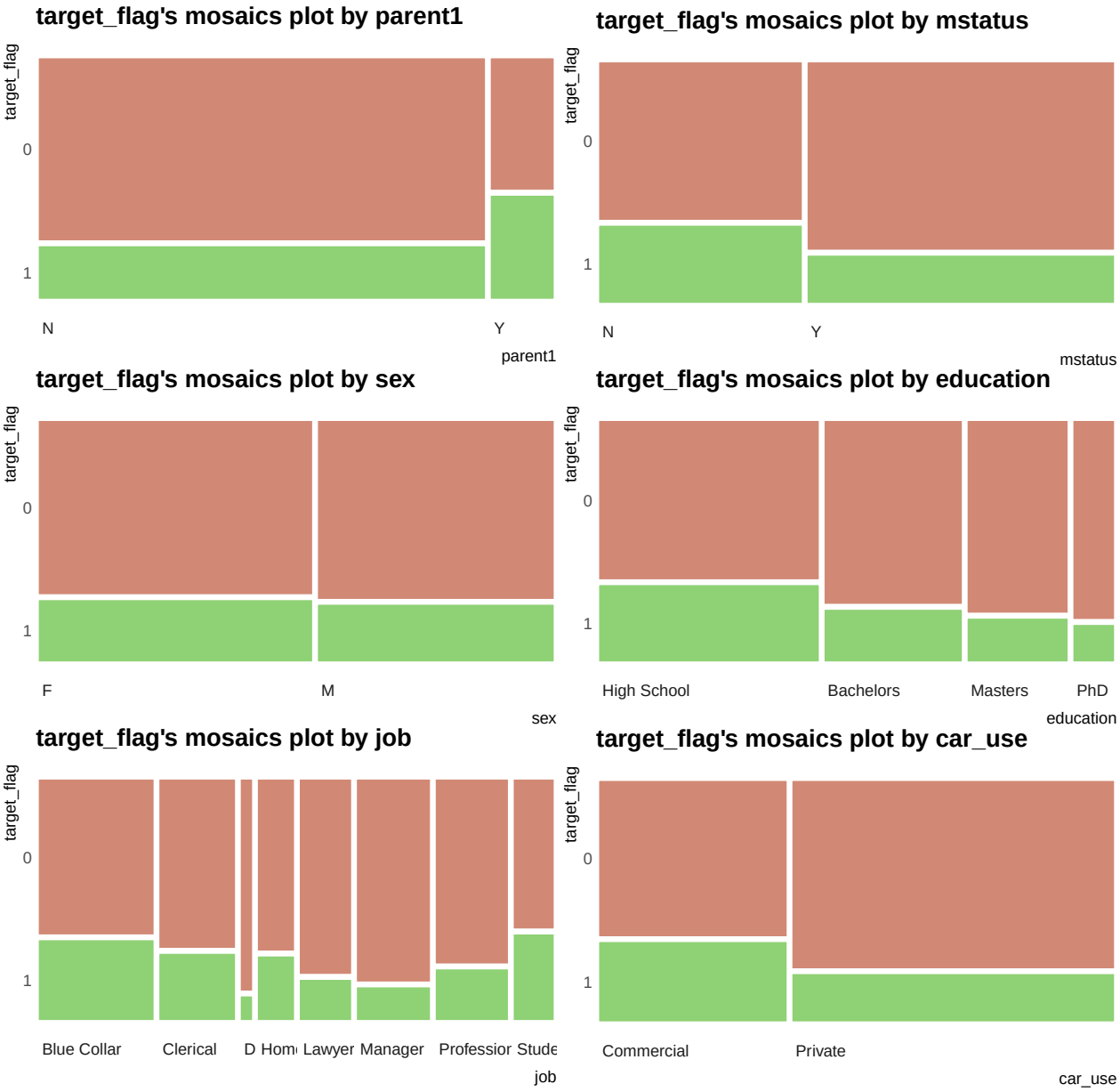
We note outlier concentrations of >5% for target_amt, kidsdriv, homekids, yoj and oldclaim.

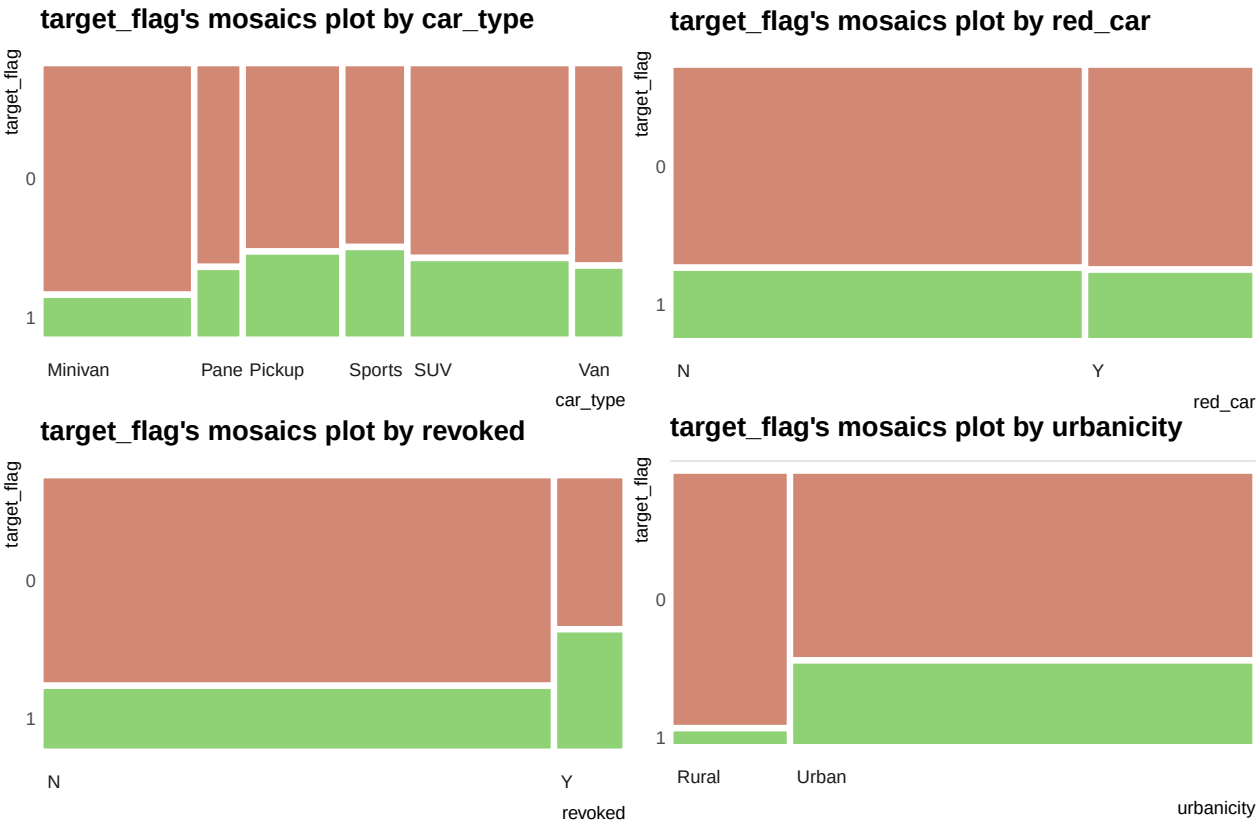
variables	outliers_cnt	outliers_ratio	outliers_mean	with_mean	without_mean
target_amt	1,620	19.851	7,039.976	1,504.325	133.318
kidsdriv	981	12.021	1.423	0.171	0.000
homekids	852	10.440	3.225	0.721	0.429
yoj	682	8.357	0.120	10.517	11.465
oldclaim	663	8.124	30,358.611	4,037.076	1,709.632
income	275	3.370	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

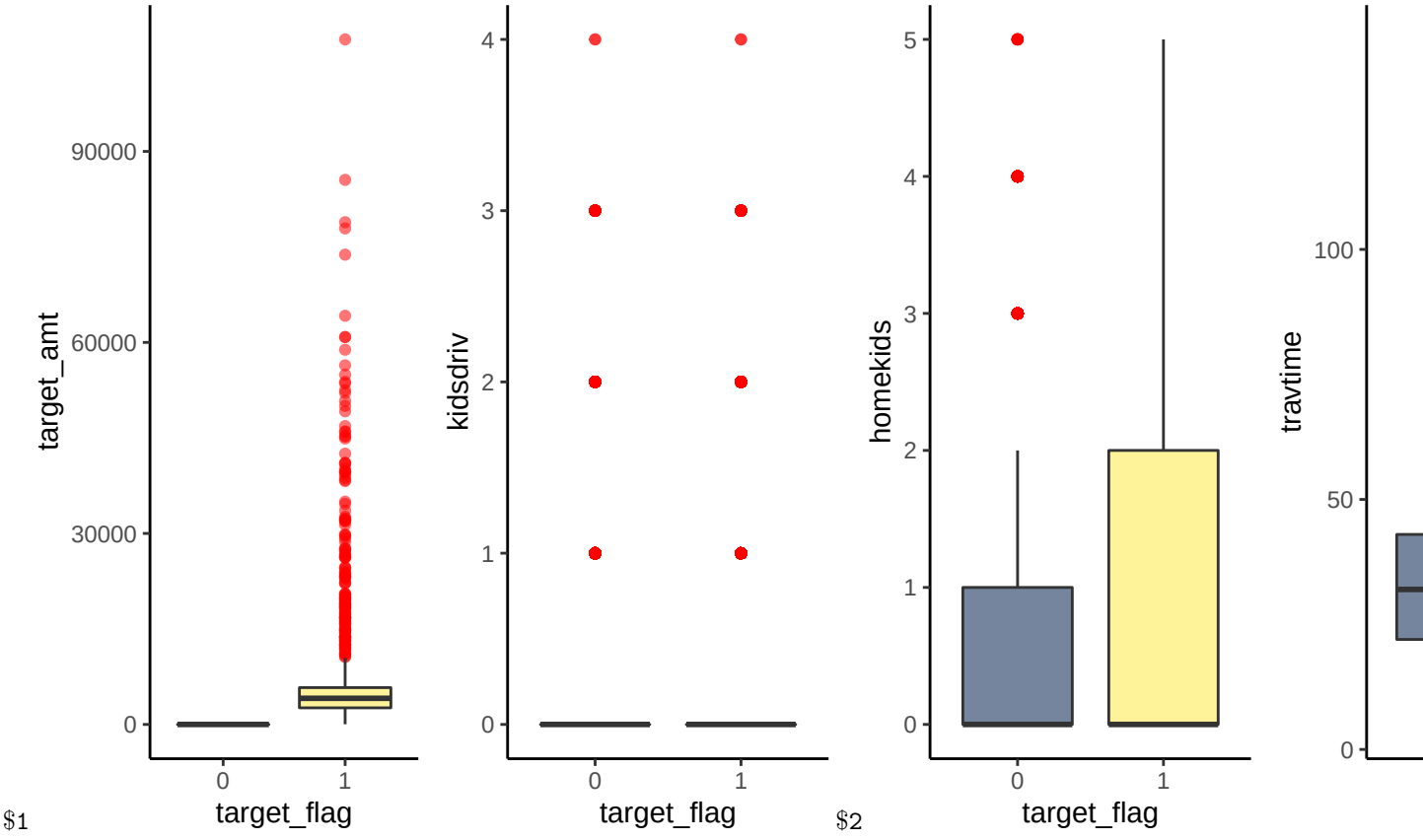
Explore relationships between response and categorical predictors

Upon reviewing the below mosaic and box plots, we can determine that the below listed variables have hardly any relationship with the response. This will be kept in mind during the variable selection phase. 'sex' 'red_car'



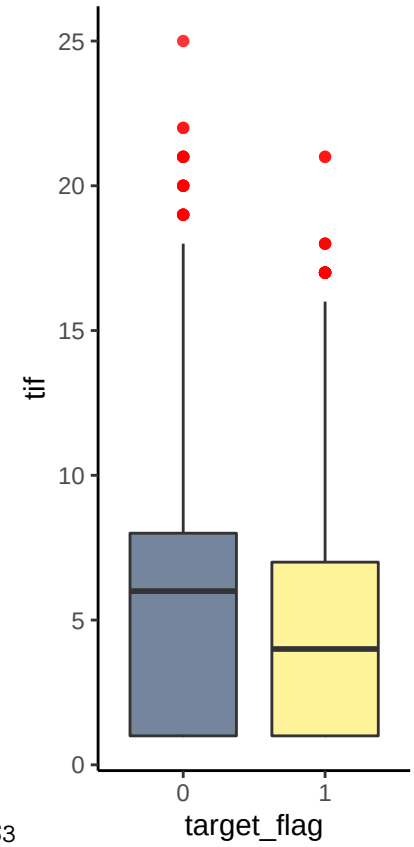


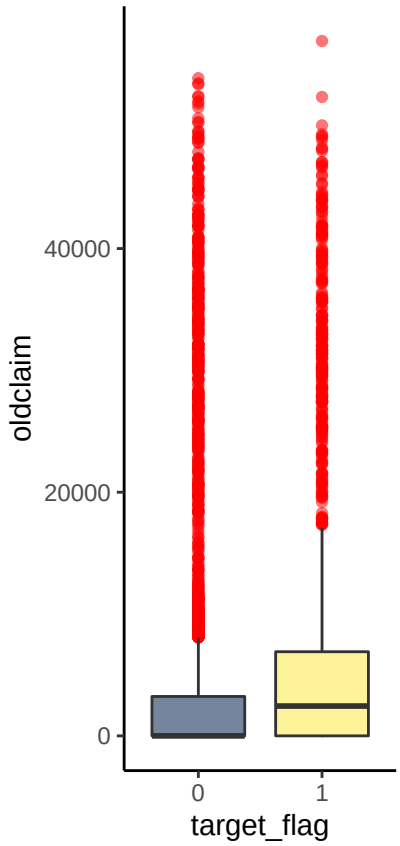
Box plot for numerical variables

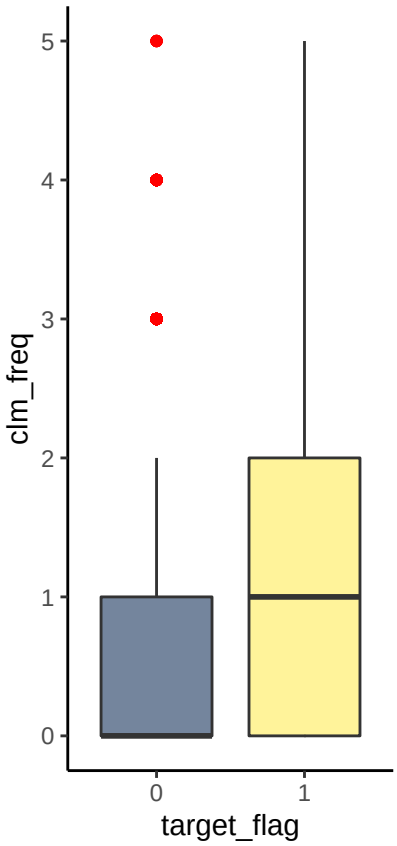


$1

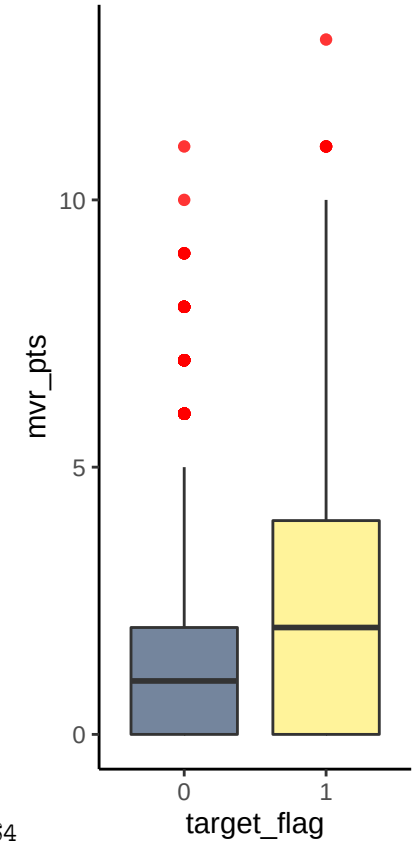
$2

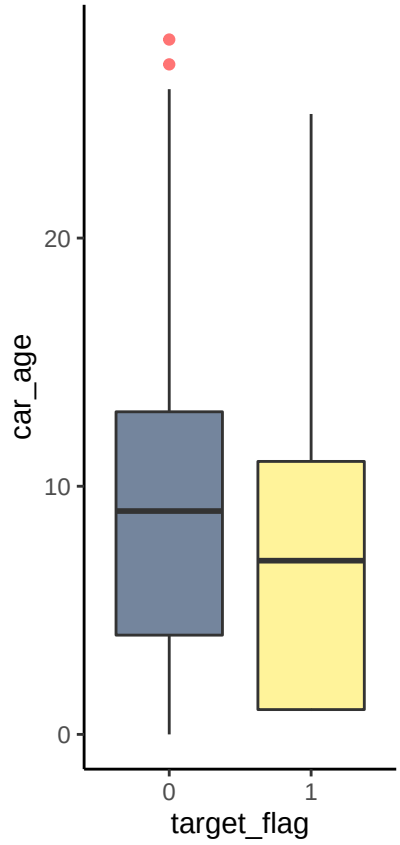

tif
25
20
15
10
5
0
0
1
target_flag

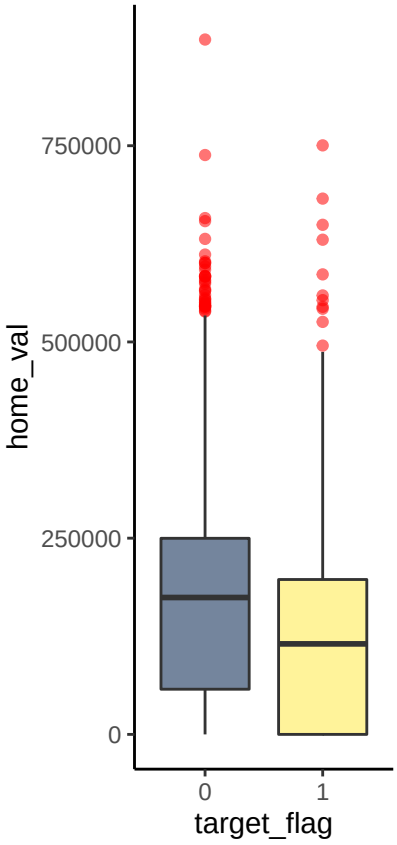

oldclaim
40000
20000
0
0
1
target_flag


clm_freq
5
4
3
2
1
0
0
1
target_flag

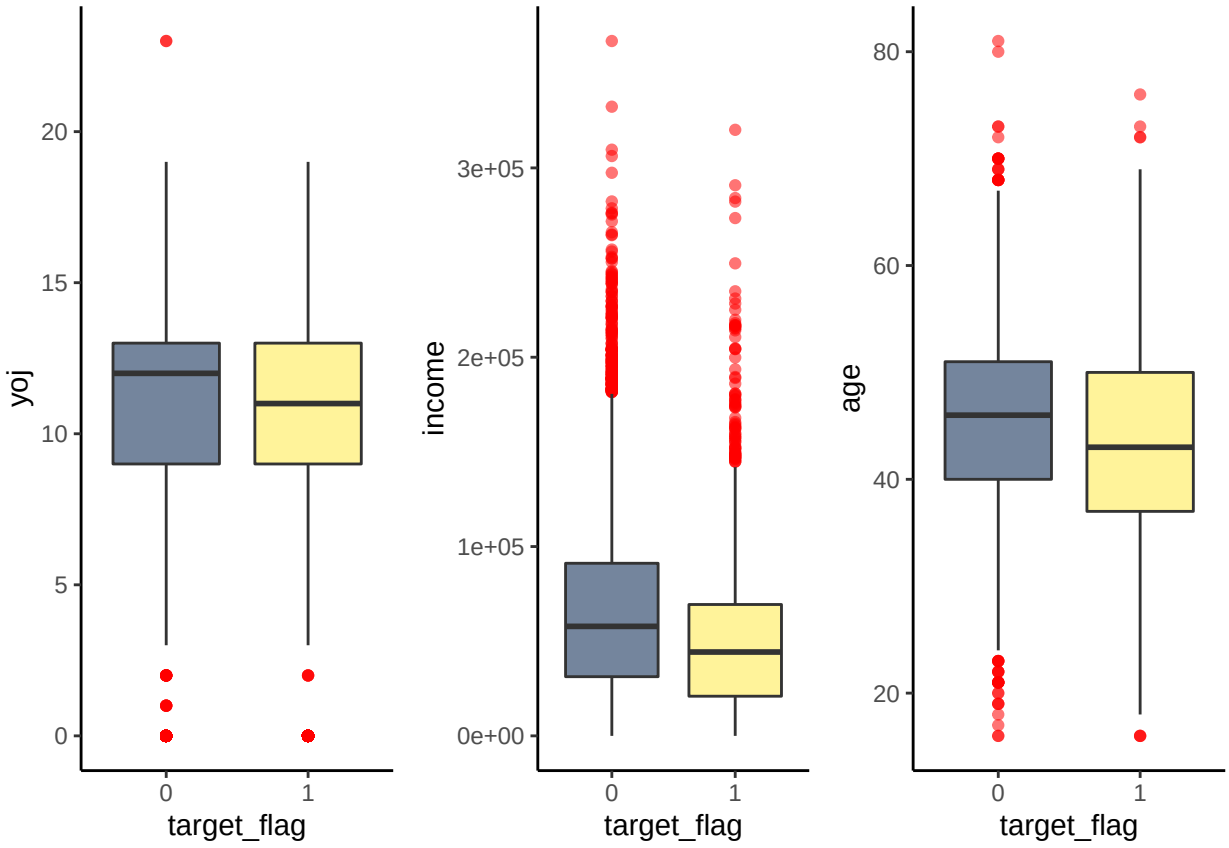
$3


mvr_pts
10
5
0
0
1
target_flag


car_age
20
10
0
0
1
target_flag


home_val
750000
500000
250000
0
0
1
target_flag

$4



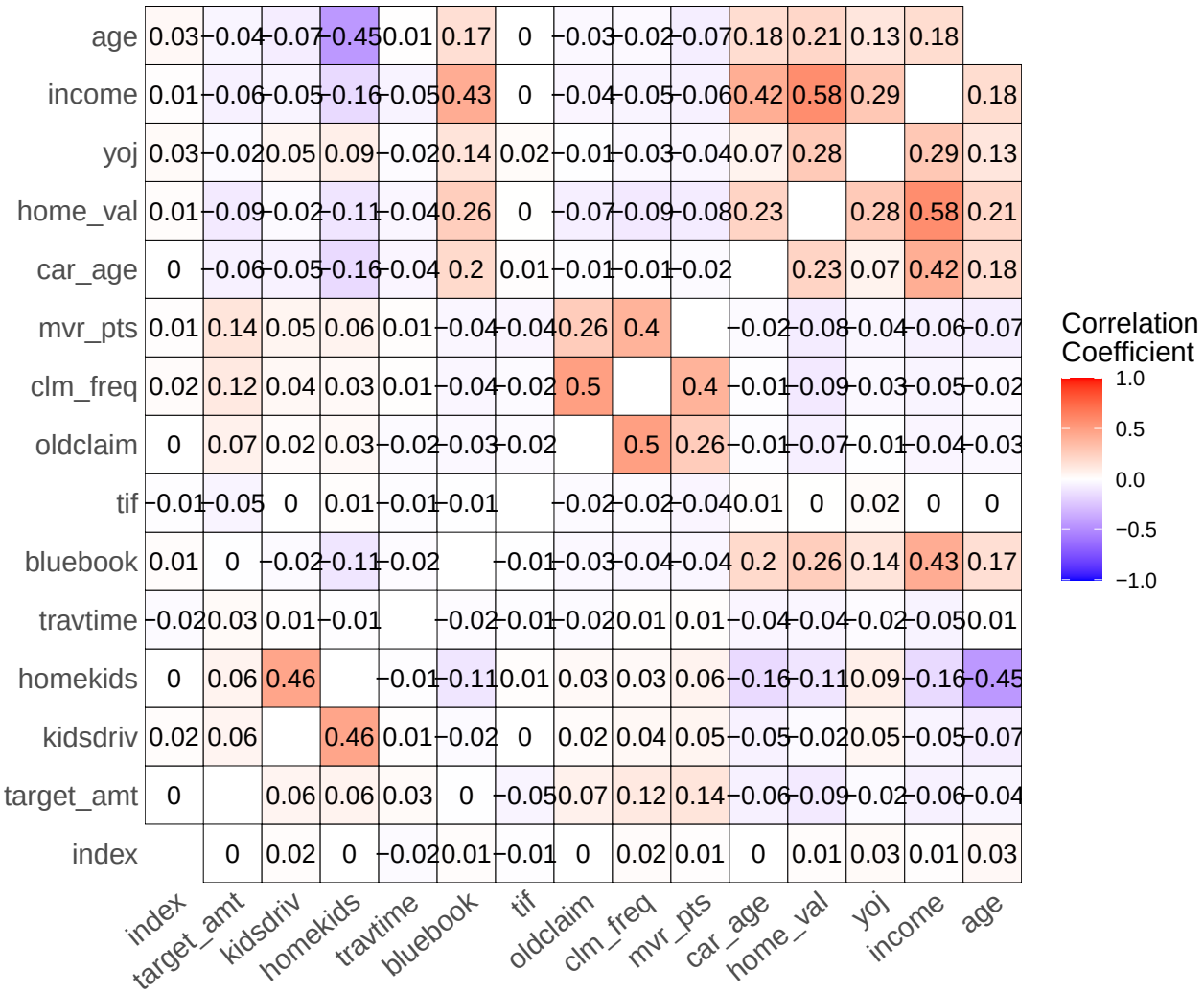
$5
attr(,“class”) [1] “list” “ggarrange”

Co-variance

We establish that there is only one pair of predictors that have a co-variance of >.5; ‘home_val’ & ‘income’. We may consider combining into an interaction term, or possible removing one from the model. We also note that many of the correlations against the target variable appear low; which is consistent with the above plots.

A tibble: 2 x 3

var1 var2 coef_corr 1 income home_val 0.581 2 home_val income 0.581


age 0.03 −0.04 −0.07 −0.45 0.01 0.17 0 −0.03 −0.02 −0.07 0.18 0.21 0.13 0.18
income 0.01 −0.06 −0.05 −0.16 −0.05 0.43 0 −0.04 −0.05 −0.06 0.42 0.58 0.29 0.18
yoj 0.03 −0.02 0.05 0.09 −0.02 0.14 0.02 −0.01 −0.03 −0.04 0.07 0.28 0.29 0.13
home_val 0.01 −0.09 −0.02 −0.11 −0.04 0.26 0 −0.07 −0.09 −0.08 0.23 0.28 0.58 0.21
car_age 0 −0.06 −0.05 −0.16 −0.04 0.2 0.01 −0.01 −0.01 −0.02 0.23 0.07 0.42 0.18
mvr_pts 0.01 0.14 0.05 0.06 0.01 −0.04 −0.04 0.26 0.4 −0.02 −0.08 −0.04 −0.06 −0.07
clm_freq 0.02 0.12 0.04 0.03 0.01 −0.04 −0.02 0.5 0.4 −0.01 −0.09 −0.03 −0.05 −0.02
oldclaim 0 0.07 0.02 0.03 −0.02 −0.03 −0.02 0.5 0.26 −0.01 −0.07 −0.01 −0.04 −0.03
tif −0.01 −0.05 0 0.01 −0.01 −0.01 −0.02 −0.02 −0.04 0.01 0 0.02 0 0
bluebook 0.01 0 −0.02 −0.11 −0.02 −0.01 −0.03 −0.04 −0.04 0.2 0.26 0.14 0.43 0.17
travtime −0.02 0.03 0.01 −0.01 −0.02 −0.01 −0.02 0.01 0.01 −0.04 −0.04 −0.02 −0.05 0.01
homekids 0 0.06 0.46 −0.01 −0.11 0.01 0.03 0.03 0.06 −0.16 −0.11 0.09 −0.16 −0.45
kidsdriv 0.02 0.06 0.46 0.01 −0.02 0 0.02 0.04 0.05 −0.05 −0.02 0.05 −0.05 −0.07
target_amt 0 0.06 0.06 0.03 0 −0.05 0.07 0.12 0.14 −0.06 −0.09 −0.02 −0.06 −0.04
index 0 0.02 0 −0.02 0.01 −0.01 0 0.02 0.01 0 0.01 0.03 0.01 0.03
index
target_amt
kidsdriv
homekids
travtime
bluebook
tif
oldclaim
clm_freq
mvr_pts
car_age
home_val
yoj
income
age
Correlation Coefficient
1.0
0.5
0.0
−0.5
−1.0

Construct Logistical Classification Model

Step 1. Assess class balance - 74% = 0, 26% =1.A 3:1 ratio really isn't a rare event issue. However, looked into weighting and various balancing approaches. They all caused the AIC to sky-rocket. Decision made to move forward with training data of similar proportion.

Table 3: All Observations

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

Step 2. Make additional factor/level adjustments following prior data evaluation

-kidsdriv [N,Y] -job [Blue Collar , Professional] -education [High School, Bachelors, Masters, PhD]

Step 3. Build a training and test data set.

Partition off 25% of the data to serve as a test set.

Table 4: Training Sample

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

Table 5: Test Sample

target_flag	n	frequency
0	1,502	0.7362745
1	538	0.2637255

Model 1: Base logistic model

This model includes all predictors and Akaiki criterion for variable selection.

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

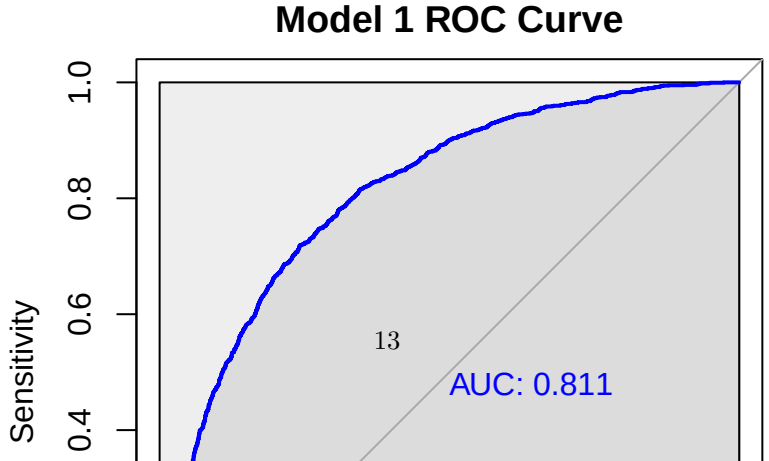
Model 1 Evaluation

Note: yoj, red_car, age, car_age are not significant in Model 1 (df_train)

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

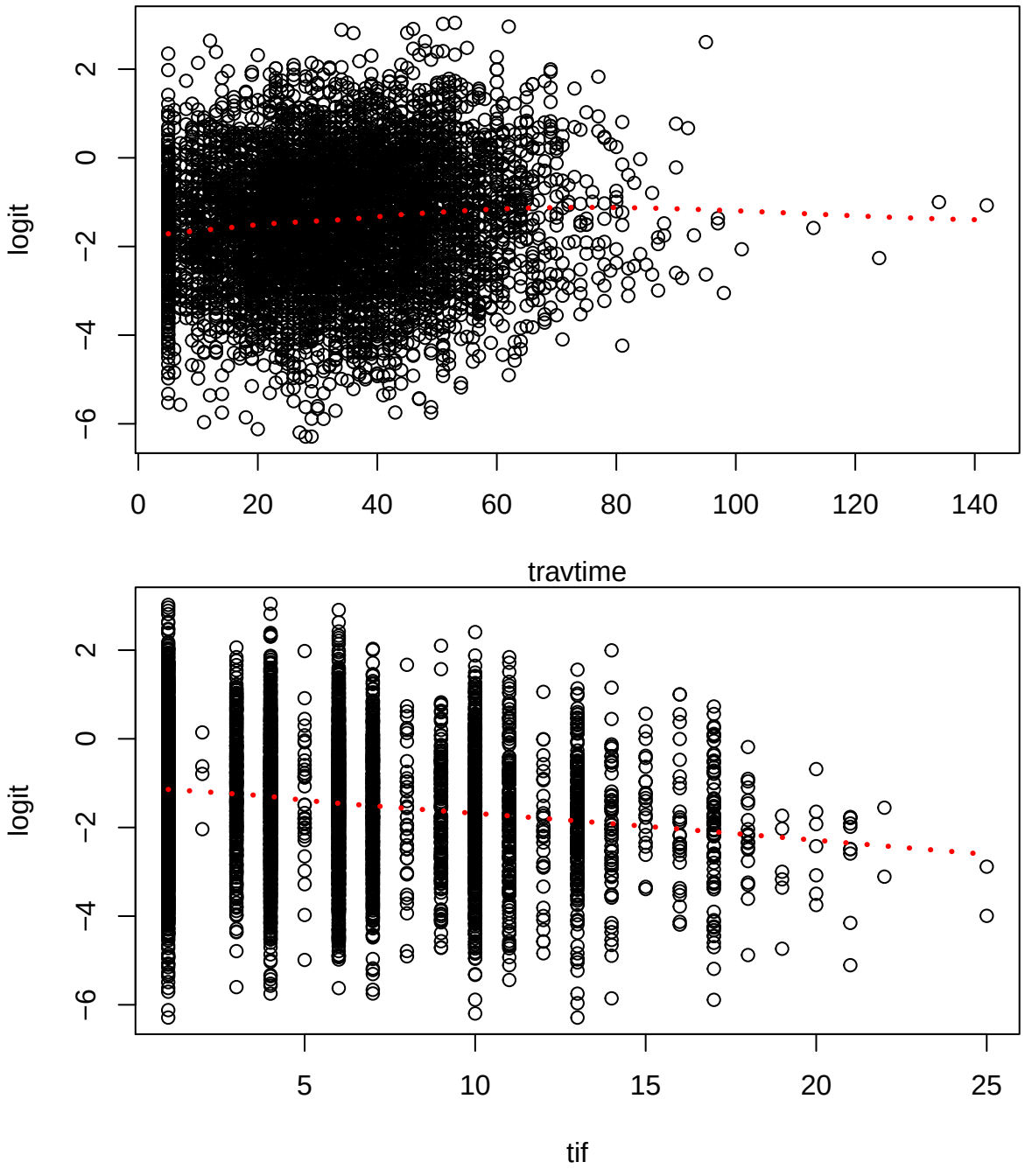


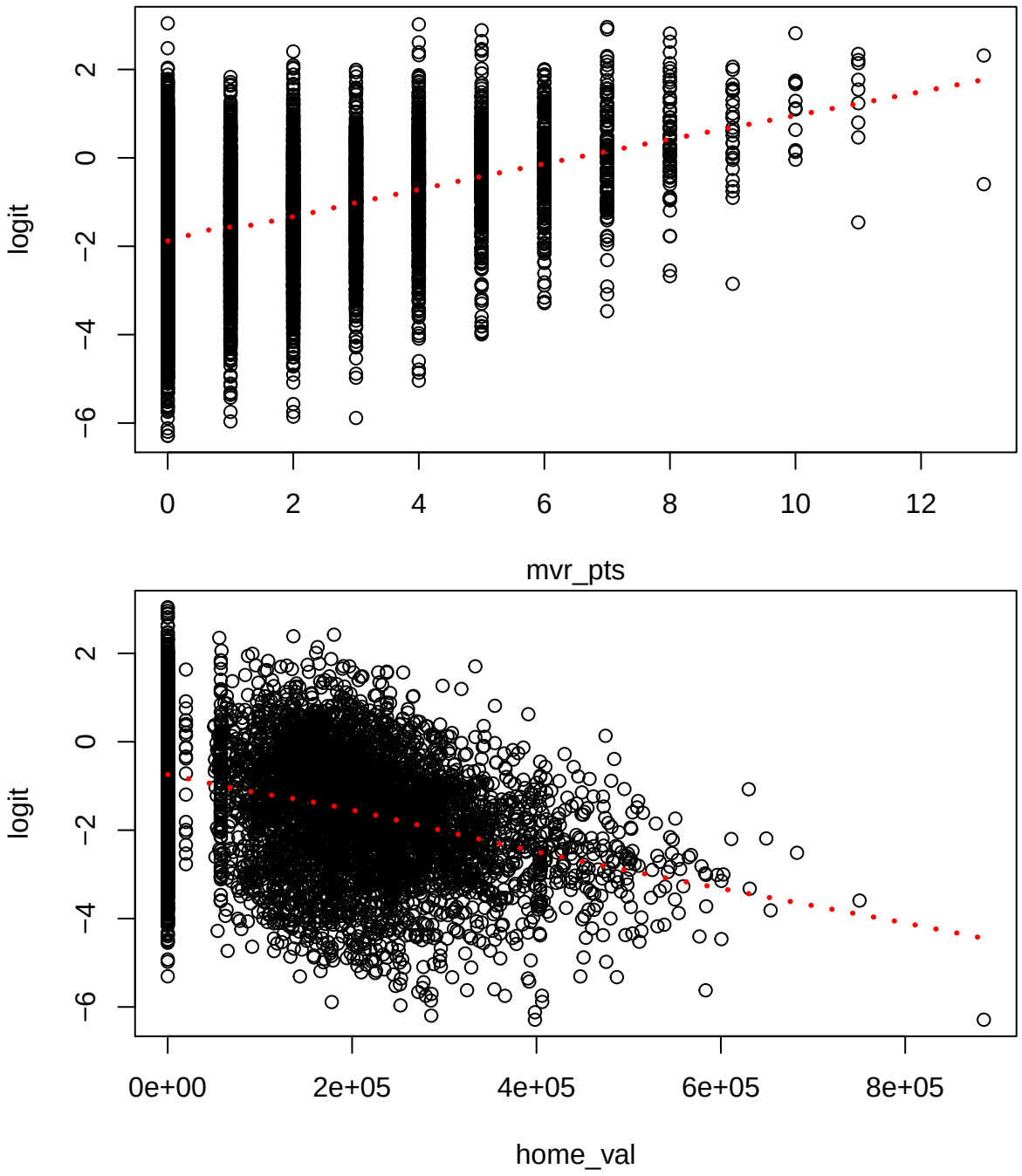
Dispersion We assess dispersion with two calculations; their results are shown below.

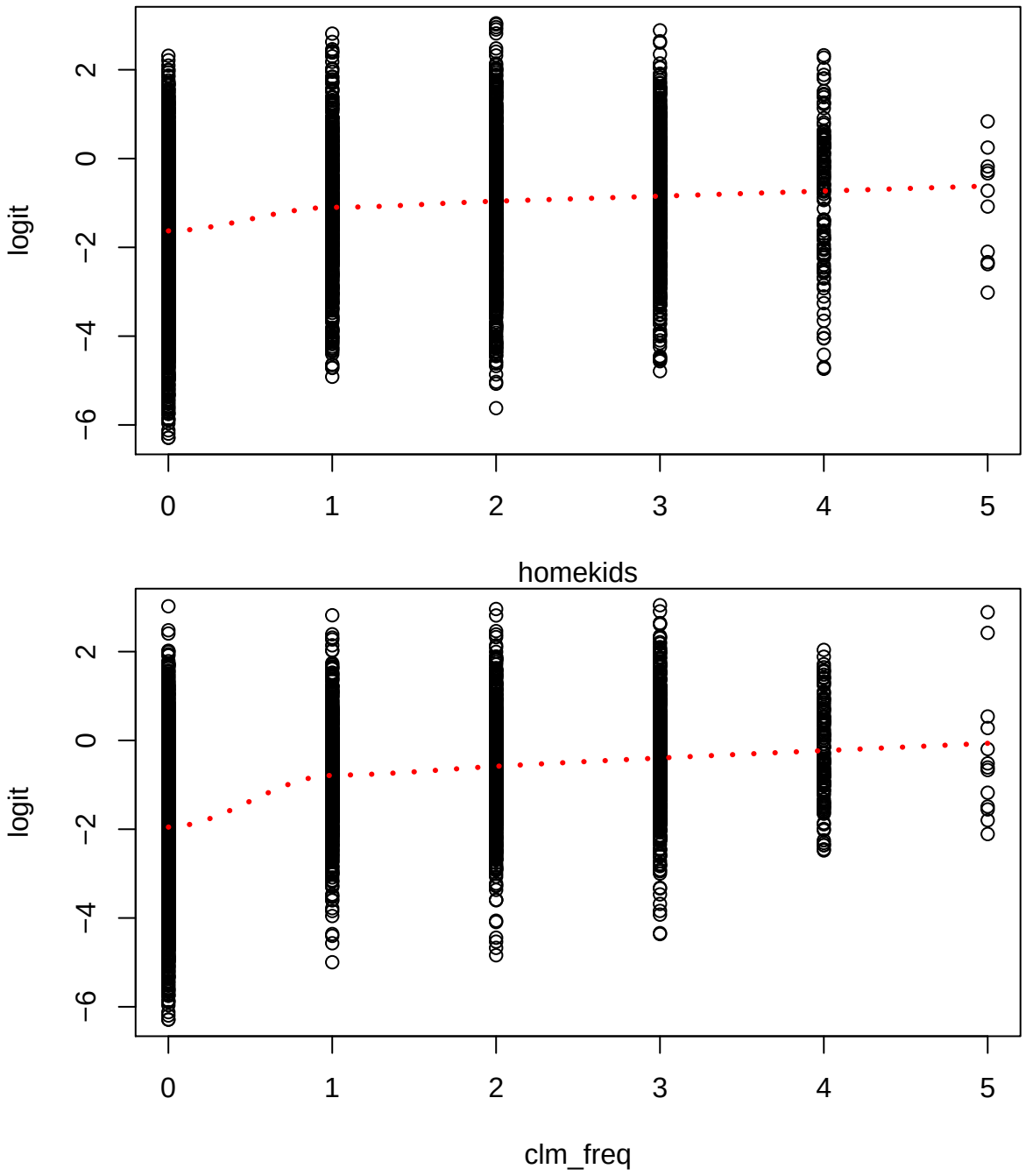
[1] "We divide the deviance by the residuals to obtain the value 0.9024. There is no overt concern since the values is not greater than 1" [1] "Next we obtain a Pearson Chi-Squared test statistic of 0.3133 This communicates that the null hypothesis is not rejected and their are no problems with dispersion."

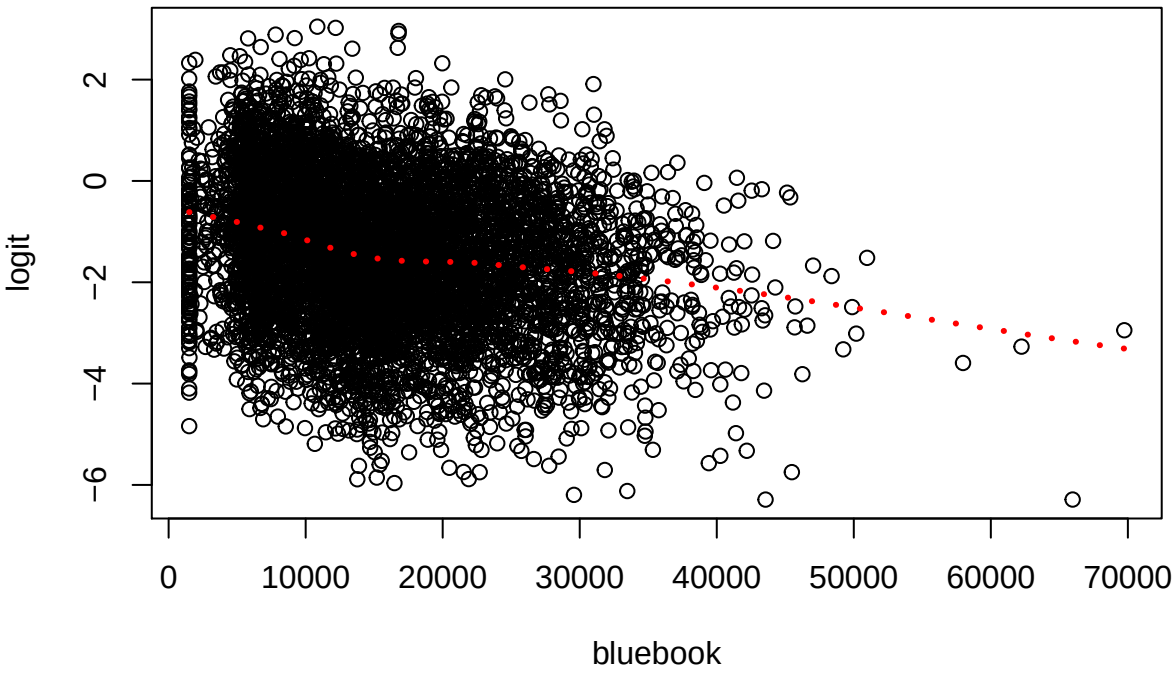
Assumption of Linearity

We can note that linearity is questionable for home-kids, but not convincing enough to remove at this time. yoj, oldclaim, or income are not considered in this diagnostic since they were not significant.




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

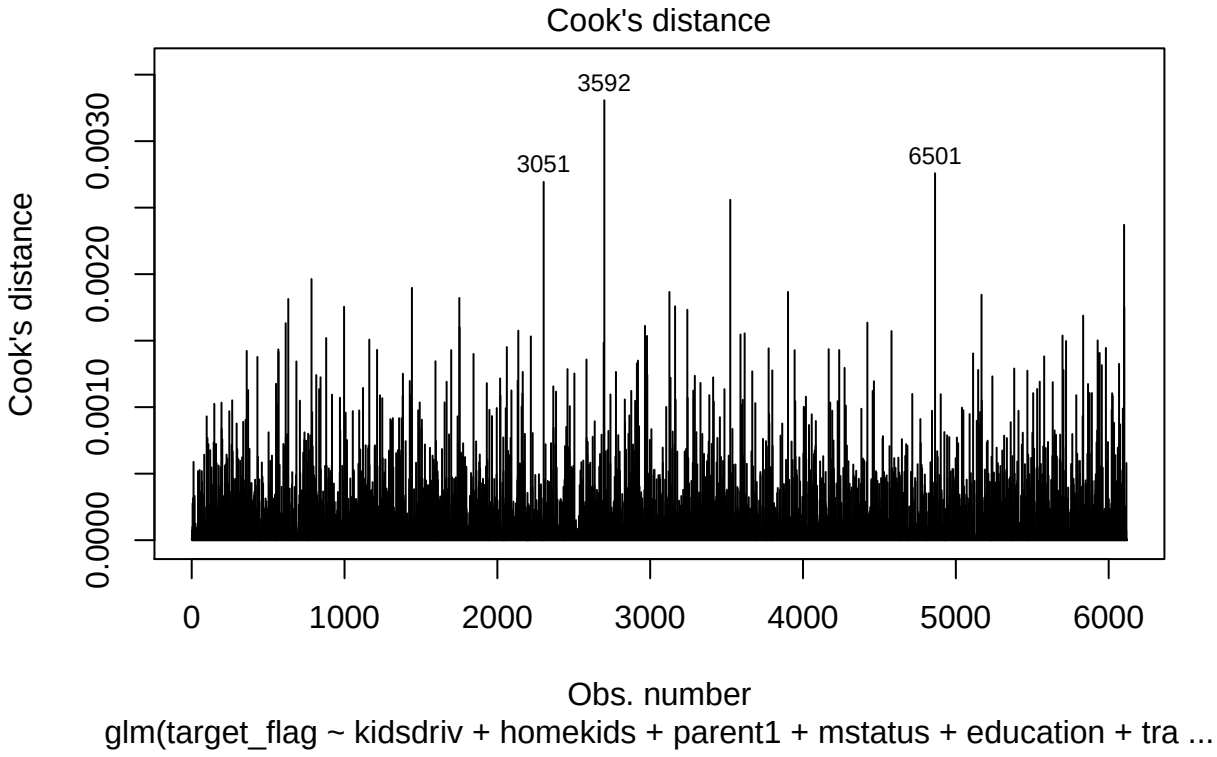

logit
2
0
−2
−4
−6
0
1
2
3
4
5
homekids
logit
2
0
−2
−4
−6
0
1
2
3
4
5
clm_freq

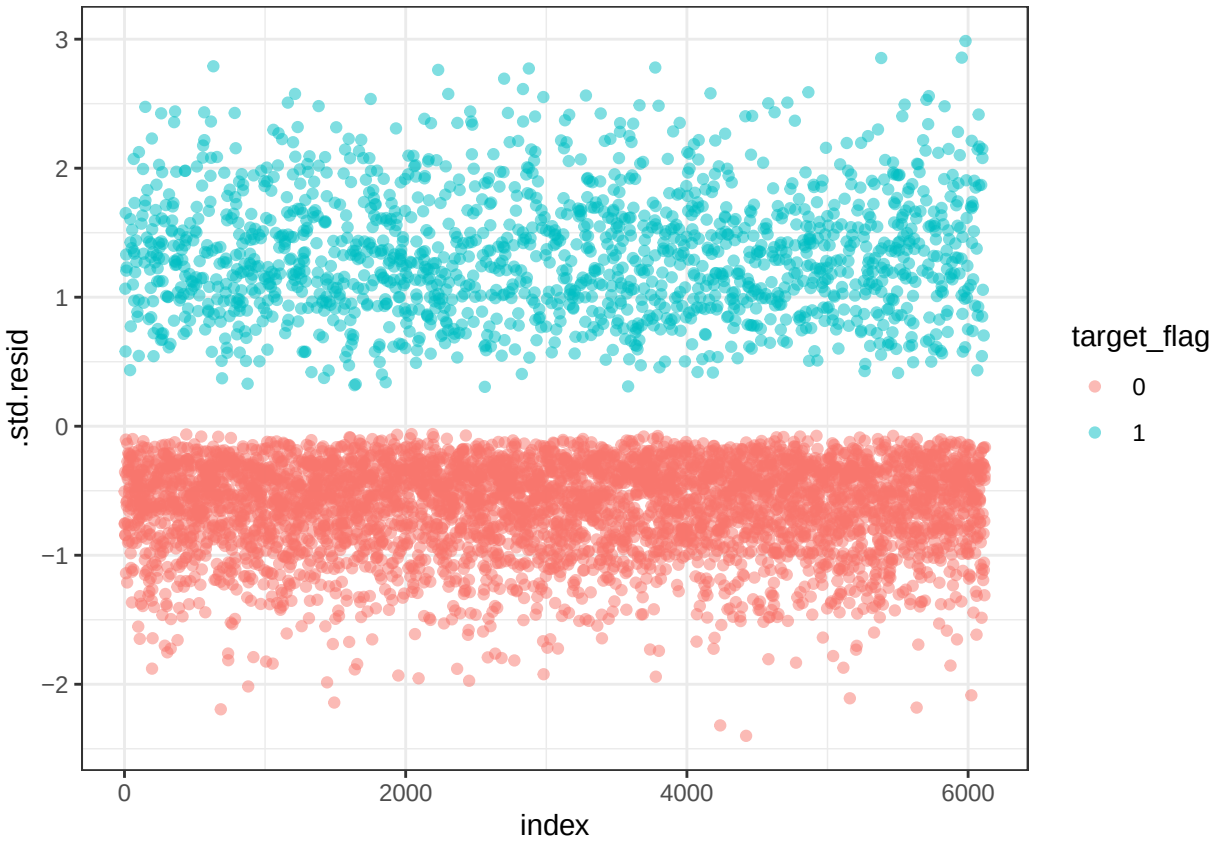


Outliers & Influential Points

Examining the standardized residuals (.std.resid) and the Cook's distance (.cooksd) using the R function augment() [broom package] we can note the below findings. -Cooks distance indicates several standout obs (3722, 3592, 6501) but no influential points (id. D >1.0) -There are no obs with std residual beyond 3 stdev - ie., no influential obs

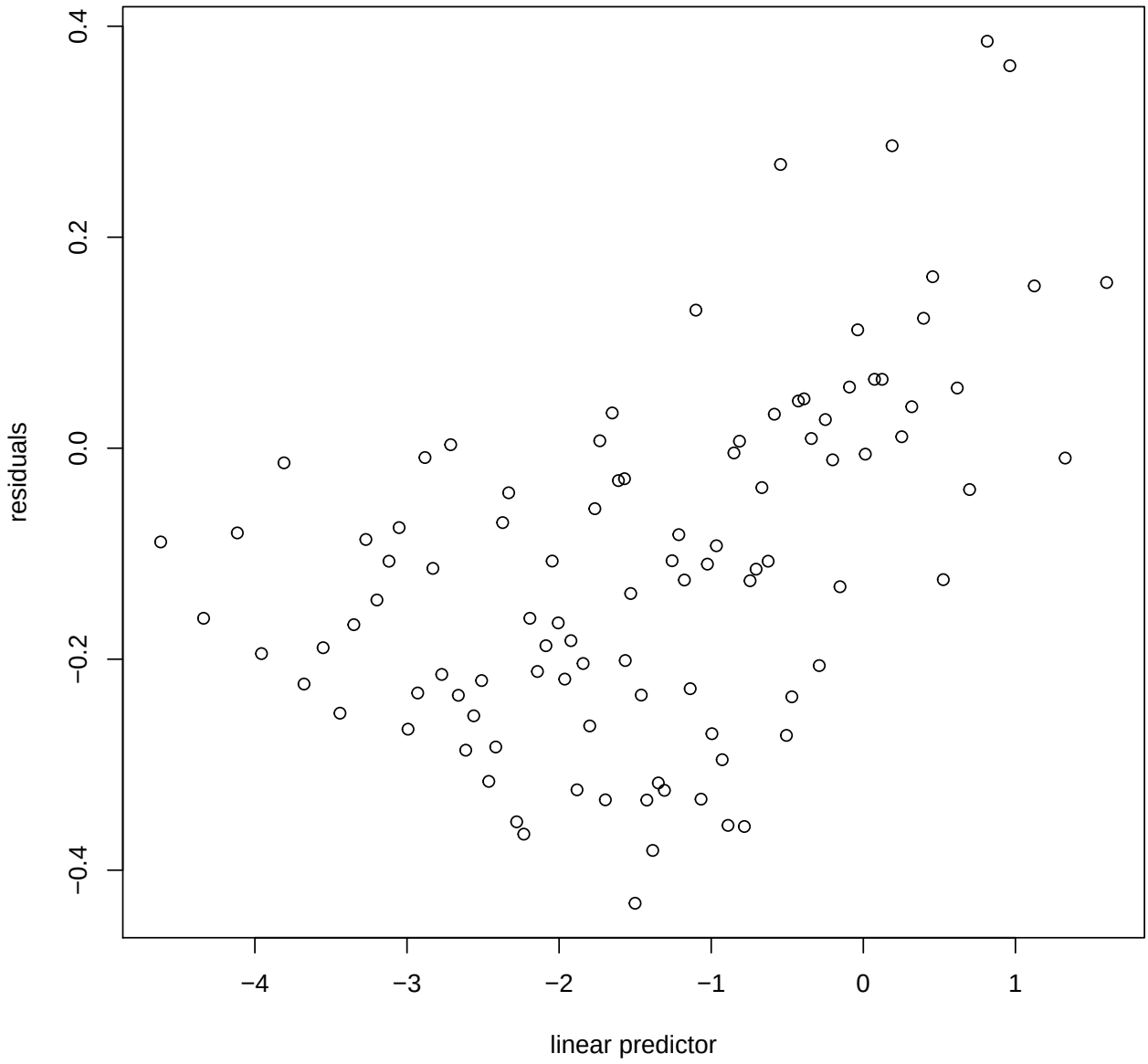
**http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/





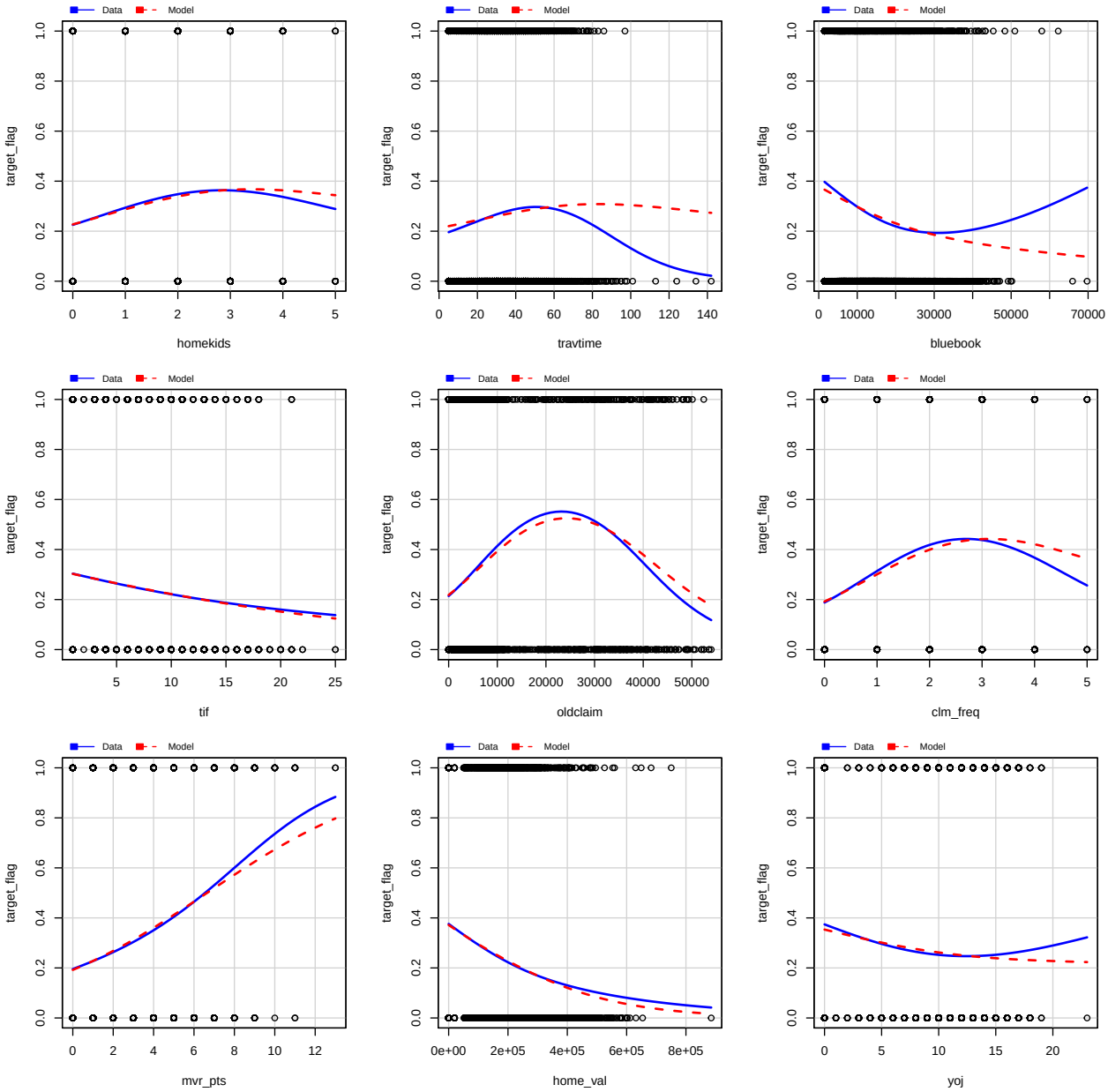
Check for Independence

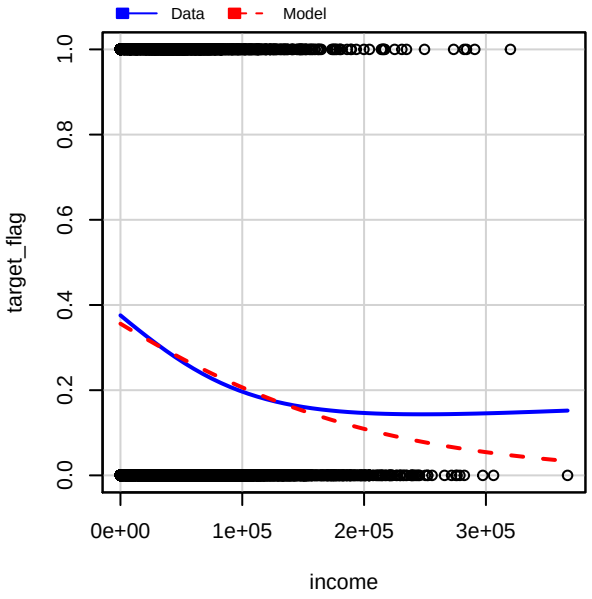
Each point on the below plot represents an aggregation of the prediction & residual values for each percentile bin of the predictions. We observe that the higher percentiles have a higher average residual. We see this as definite pattern which suggests that the model may be miss-classified.

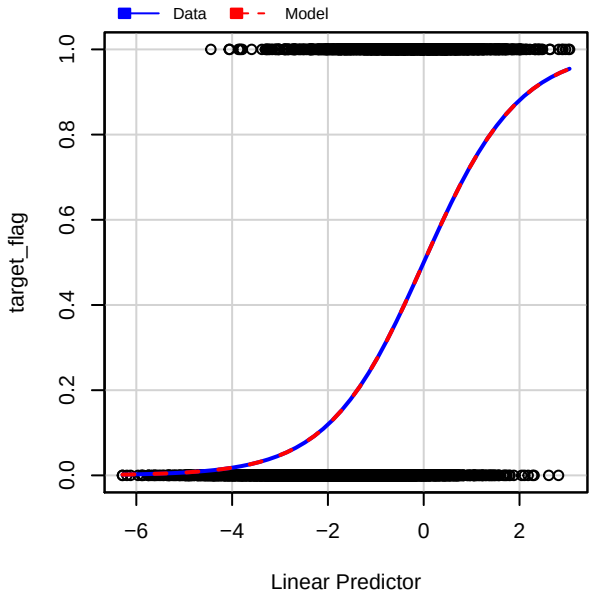


Goodness of Fit - marginal plots

We now review how the fitted logistic regression compares to the data using the below marginal plots. Findings: consider transformations for trav_time and tif. We will drop income, oldclaim, and yoj from subsequent models.


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

Model 2: Apply Predictor Transformations

The following transformations result from some trial and error:

sqrt: income log: bluebook quadratic: travtime

yoj & homekids removed per insignificance.

Note: AIC has gone down slightly relative to Model1

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

$\chi^2(24)$	1581.45
Pseudo-R^2 (Cragg-Uhler)	0.33
Pseudo-R^2 (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

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

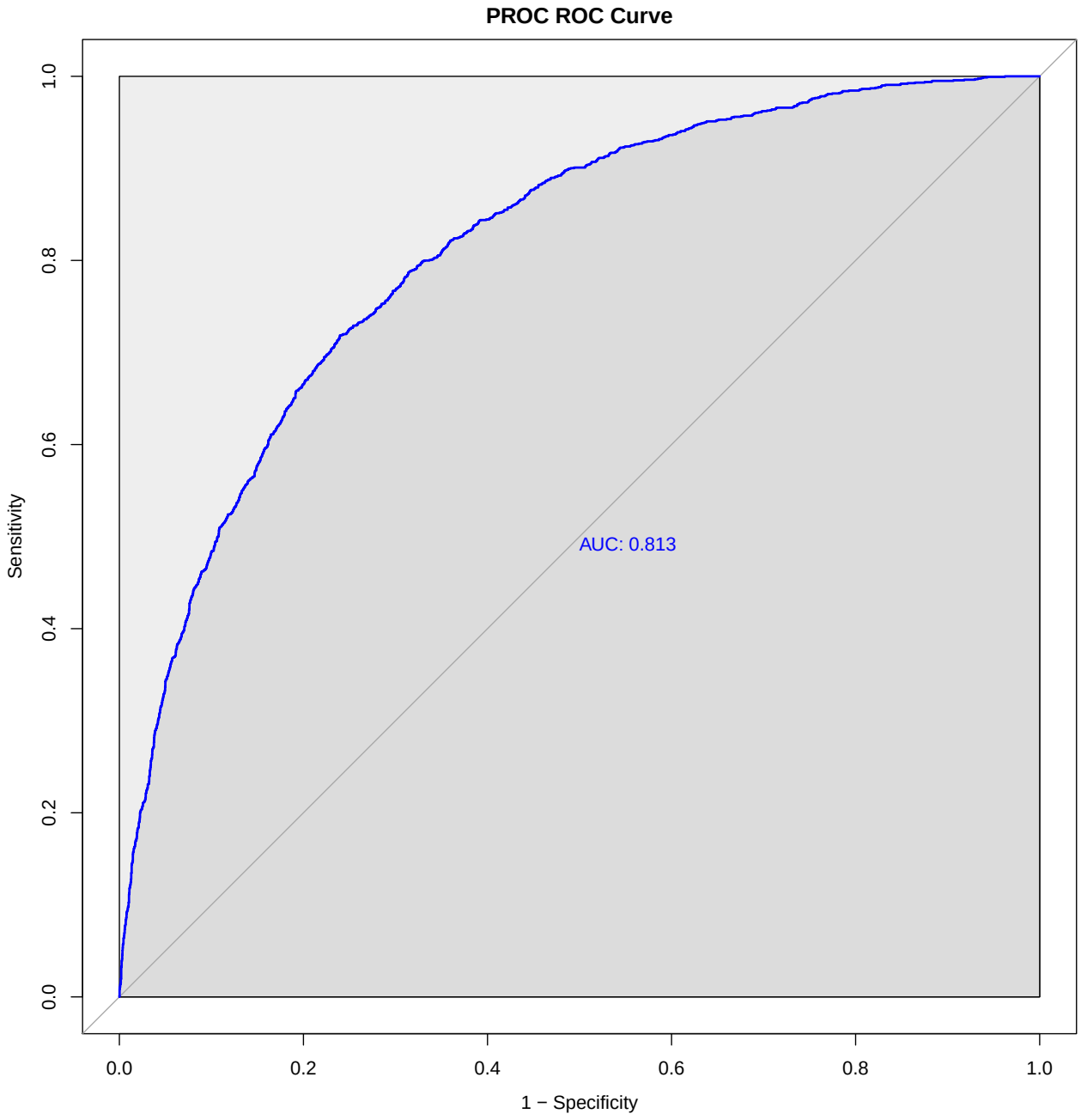
```
        Sensitivity : 0.9239
        Specificity : 0.4250
     Pos Pred Value : 0.8177
     Neg Pred Value : 0.6667
         Prevalence : 0.7363
     Detection Rate : 0.6802
```

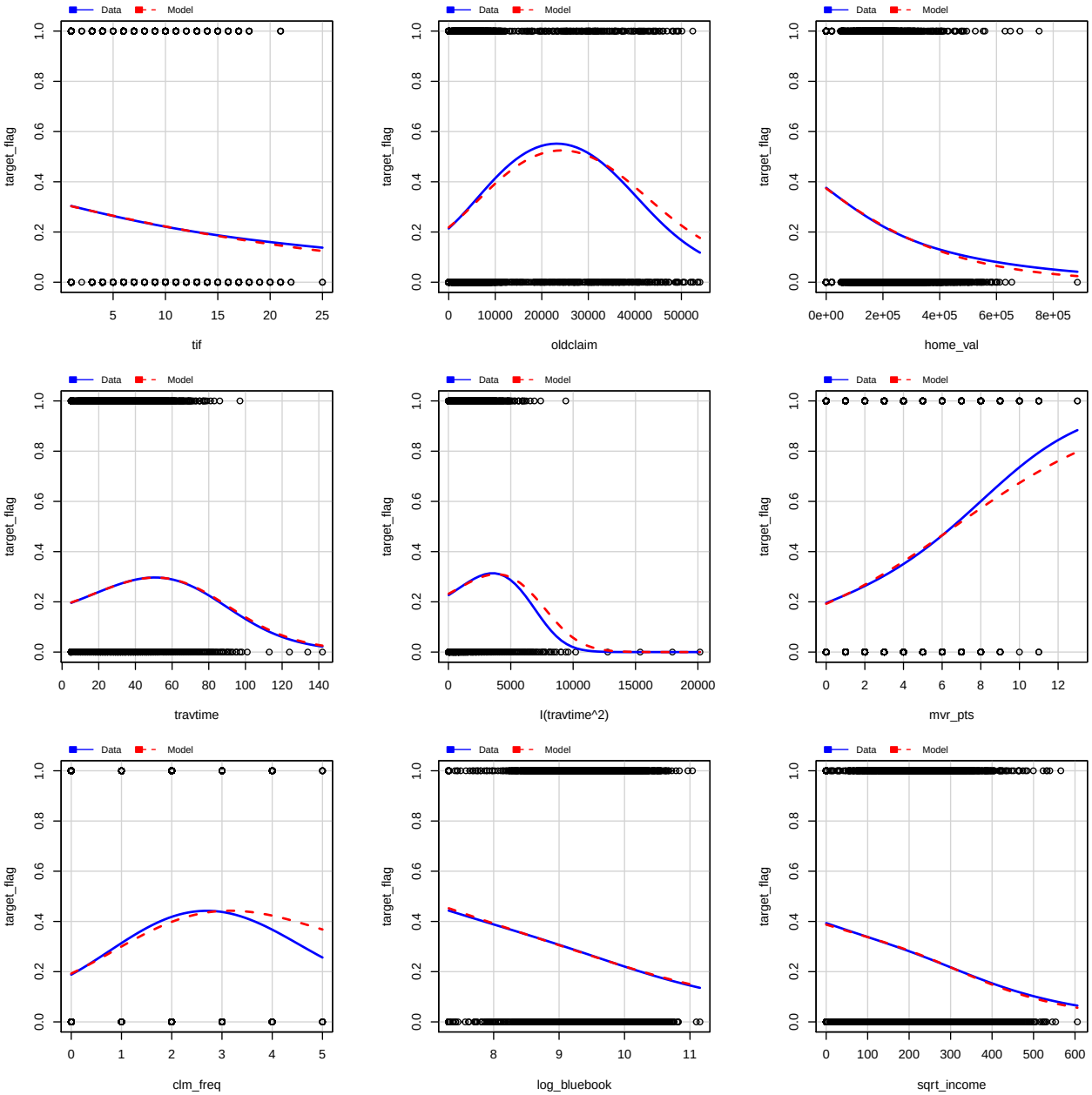
Detection Prevalence : 0.8319
Balanced Accuracy : 0.6745

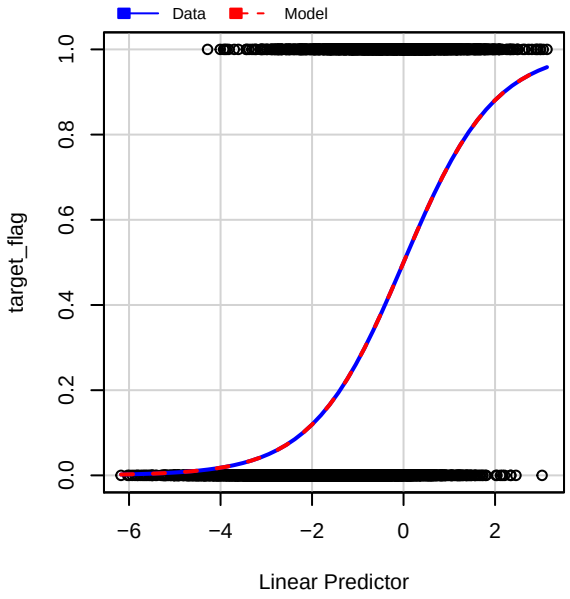
```
   'Positive' Class : 0
```

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

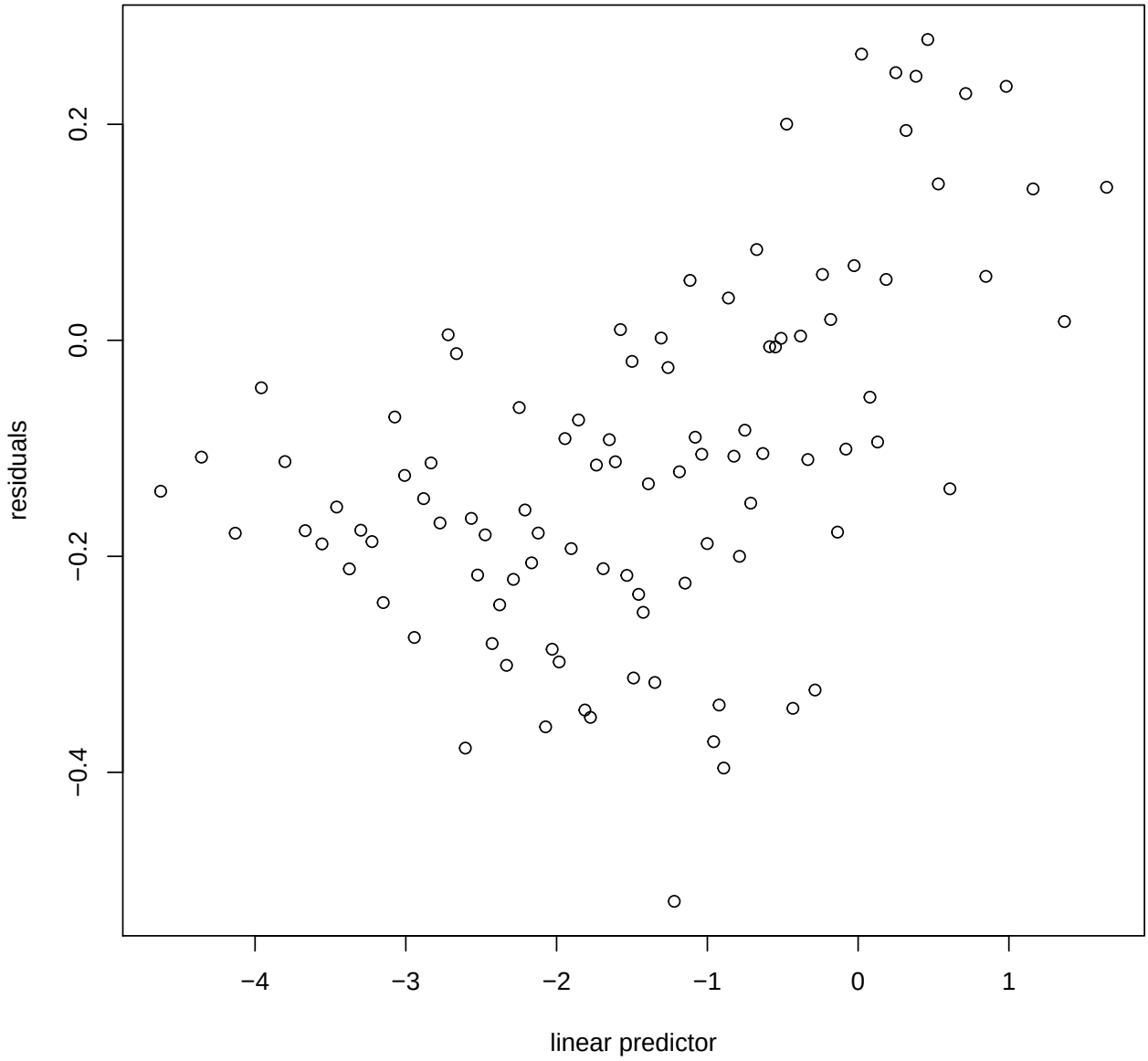

PROC ROC Curve
Sensitivity
1 − Specificity
0.0
0.2
0.4
0.6
0.8
1.0
AUC: 0.813


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

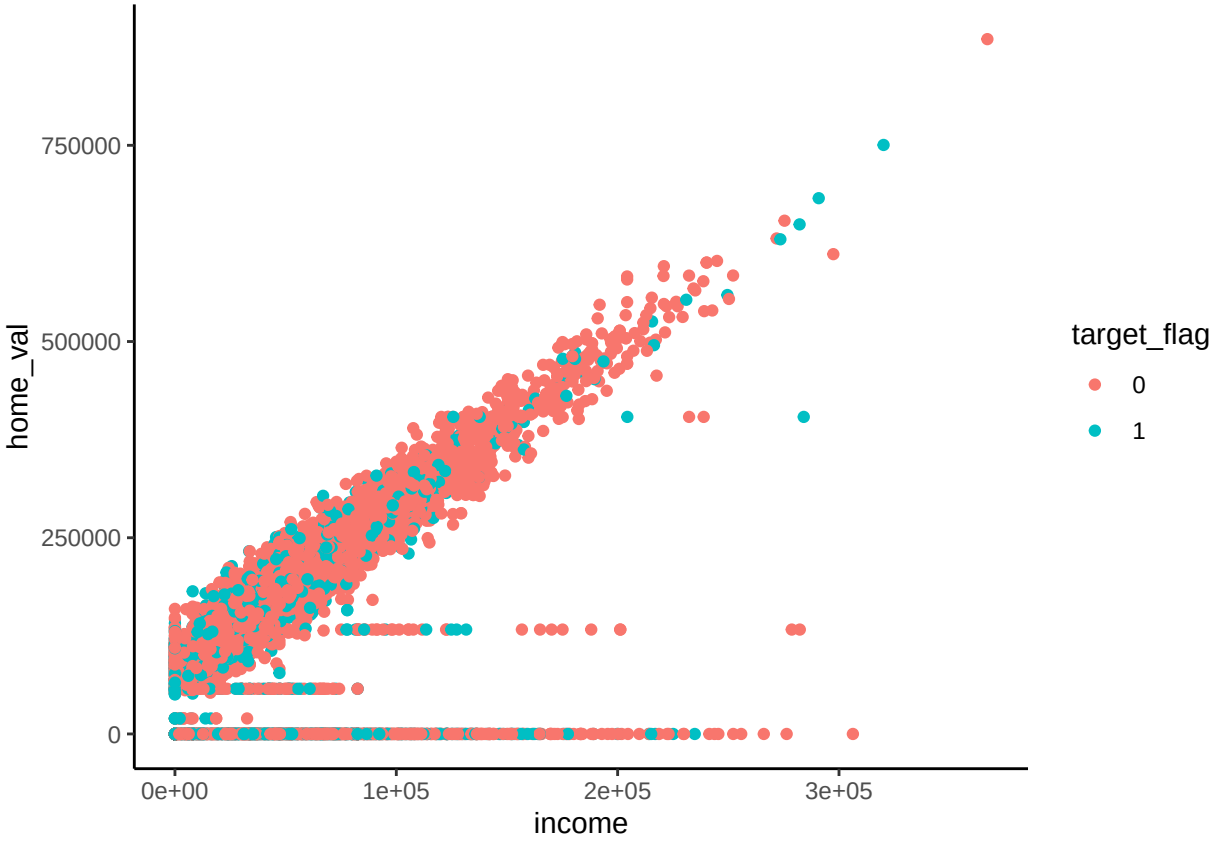
Still seeing pattern - possible mis-specification



Model 3 - Feature engineering and Interactions among predictor variables

Establish new variable of liquidity = (home_val+1)/(income+1) to account for the co-variance that was discovered earlier. Some home_val observations are = 0, likely indicating renters, so +1 is added to the variables to avoid 0 division NaN results.

The below factors are applied in an attempt to capture important groupings noticed in histograms. -mvr_pts= factor("none", "low", "high") -liquidity= factor("low", "high") -tif factor=("low", "moderate", "high") -clm_freq= factor("none","moderate","high")



Includes interactions, transformation (bluebook), factored variables, and feature engineering

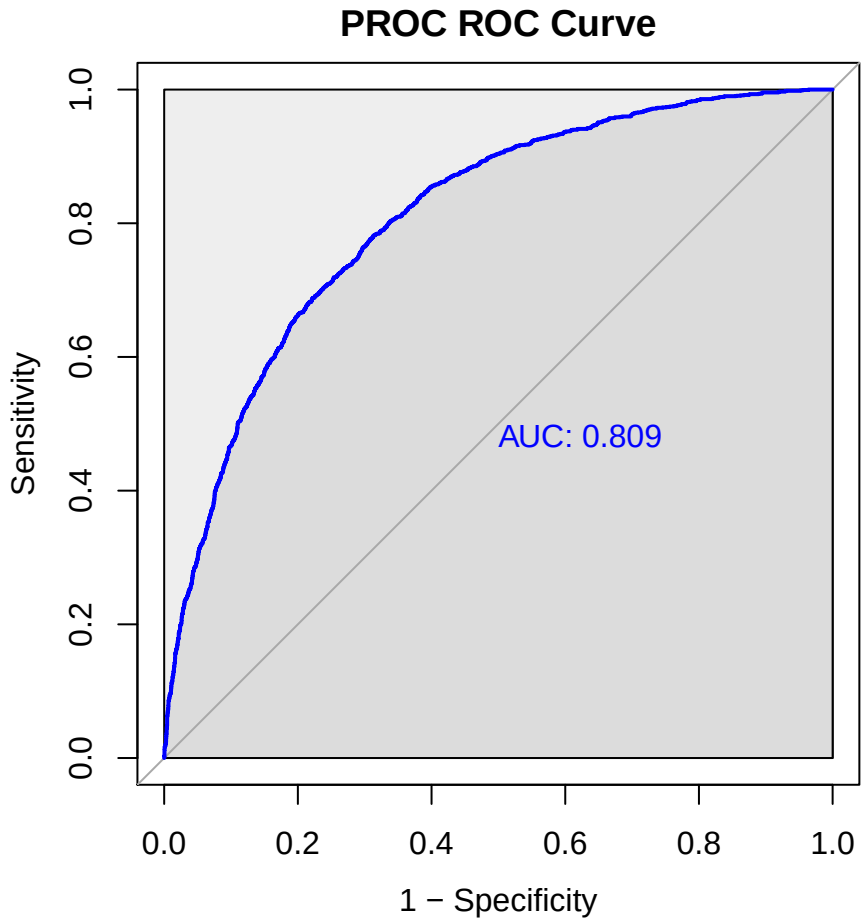
Observations	6120
Dependent variable	target_flag
Type	Generalized linear model
Family	binomial
Link	logit

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

Diagnostics

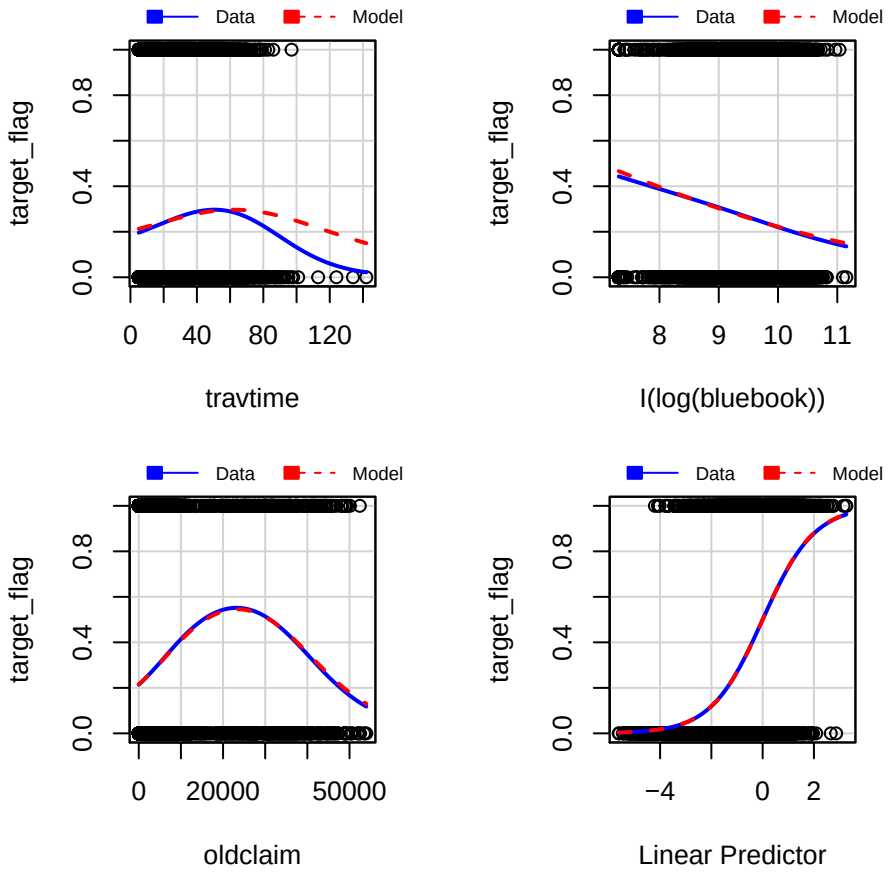
A tibble: 1 x 12

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



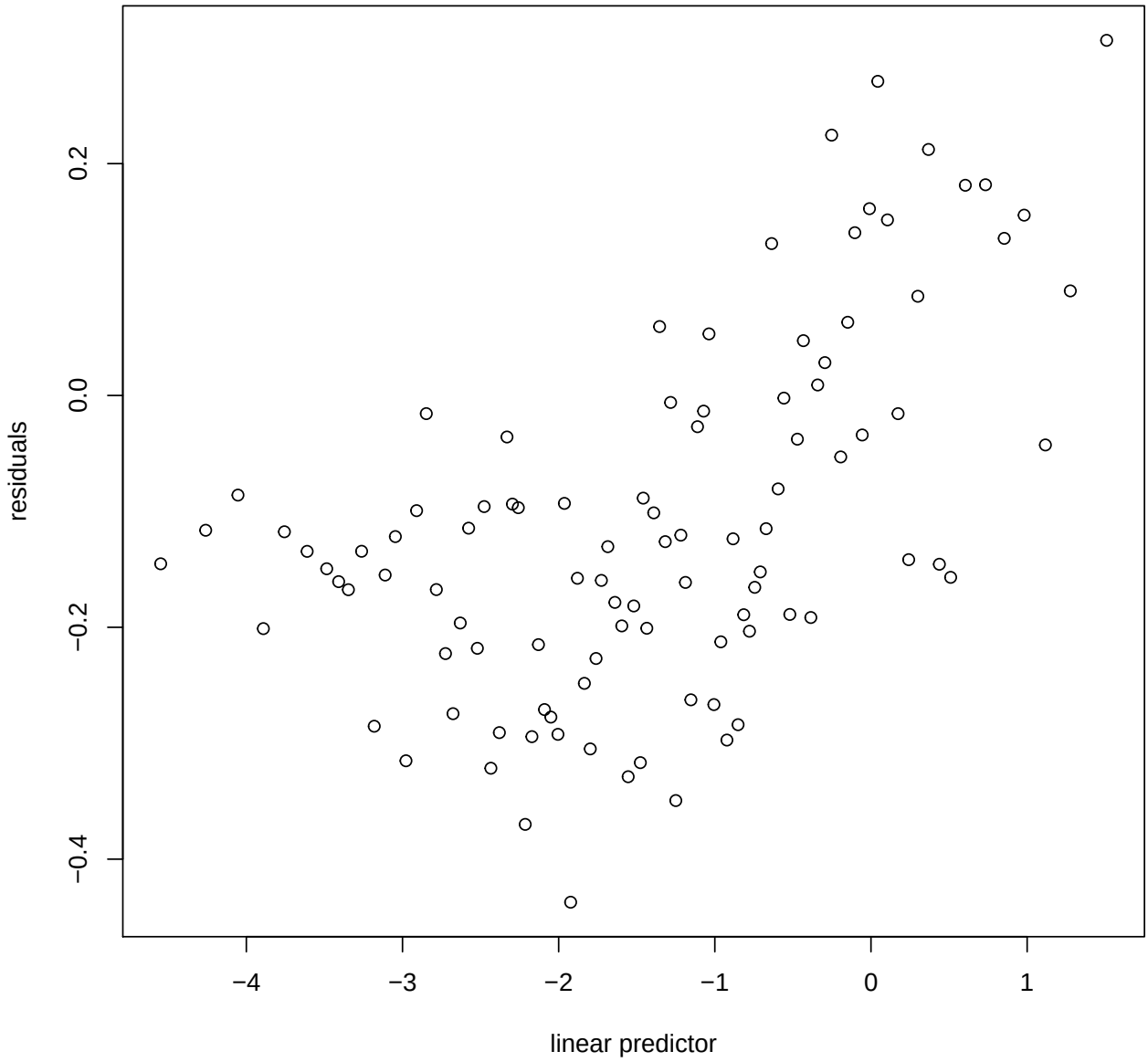
Dispersion

No evidence of significant dispersion

[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 Pearson Chi-Squared test statistic of 0.8729 This communicates that the null hypothesis is not rejected and their are no problems with dispersion."

Independence

We note that the patterns in residuals still suggest mis-specification. Further investigation of the context of the data collection is suggested.



Logistical Classification Model Selection

model performance similar across all cases. Model1 had the highest accuracy. Model2 has the lowest AIC and predictor numbers.

The models do well at predicting no crashes but performs less well at predicting crashes with a .5 threshold. Given the payout risk - a threshold of ~0.3 might be advisable.

model	predictors	sensitivity	specificity	pos_rate	neg_rate	precision	recall	f1	
Base Model: base variables	25	0.9227696	0.4144981	0.8148148	0.6578171	0.8148148	0.9227696	0.8654387	0.
transformation Model: reduced variables	24	0.9238793	0.4250310	0.8177175	0.6666667	0.8177175	0.9238793	0.8675628	0.
Feature_Eng+Transform Model: reduced variables	30	0.9243231	0.3983891	0.8109424	0.6534553	0.8109424	0.9243231	0.8639286	0.

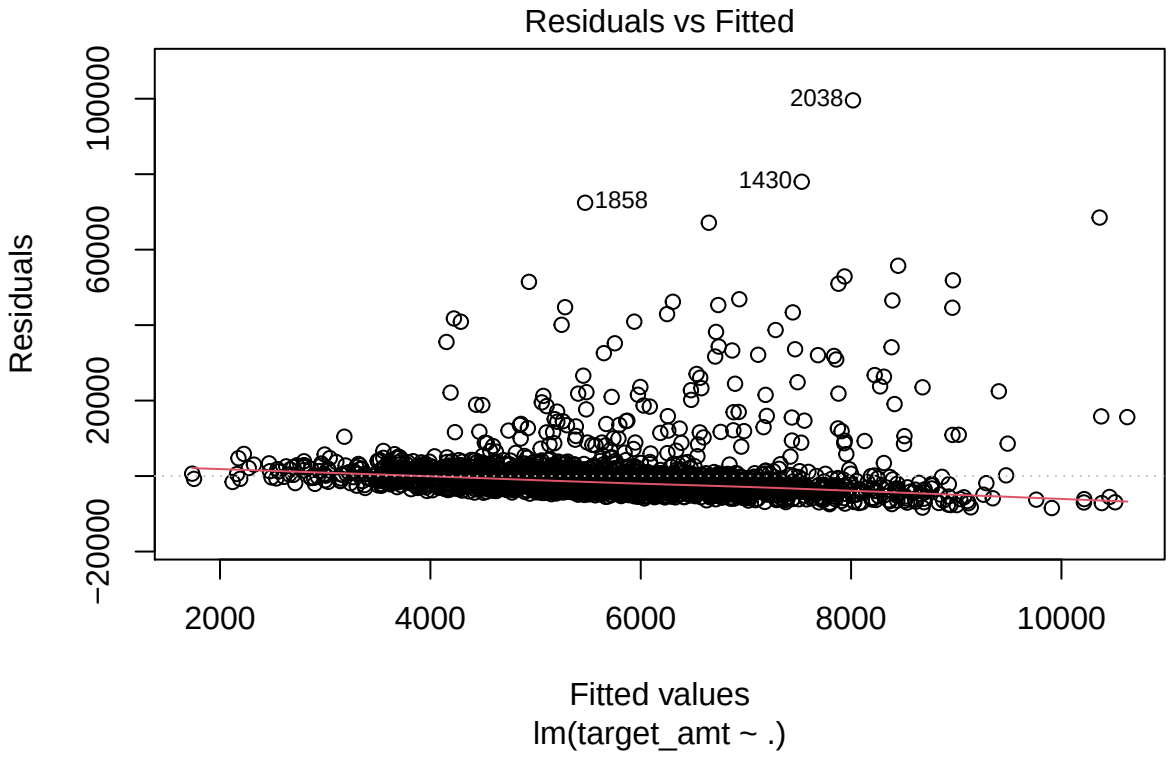
Construct Linear Regression Model

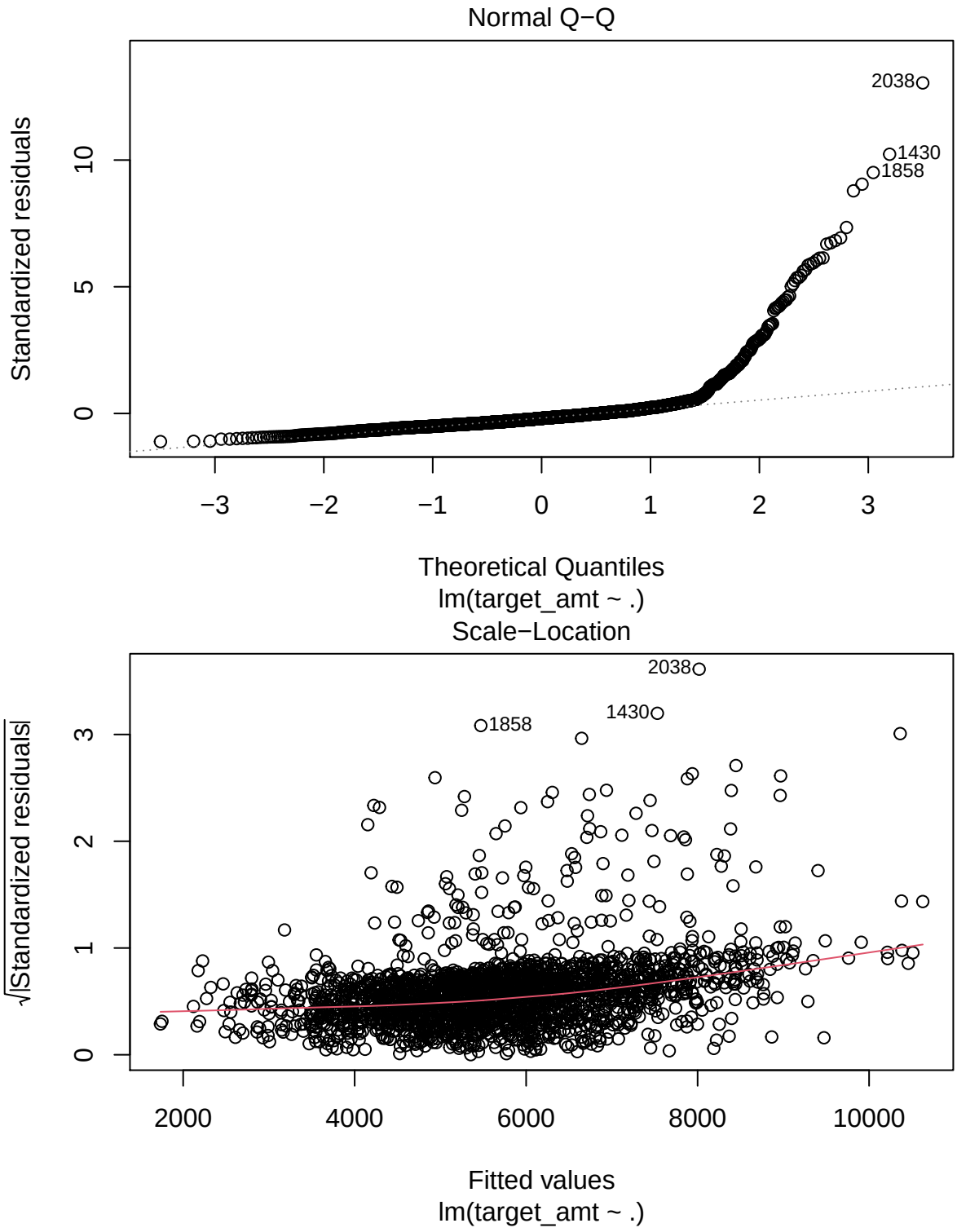
Firstly, we will like to how the saturated model performs under the standard Gaussian assumptions. We find there are only four variables with significant p-values; and the r-squared is very low. Also the residual plots fail the required assumptions regarding the normal distribution and constant variance.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

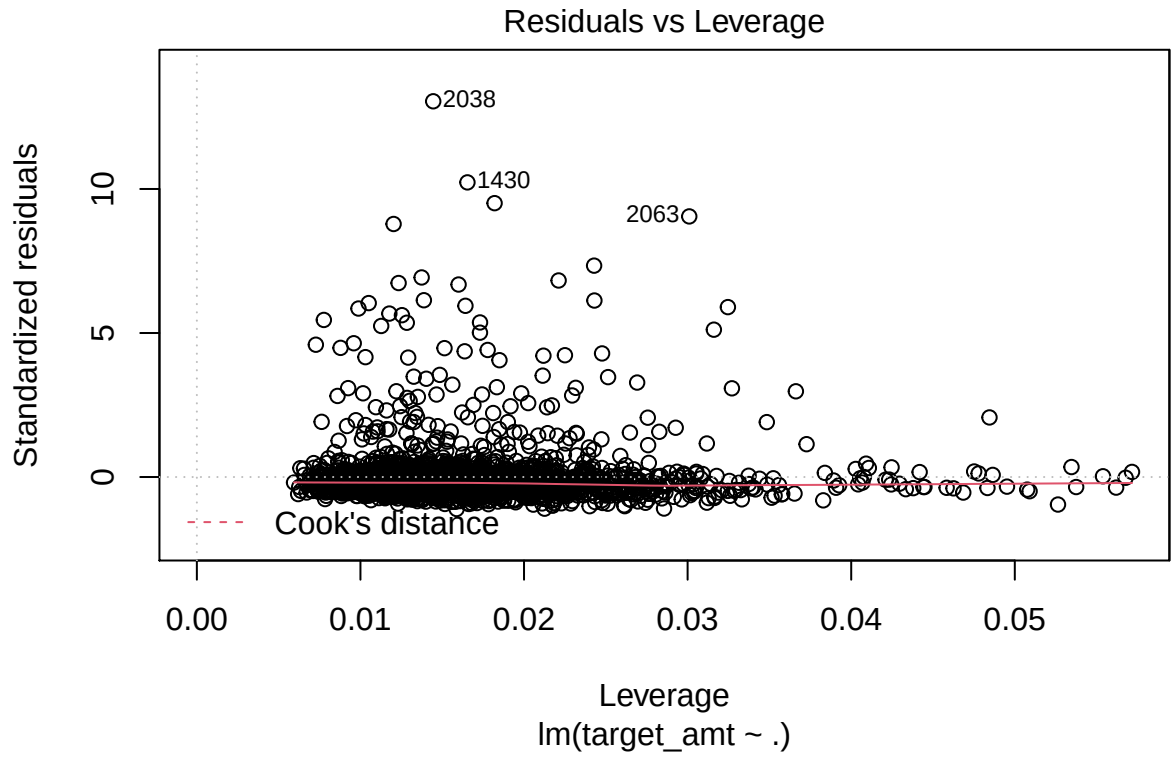
Model 1: Base Linear Model

Observations	2152
Dependent variable	target_amt
Type	OLS linear regression

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

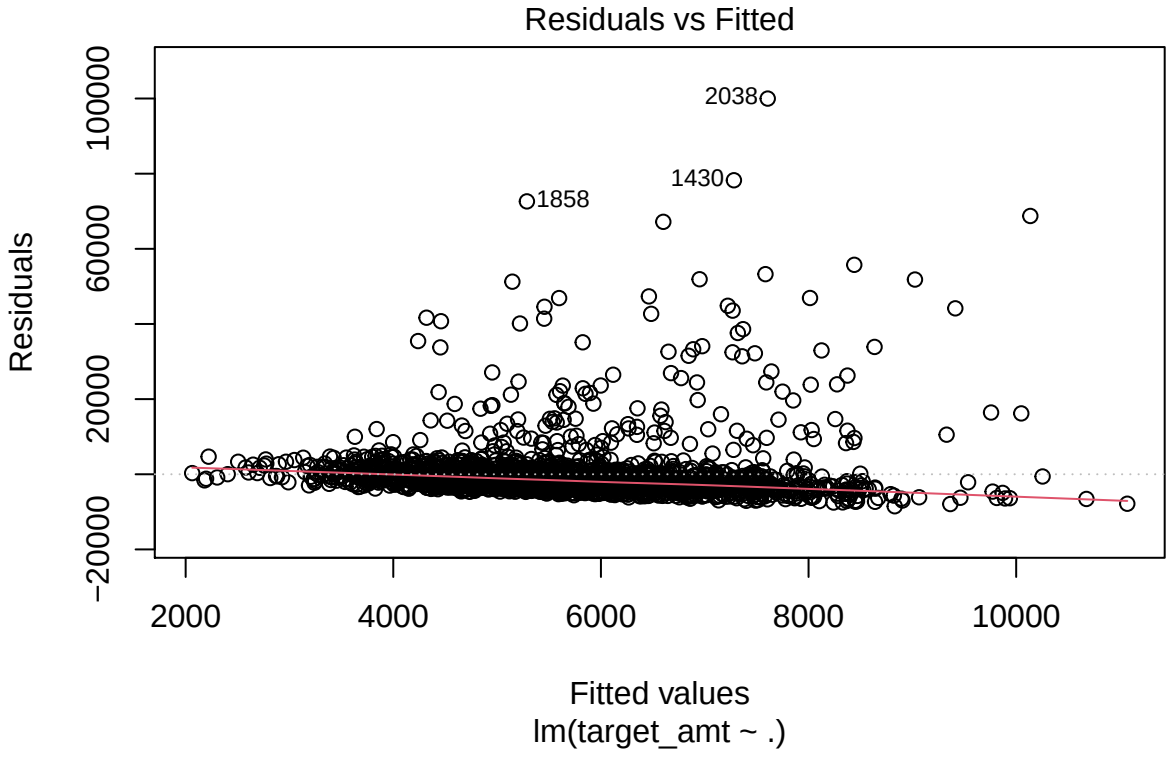


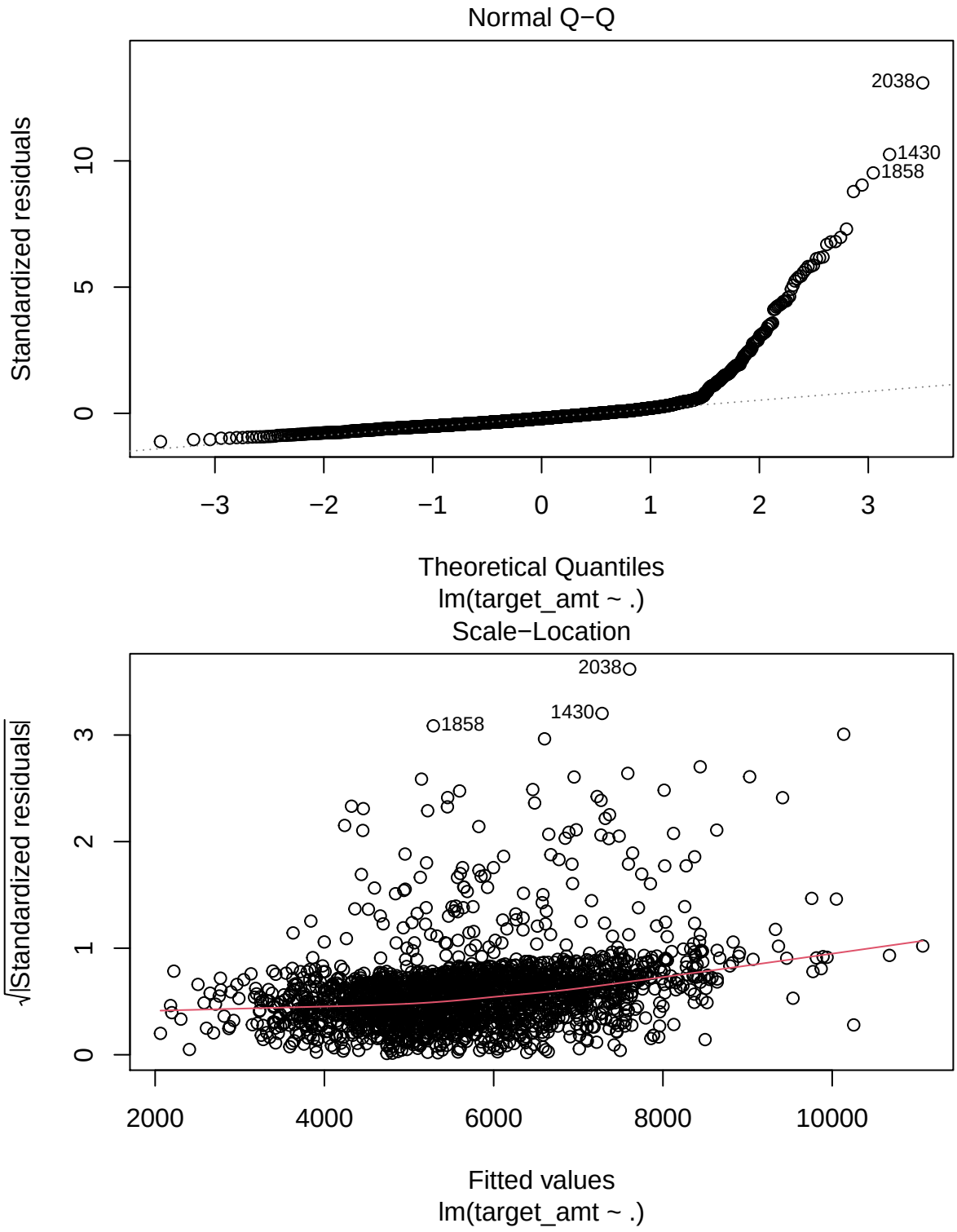

Normal Q−Q
Standardized residuals
2038
1430
1858
Theoretical Quantiles
lm(target_amt ~ .)
Scale−Location
√|Standardized residuals|
2038
1430
1858
Fitted values
lm(target_amt ~ .)

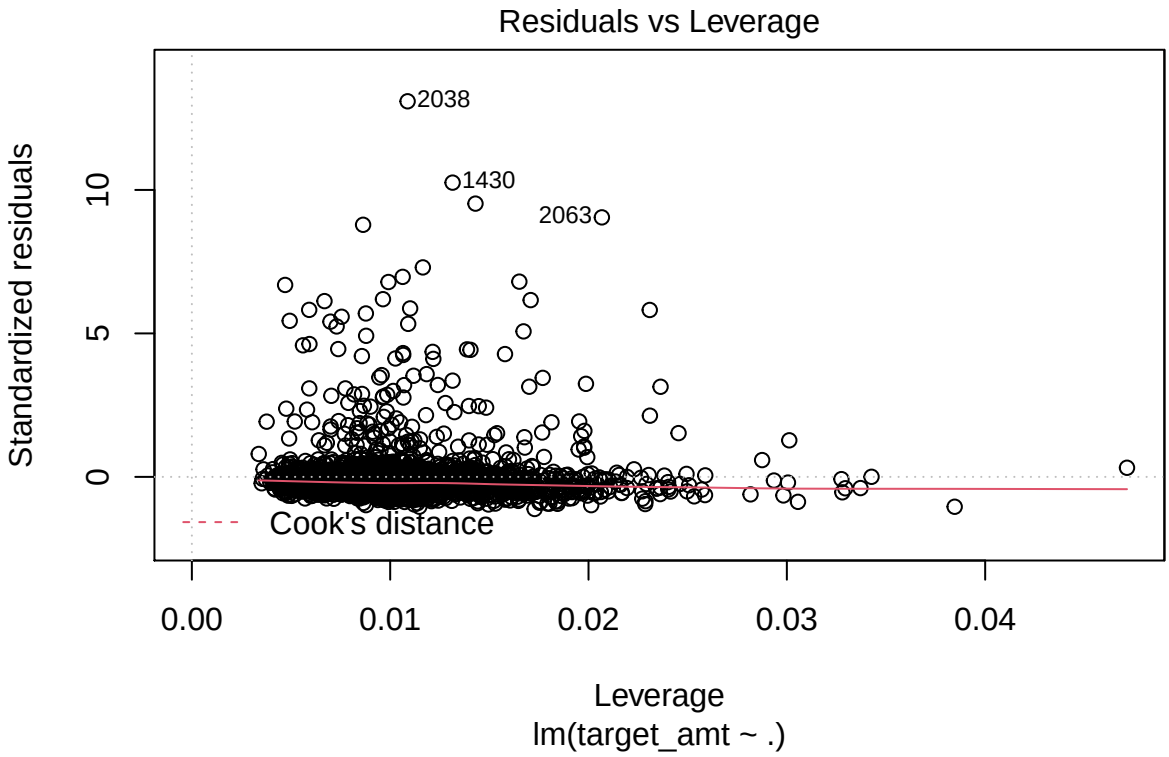


Cost Model 2: Feature Reduction

By removing some of the variables we earmarked earlier in the analysis, we can see a reduction in the Residual Standard Error. -parent1, -age, -homekids, - kidsdriv, -red_car, -urbanicity,-job



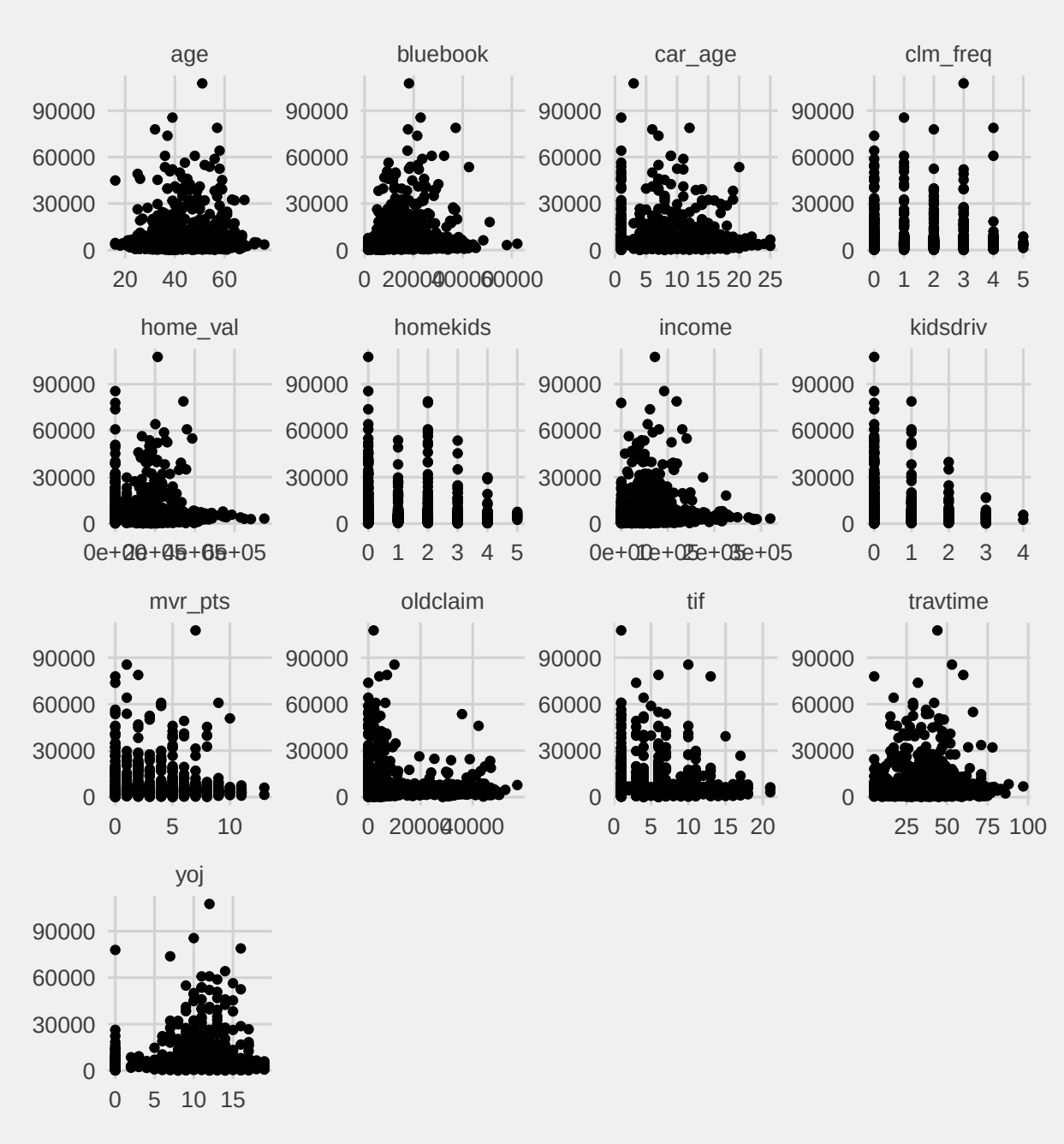

Normal Q−Q
Standardized residuals
2038
1430
1858
Theoretical Quantiles
lm(target_amt ~ .)
Scale−Location
√|Standardized residuals|
2038
1430
1858
Fitted values
lm(target_amt ~ .)


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

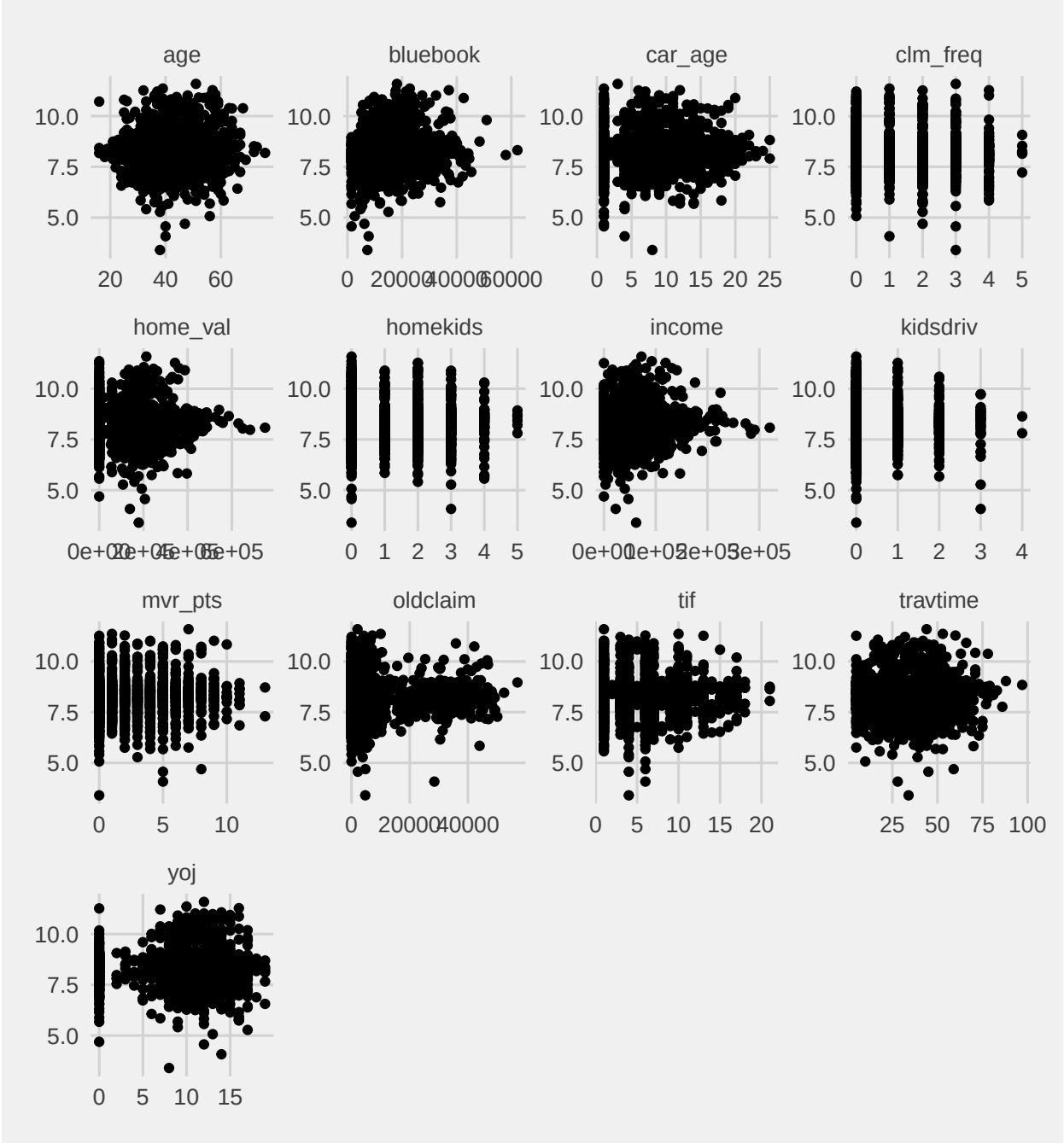
Cost Model 3: Transformation & Weights

We attempt to correct the heteroscadasticity of the residual plots through transformations.

Lets review the linearity from the plots below.

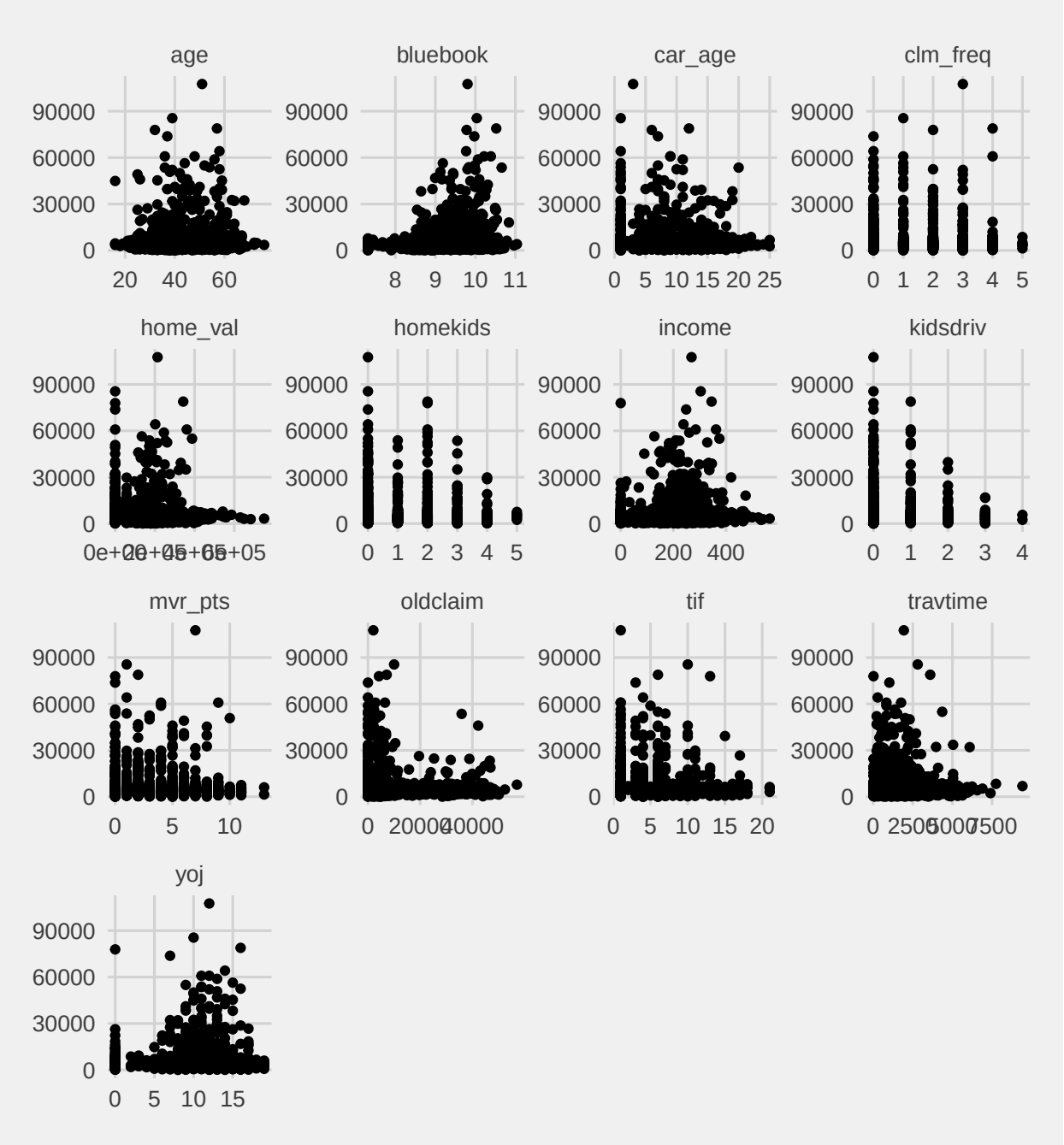


Now with a log transformation on the response. We can see that this evaporates much of the linearity we no-



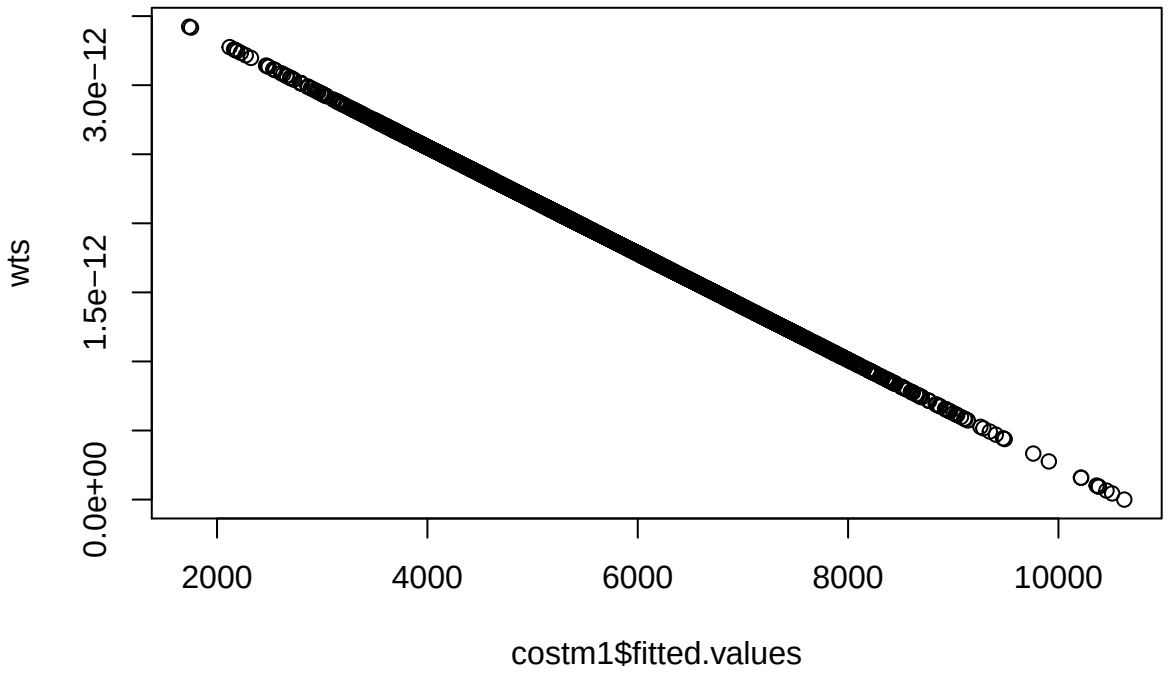
ticed above.

Below is another set of plots that feature the below predictor transformations. We can iterate through additional attempts, but it appears that the relationship sqrt: income log: bluebook quadratic: travtime



The last component to our thrid model is the application of weights. The first step shown below is the calculation these weight coefficients. 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. By taking the absolute value of this regression; we can place less value on the observations with greater variance.

Below is a plot of how the weights 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.

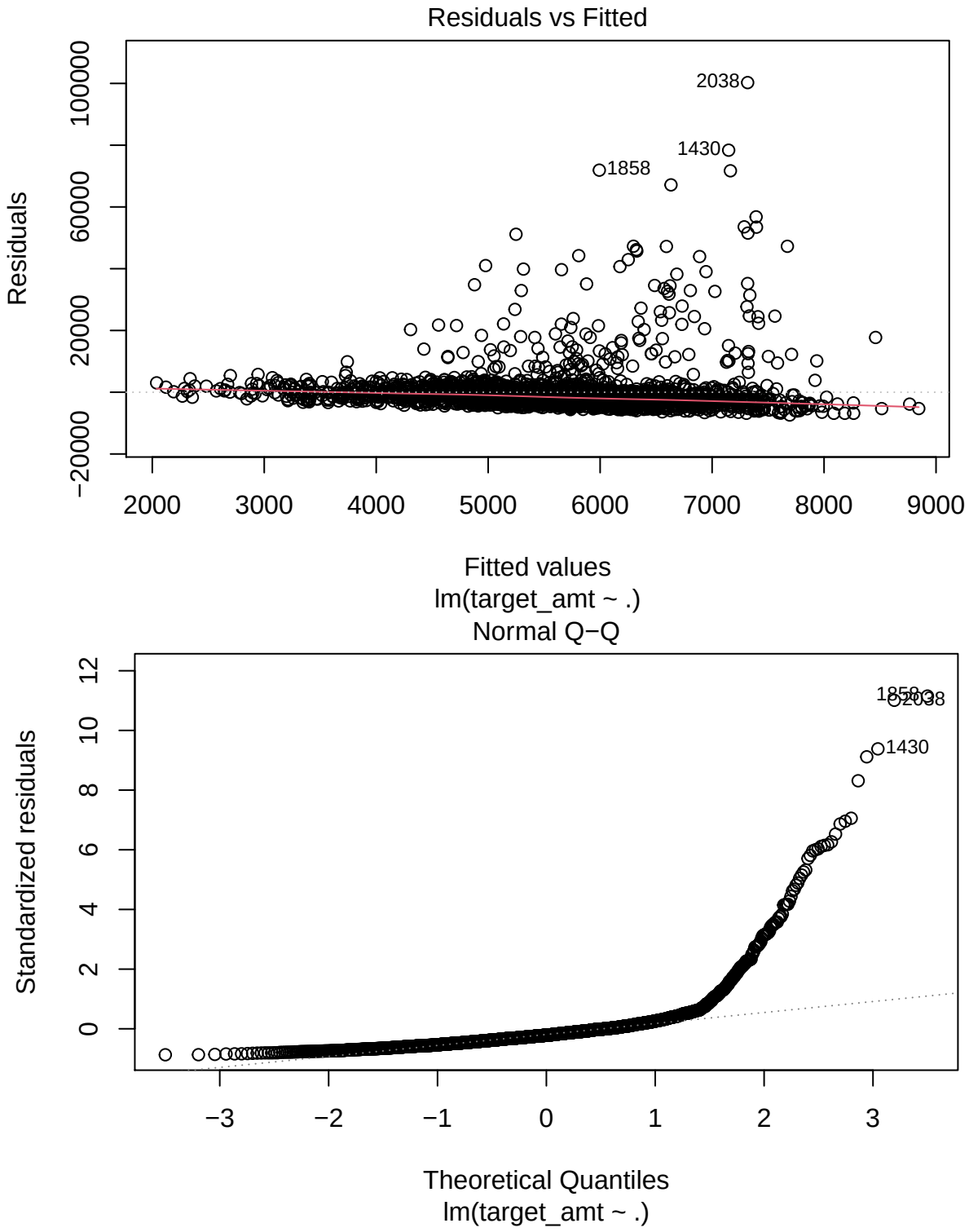
Call: lm(formula = target_amt ~ ., data = dfcrashm3, weights = wts)

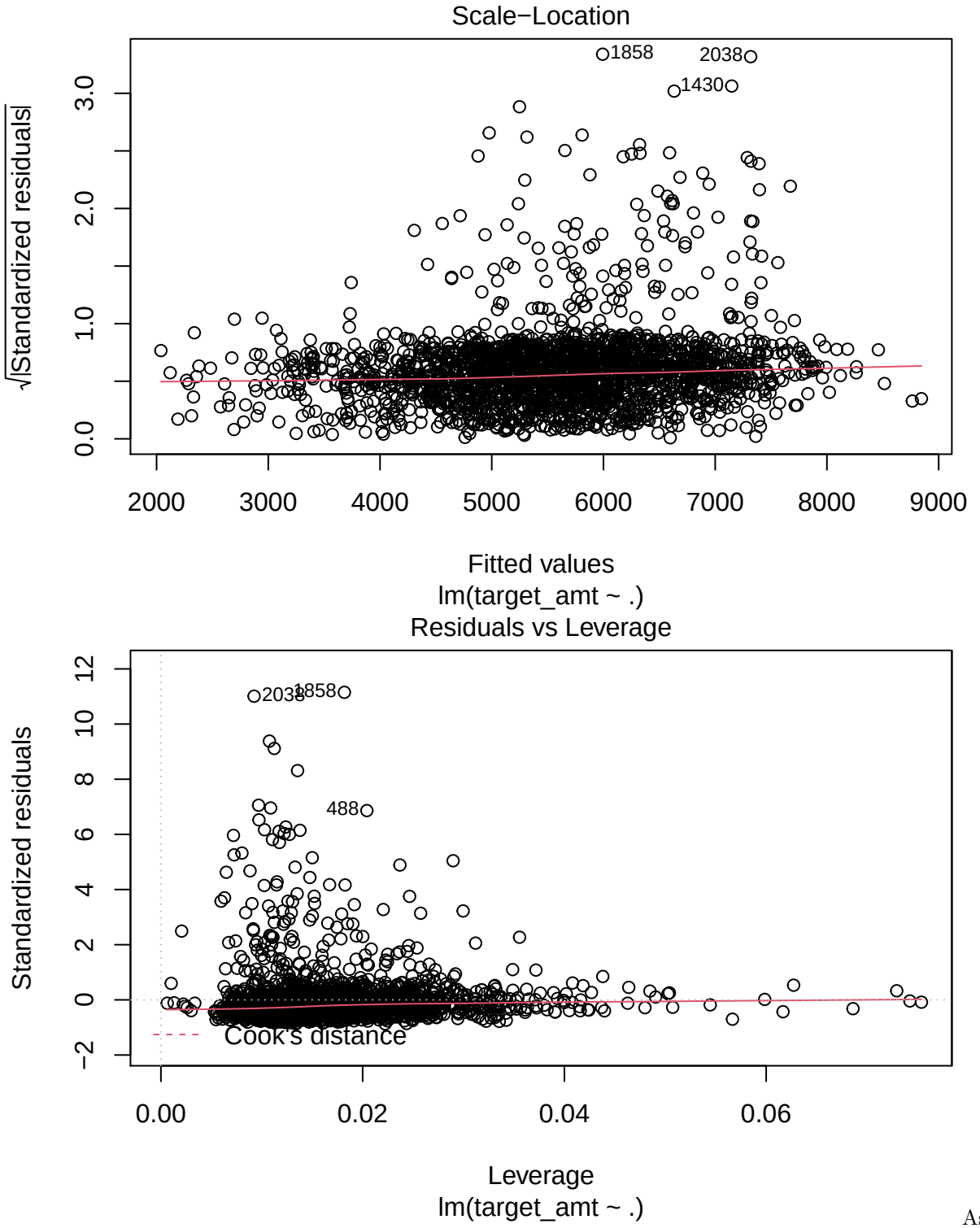
Weighted Residuals: Min 1Q Median 3Q -0.007905 -0.003983 -0.001955 0.000526 Max 0.101315

Coefficients: Estimate (Intercept) -6.433e+03 kidsdriv -1.457e+02 homekids 1.173e+02 parent1Y 7.492e+01 mstatusY -4.046e+02 sexM 8.351e+02 educationBachelors 1.875e+02 educationMasters 7.788e+02 educationPhD 1.687e+03 travtime -7.691e-02 car_usePrivate -1.848e+02 bluebook 1.237e+03 tif -7.845e+00 car_typePanel Truck -1.891e+02 car_typePickup -1.243e+02 car_typeSports Car 7.211e+02 car_typeSUV 4.667e+02 car_typeVan 1.716e+02 red_carY -5.020e+01 oldclaim 2.616e-02 clm_freq -1.899e+02 revokedY -1.043e+03 mvr_pts 9.614e+01 urbanicityUrban 3.184e+02 car_age -6.369e+01 home_val 7.776e-04 yoj 3.150e+01 income -2.587e+00 age 1.019e+01 jobClerical -1.203e+02 jobDoctor -1.594e+03 jobHome Maker -4.970e+02 jobLawyer 5.135e+01 jobManager -1.222e+03 jobProfessional 3.912e+02 jobStudent -2.244e+02 Std. Error t value (Intercept) 2.857e+03 -2.252 kidsdriv 2.723e+02 -0.535 homekids 1.788e+02 0.656 parent1Y 5.166e+02 0.145 mstatusY 4.444e+02 -0.910 sexM 5.649e+02 1.478 educationBachelors 4.420e+02 0.424 educationMasters 7.957e+02 0.979 educationPhD 1.071e+03 1.575 travtime 1.285e-01 -0.598 car_usePrivate 4.208e+02 -0.439 bluebook 2.758e+02 4.484 tif 3.657e+01 -0.215 car_typePanel Truck 8.461e+02 -0.223 car_typePickup 5.058e+02 -0.246 car_typeSports Car 6.316e+02 1.142 car_typeSUV 5.510e+02 0.847 car_typeVan 7.133e+02 0.241 red_carY 4.477e+02 -0.112 oldclaim 1.925e-02 1.359 clm_freq 1.366e+02 -1.390 revokedY 4.262e+02 -2.448 mvr_pts 6.120e+01 1.571 urbanicityUrban 6.463e+02 0.493 car_age 4.043e+01 -1.575 home_val 1.873e-03 0.415 yoj 4.421e+01 0.712 income 3.042e+00 -0.850 age 1.841e+01 0.554 jobClerical 5.061e+02 -0.238 jobDoctor 1.524e+03 -1.046 jobHome Maker 7.969e+02 -0.624 jobLawyer 9.238e+02 0.056 jobManager 6.933e+02 -1.763 jobProfessional 5.950e+02 0.657 jobStudent 6.898e+02 -0.325 Pr(>|t|)
(Intercept) 0.0244 *
kidsdriv 0.5927
homekids 0.5119
parent1Y 0.8847
mstatusY 0.3627
sexM 0.1395
educationBachelors 0.6715
educationMasters 0.3278

educationPhD 0.1153
travtime 0.5496
car_usePrivate 0.6606
bluebook 7.72e-06 ** *tif 0.8302*
car_typePanel Truck 0.8232
car_typePickup 0.8059
car_typeSports Car 0.2537
car_typeSUV 0.3971
car_typeVan 0.8100
red_carY 0.9107
oldclaim 0.1744
clm_freq 0.1646
revokedY 0.0144
mvr_pts 0.1164
urbanicityUrban 0.6223
car_age 0.1154
home_val 0.6780
yoj 0.4763
income 0.3953
age 0.5798
jobClerical 0.8122
jobDoctor 0.2958
jobHome Maker 0.5329
jobLawyer 0.9557
jobManager 0.0781 .
jobProfessional 0.5110
jobStudent 0.7450
— Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009169 on 2115 degrees of freedom Multiple R-squared: 0.02399, Adjusted R-squared: 0.007836 F-statistic: 1.485 on 35 and 2115 DF, p-value: 0.03385


Residuals vs Fitted
Residuals
Fitted values
lm(target_amt ~ .)
2038
1430
1858
−20000
20000
60000
100000
2000
3000
4000
5000
6000
7000
8000
9000
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(target_amt ~ .)
1858
2038
1430
0
2
4
6
8
10
12
−3
−2
−1
0
1
2
3

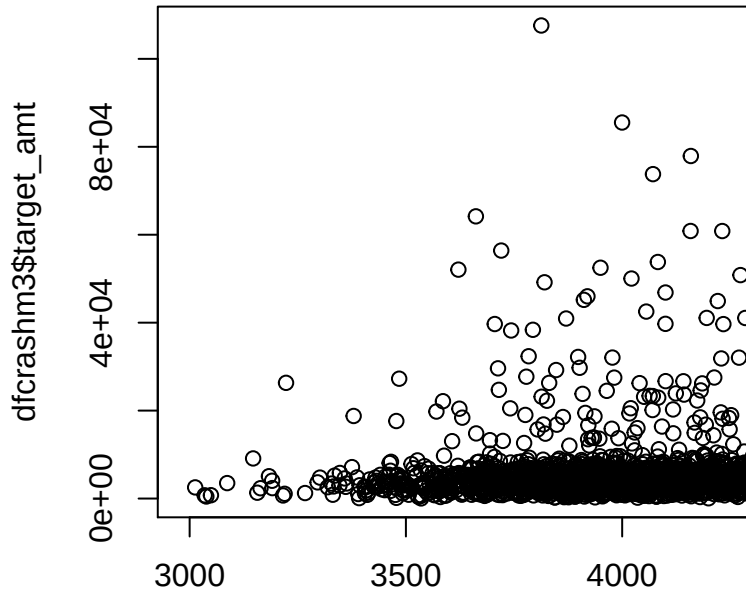


As an alternative approach to weighting the model; we implement a Robust Linear Regression. The function uses Iteratively Reweighted Least Squares(IRLS) to maximize the likelihood estimation. The Residual standard error is reduced again.

Call: rlm(formula = target_amt ~ ., data = dfcrashm3, weights = wts, method = "MM") Residuals: Min 1Q Median -0.0064007 -0.0018924 0.0001064 3Q Max 0.0024489 0.1039949

Coefficients: Value

(Intercept) 1615.9423 kidsdriv -89.1707 homekids 104.6916 parent1Y -319.9883 mstatusY -320.6404 sexM -28.3404 educationBachelors -283.5078 educationMasters 103.1805 educationPhD 192.9389 travtime 0.0074 car_usePrivate 54.6425 bluebook 230.5055 tif 4.9115 car_typePanel Truck 171.8559 car_typePickup 126.1295 car_typeSports Car 51.7437 car_typeSUV 35.3350 car_typeVan 10.3598 red_carY -10.4741 oldclaim 0.0044 clm_freq -83.0664 revokedY 18.1141 mvr_pts 55.3400 urbanicityUrban 287.3838 car_age 7.8953 home_val 0.0007 yoj -2.8463 income -1.1106 age 2.8436 jobClerical -13.6413 jobDoctor -237.7416 jobHome Maker -269.2722 jobLawyer -239.6008 jobManager 32.2356 jobProfessional 121.3543 jobStudent 85.7395 Std. Error (Intercept) 959.1154 kidsdriv 91.4198 homekids 60.0427 parent1Y 173.4189 mstatusY 149.1890 sexM 189.6396 educationBachelors 148.3931 educationMasters 267.1438 educationPhD 359.5515 travtime 0.0431 car_usePrivate 141.2734 bluebook 92.5910 tif 12.2767 car_typePanel Truck 284.0552 car_typePickup 169.8115 car_typeSports Car 212.0460 car_typeSUV 184.9846 car_typeVan 239.4623 red_carY 150.3059 oldclaim 0.0065 clm_freq 45.8690 revokedY 143.0876 mvr_pts 20.5470 urbanicityUrban 216.9792 car_age 13.5732 home_val 0.0006 yoj 14.8421 income 1.0214 age 6.1797 jobClerical 169.9130 jobDoctor 511.7752 jobHome Maker 267.5485 jobLawyer 310.1384 jobManager 232.7372 jobProfessional 199.7532 jobStudent 231.5942 t value
(Intercept) 1.6848 kidsdriv -0.9754 homekids 1.7436 parent1Y -1.8452 mstatusY -2.1492 sexM -0.1494 educationBachelors -1.9105 educationMasters 0.3862 educationPhD 0.5366 travtime 0.1722 car_usePrivate 0.3868 bluebook 2.4895 tif 0.4001 car_typePanel Truck 0.6050 car_typePickup 0.7428 car_typeSports Car 0.2440 car_typeSUV 0.1910 car_typeVan 0.0433 red_carY -0.0697 oldclaim 0.6759 clm_freq -1.8109 revokedY 0.1266 mvr_pts 2.6933 urbanicityUrban 1.3245 car_age 0.5817 home_val 1.1831 yoj -0.1918 income -1.0873 age 0.4602 jobClerical -0.0803 jobDoctor -0.4645 jobHome Maker -1.0064 jobLawyer -0.7726 jobManager 0.1385 jobProfessional 0.6075 jobStudent 0.3702



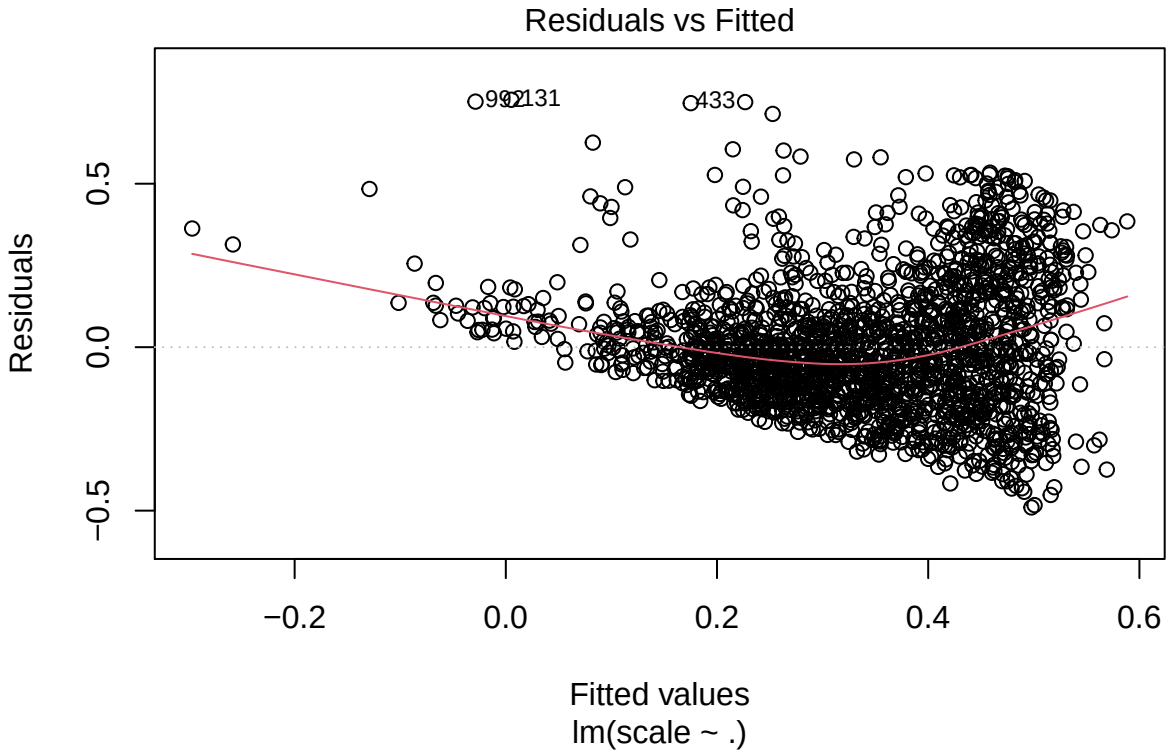
Residual standard error: 0.003109 on 2116 degrees of freedom

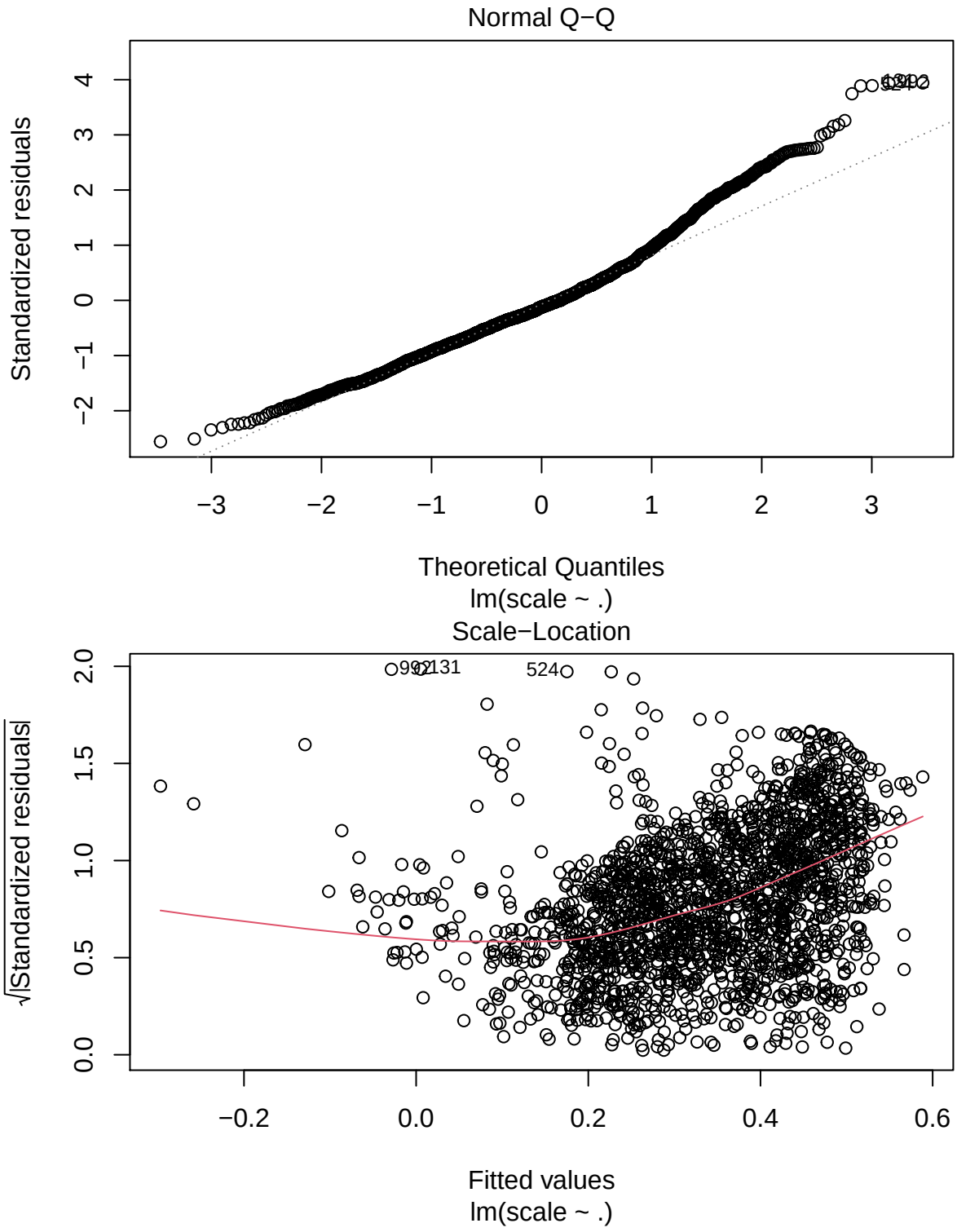
Cost Model 5 Target Interaction Term

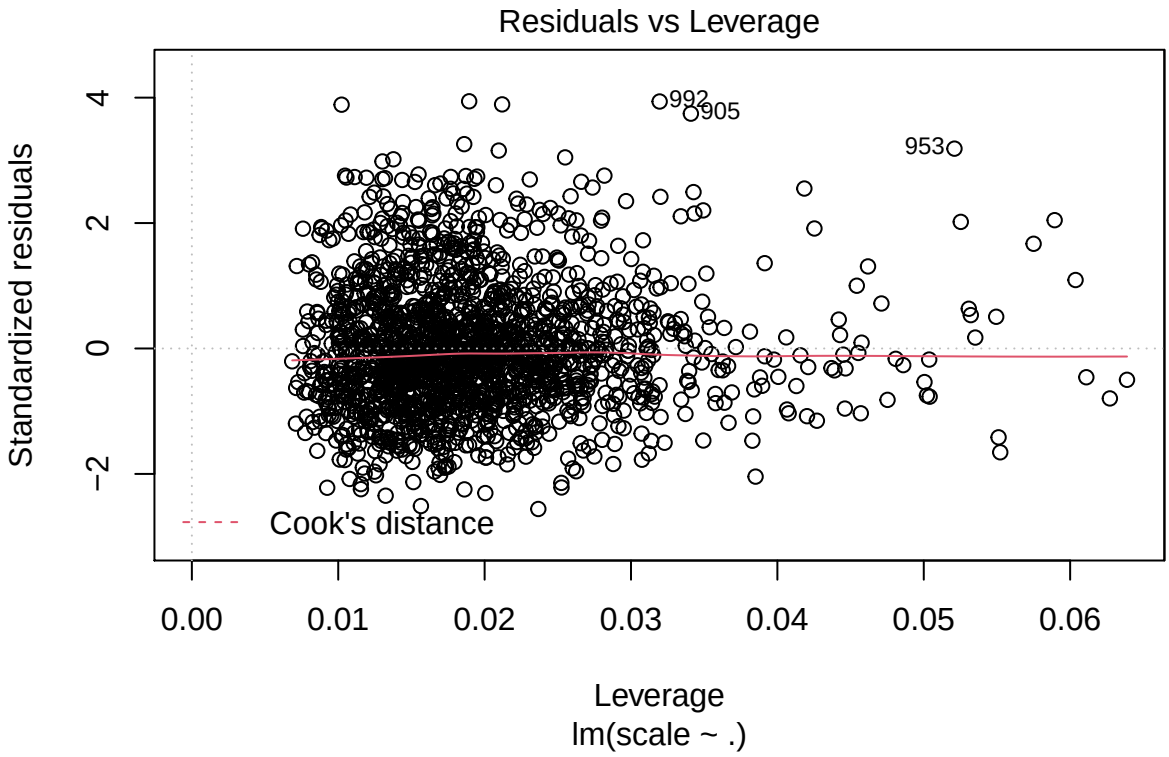
Although there has been some improvements across the above models; the R^2 is still much lower than we can be satisfied with. We now move to rethink the target variable. It stands to reason that the cost of a crash is mostly a function of the value of the car; and the p-values from the above models tell that story. Rather than regressing on the cost, which renders most predictors useless, we regression on intensity of the accident. We can represent that intensity as the 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(scale ~ .)
Scale−Location
√|Standardized residuals|
524
Fitted values
lm(scale ~ .)



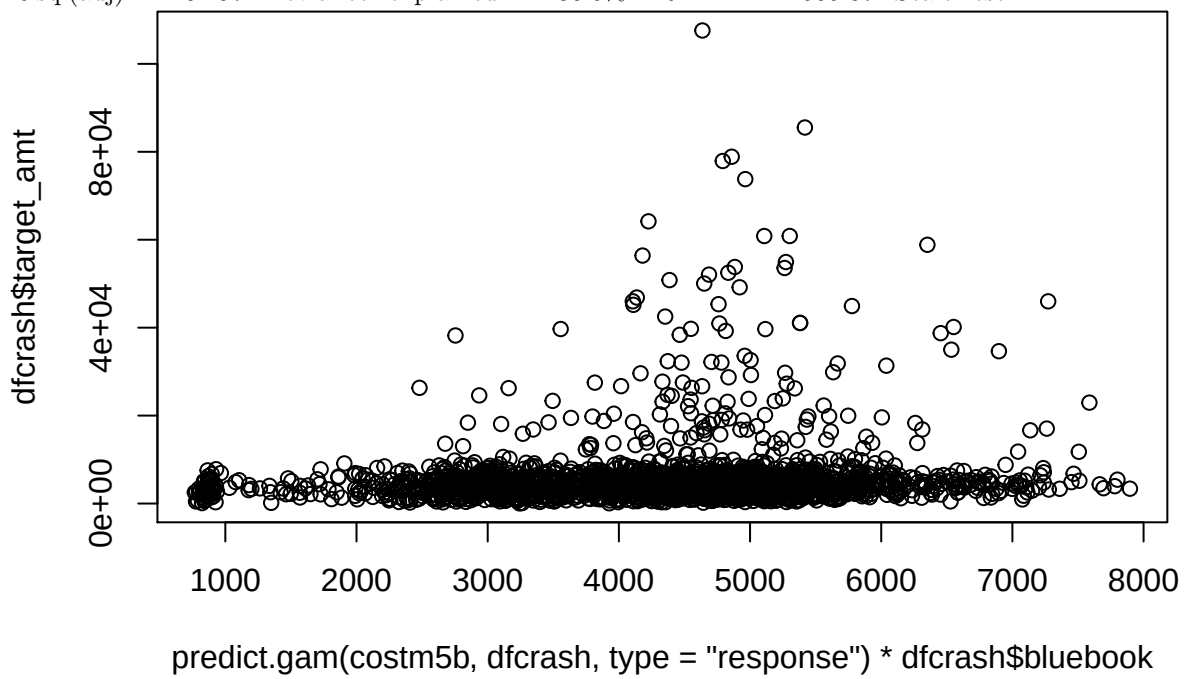
Family: Beta regression(4.732) Link function: logit

Formula: scale ~ age + sex + mstatus + tif + red_car + car_age + home_val + parent1 + mstatus + education + car_use + tif + car_type + oldclaim + revoked + urbanicity + home_val + job + travtime + bluebook + mvr_pts + clm_freq + income

Parametric coefficients: Estimate (Intercept) 4.484e-01 age -4.135e-03 sexM -4.872e-02 mstatusY -7.526e-02 tif -1.092e-03 red_carY 5.826e-02 car_age 2.444e-03 home_val 1.336e-07 parent1Y -1.468e-02 educationBachelors -8.074e-02 educationMasters 6.031e-02 educationPhD 1.661e-01 car_usePrivate 5.756e-02 car_typePanel Truck 4.565e-01 car_typePickup 1.698e-01 car_typeSports Car 1.068e-01 car_typeSUV 7.940e-02 car_typeVan 2.911e-02 oldclaim 3.898e-06 revokedY -3.773e-02 urbanicityUrban 7.898e-02 jobClerical 3.634e-02 jobDoctor 1.597e-01 jobHome Maker 1.304e-02 jobLawyer -1.103e-01 jobManager 1.361e-02 jobProfessional 8.667e-02 jobStudent 1.376e-02 travtime -1.815e-04 bluebook -6.904e-05 mvr_pts 9.131e-03 clm_freq -3.487e-02 income -1.331e-06 Std. Error z value (Intercept) 1.812e-01 2.475 age 2.352e-03 -1.758 sexM 7.807e-02 -0.624 mstatusY 5.678e-02 -1.325 tif 5.116e-03 -0.213 red_carY 5.931e-02 0.982 car_age 5.401e-03 0.453 home_val 2.536e-07 0.527 parent1Y 6.017e-02 -0.244 educationBachelors 6.051e-02 -1.334 educationMasters 1.057e-01 0.570 educationPhD 1.385e-01 1.199 car_usePrivate 5.843e-02 0.985 car_typePanel Truck 1.137e-01 4.016 car_typePickup 7.086e-02 2.396 car_typeSports Car 8.998e-02 1.186 car_typeSUV 7.933e-02 1.001 car_typeVan 9.140e-02 0.318 oldclaim 2.700e-06 1.443 revokedY 6.074e-02 -0.621 urbanicityUrban 9.066e-02 0.871 jobClerical 6.912e-02 0.526 jobDoctor 2.032e-01 0.786 jobHome Maker 9.598e-02 0.136 jobLawyer 1.229e-01 -0.897 jobManager 9.481e-02 0.144 jobProfessional 7.619e-02 1.138 jobStudent 8.016e-02 0.172 travtime 1.324e-03 -0.137 bluebook 3.895e-06 -17.726 mvr_pts 8.215e-03 1.112 clm_freq 1.886e-02 -1.849 income 8.394e-07 -1.585 Pr(>|z|)
(Intercept) 0.0133 *
age 0.0787 .
sexM 0.5326
mstatusY 0.1850
tif 0.8310
red_carY 0.3260
car_age 0.6509
home_val 0.5985

parent1Y 0.8073
educationBachelors 0.1821
educationMasters 0.5684
educationPhD 0.2304
car_usePrivate 0.3245
car_typePanel Truck 5.91e-05 ***car_typePickup 0.0166***
car_typeSports Car 0.2354
car_typeSUV 0.3169
car_typeVan 0.7501
oldclaim 0.1489
revokedY 0.5345
urbanicityUrban 0.3837
jobClerical 0.5991
jobDoctor 0.4319
jobHome Maker 0.8920
jobLawyer 0.3695
jobManager 0.8858
jobProfessional 0.2553
jobStudent 0.8637
travtime 0.8909
bluebook < 2e-16 * mvr_pts 0.2664
clm_freq 0.0644 .
income 0.1129
— Signif. codes:
0 ‘***’ 0.001 ’**’ 0.01 ’*’ 0.05 ‘.’ 0.1 ’ ’ 1

R-sq.(adj) = 0.297 Deviance explained = 30.5% -REML = -505.87 Scale est. = 1 n = 1873



	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	1873
Dependent variable	scale
Type	OLS linear regression

F(35,1837)	21.58
R^2	0.29
Adj. R^2	0.28

	Est.	S.E.	t val.	p
(Intercept)	0.55	0.04	12.79	0.00
kidsdriv	-0.00	0.01	-0.04	0.97
homekids	0.00	0.01	0.65	0.51
parent1Y	-0.00	0.02	-0.22	0.83
mstatusY	-0.02	0.01	-1.49	0.14
sexM	-0.01	0.02	-0.54	0.59
educationBachelors	-0.01	0.01	-1.00	0.32
educationMasters	0.01	0.02	0.58	0.56
educationPhD	0.05	0.03	1.75	0.08
travtime	0.00	0.00	0.24	0.81
car_usePrivate	0.01	0.01	0.82	0.41
bluebook	-0.00	0.00	-17.84	0.00
tif	-0.00	0.00	-0.74	0.46
car_typePanel Truck	0.10	0.03	3.88	0.00
car_typePickup	0.04	0.02	2.66	0.01
car_typeSports Car	0.02	0.02	1.08	0.28
car_typeSUV	0.02	0.02	0.96	0.34
car_typeVan	-0.01	0.02	-0.35	0.73
red_carY	0.01	0.01	0.64	0.52
oldclaim	0.00	0.00	1.62	0.11
clm_freq	-0.01	0.00	-2.16	0.03
revokedY	-0.01	0.01	-0.44	0.66
mvr_pts	0.00	0.00	0.98	0.33
urbanicityUrban	0.02	0.02	0.79	0.43
car_age	-0.00	0.00	-0.02	0.98
home_val	0.00	0.00	0.80	0.42
yoj	0.00	0.00	1.19	0.24
income	-0.00	0.00	-1.49	0.14
age	-0.00	0.00	-1.17	0.24
jobClerical	0.01	0.02	0.88	0.38
jobDoctor	0.02	0.05	0.34	0.74
jobHome Maker	0.01	0.02	0.27	0.79
jobLawyer	-0.02	0.03	-0.85	0.40
jobManager	0.00	0.02	0.03	0.98
jobProfessional	550.02	0.02	0.91	0.37
jobStudent	0.02	0.02	0.99	0.32

Standard errors: OLS