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 summarry from the context of an insurance provider.

In this report we will:

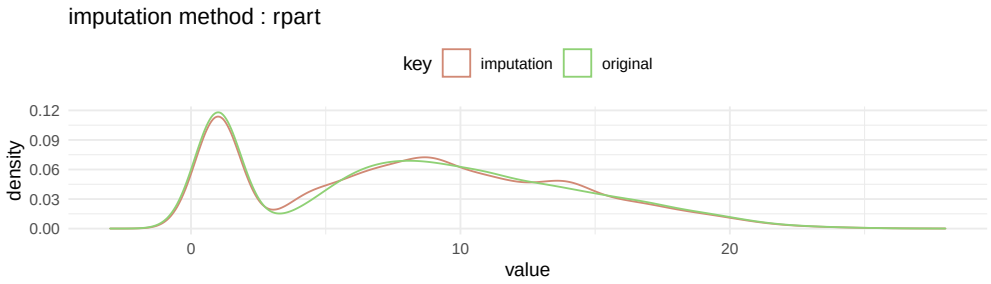
- Explore the data
- Transform data to address multicollenearity and meet variable distribution needs
- Compare different models and select the most accurate model
- Test our model on the evaluation dataset

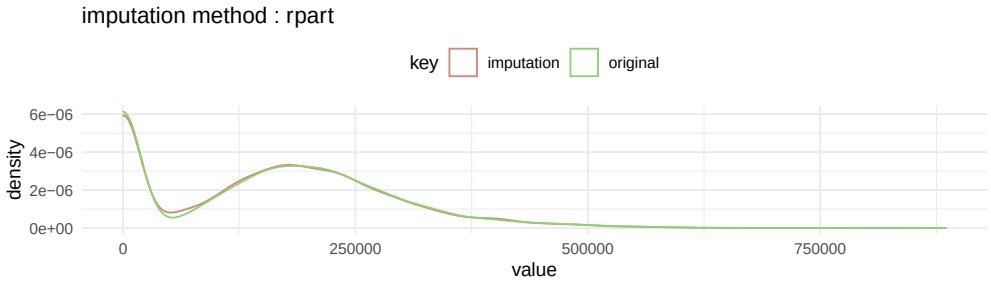
Data cleaning

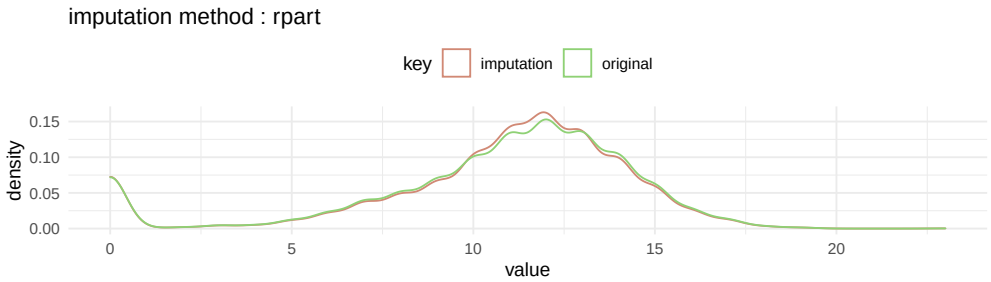
variables	types	missing_count	missing_percent
job	factor	526	6.4452886
car_age	numeric	510	6.2492342
home_val	numeric	464	5.6855777
yoj	numeric	454	5.5630437
income	numeric	445	5.4527631
age	numeric	6	0.0735204

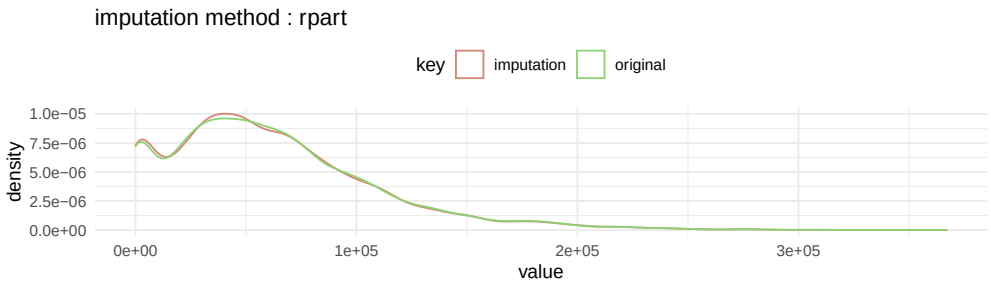
We move to impute the missing data:
Recursive Partitioning and Regression Trees is used to impute the numerical variable.
Multivariate Imputation by Chained Equations is used to impute the categorical variable.

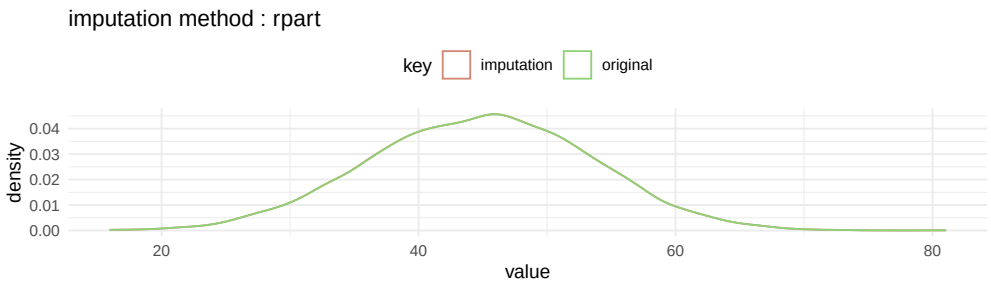
The following plots confirm the imputation follows the nature of the existing data, so we a confident the results our analysis are not affected.

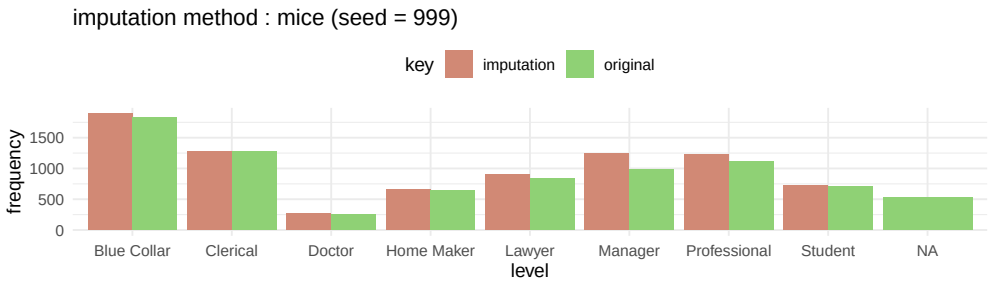

imputation method : rpart
key
imputation
original
density
0.12
0.09
0.06
0.03
0.00
0
10
20
value


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 bimodal distribution, with centers around zero and more normal appearring 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 obviouse 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.

Numerical Distributions [1] "target_amt" "kidsdriv" "homekids" "oldclaim" "clm_freq"
[6] "mvr_pts" "yoj" "income"

vars	statistic	p_value	sample
index	0.955	8.40e-37	5,000
target_amt	0.337	2.25e-86	5,000
kidsdriv	0.370	3.10e-85	5,000
homekids	0.679	7.22e-71	5,000
travtime	0.980	9.91e-26	5,000
bluebook	0.961	1.16e-34	5,000
tif	0.886	2.03e-51	5,000
oldclaim	0.515	1.64e-79	5,000
clm_freq	0.711	9.08e-69	5,000
mvr_pts	0.798	9.71e-62	5,000
car_age	0.939	3.13e-41	5,000
home_val	0.923	5.74e-45	5,000
yoj	0.871	1.36e-53	5,000
income	0.922	4.90e-45	5,000
age	0.998	5.36e-05	5,000

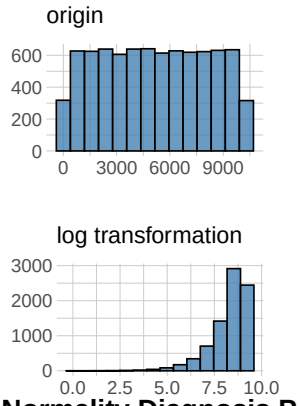
variables	levels	N	freq	ratio	rank
target_flag	0	8,161	6,008	73.618429	1
target_flag	1	8,161	2,153	26.381571	2
parent1	N	8,161	7,084	86.803088	1
parent1	Y	8,161	1,077	13.196912	2
mstatus	Y	8,161	4,894	59.968141	1
mstatus	N	8,161	3,267	40.031859	2
sex	F	8,161	4,375	53.608626	1

variables	levels	N	freq	ratio	rank
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

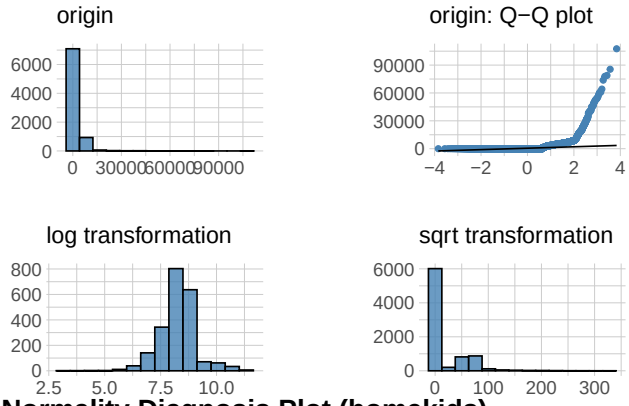
variables	min	mean	median	max	zero	minus
index	1	5,151.8676633	5,133	10,302	0	0
target_amt	0	1,504.3248376	0	107,586	6,008	0
kidsdriv	0	0.1710575	0	4	7,180	0

variables	min	mean	median	max	zero	minus
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	-3	8.3439529	8	28	3	1
home_val	0	154,903.496997	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

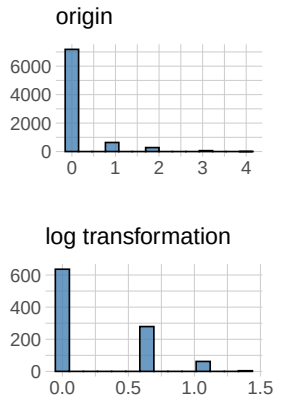
Normality Diagnosis Plot (index)



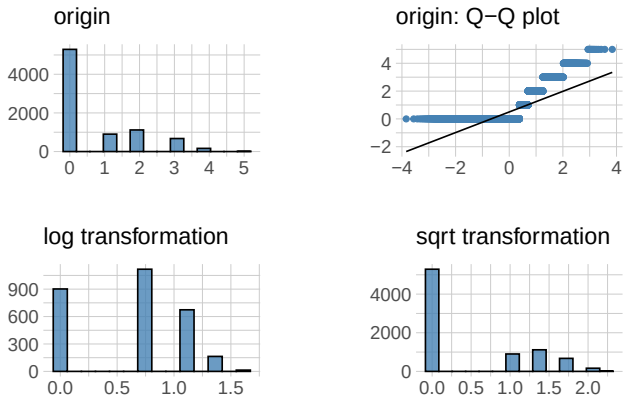
Normality Diagnosis Plot (target_amt)



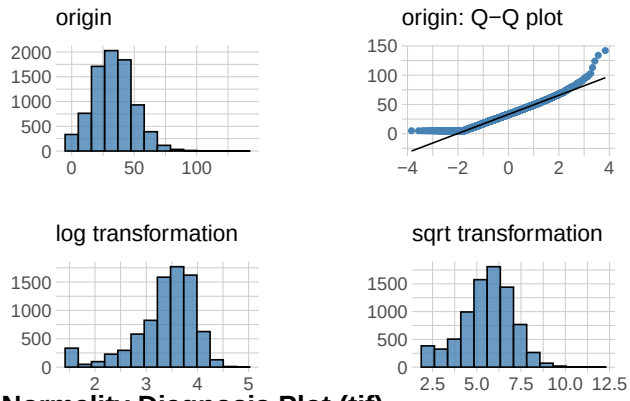
Normality Diagnosis Plot (kidsdriv)



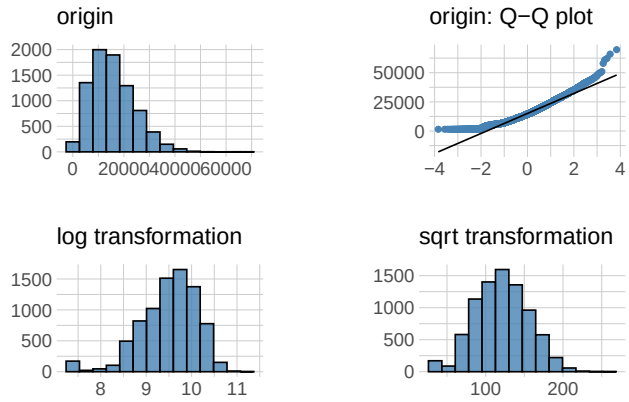
Normality Diagnosis Plot (homekids)



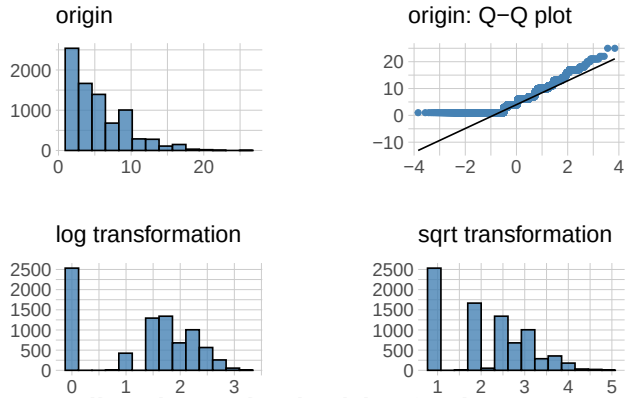
Normality Diagnosis Plot (travtime)



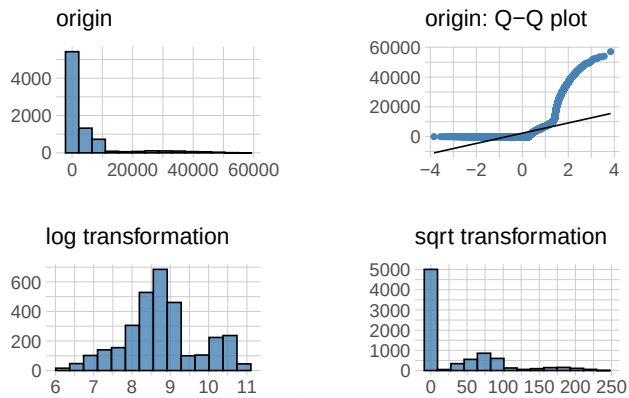
Normality Diagnosis Plot (bluebook)



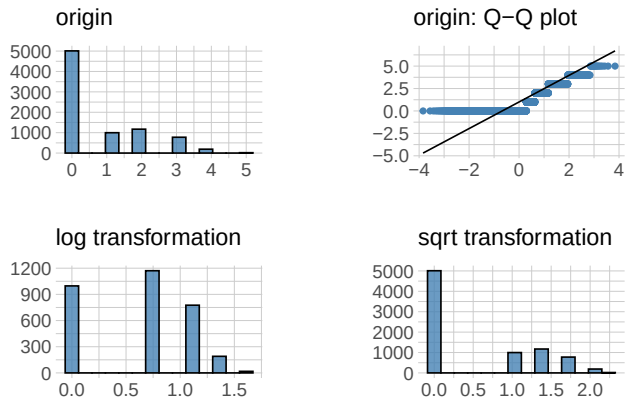
Normality Diagnosis Plot (tif)



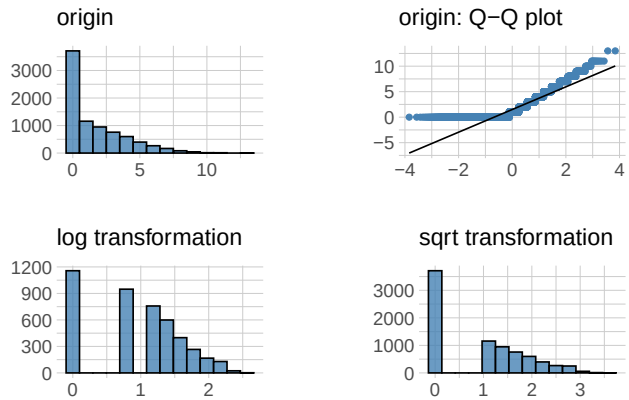
Normality Diagnosis Plot (oldclaim)

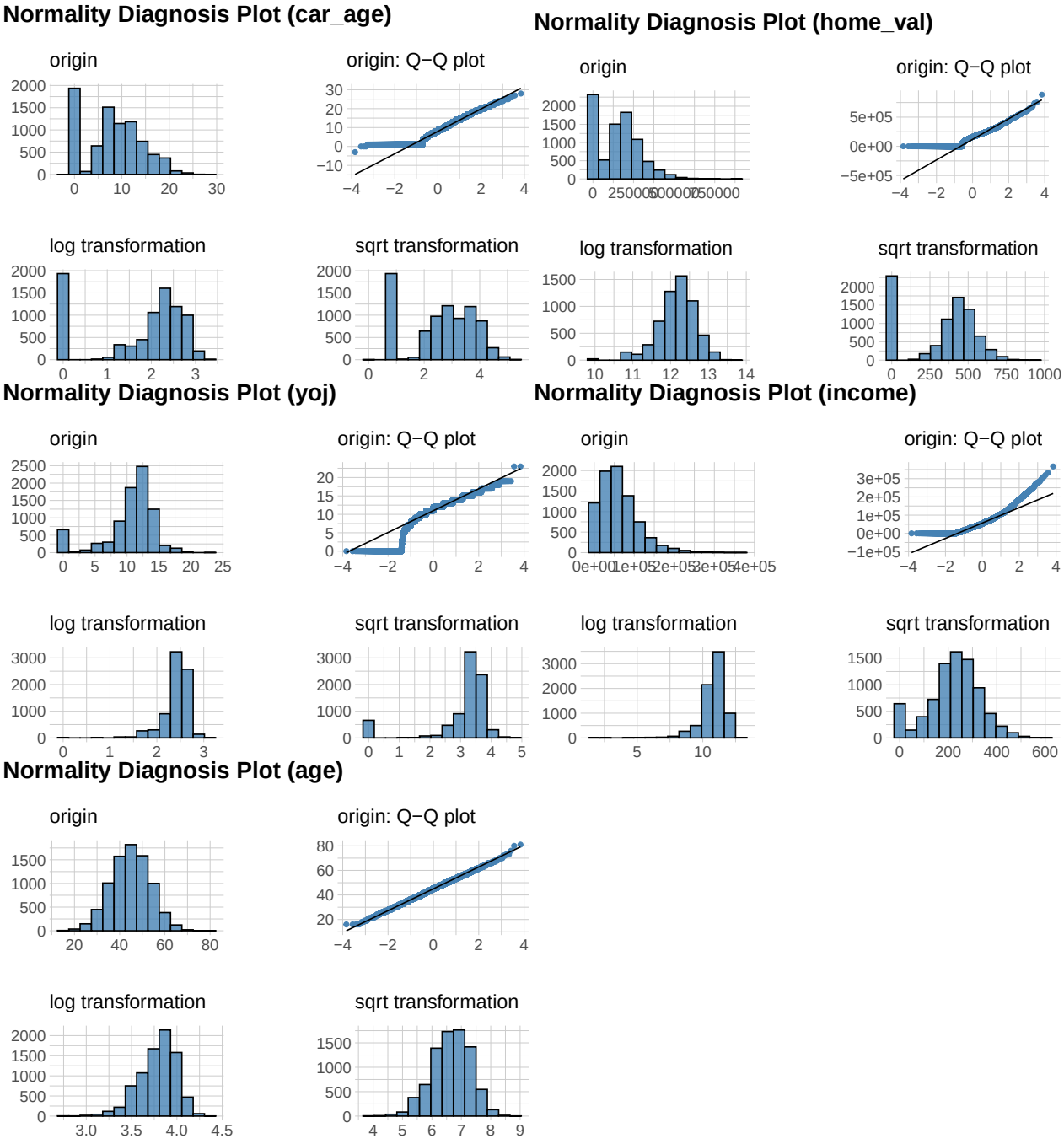


Normality Diagnosis Plot (clm_freq)



Normality Diagnosis Plot (mvr_pts)





Outliers

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

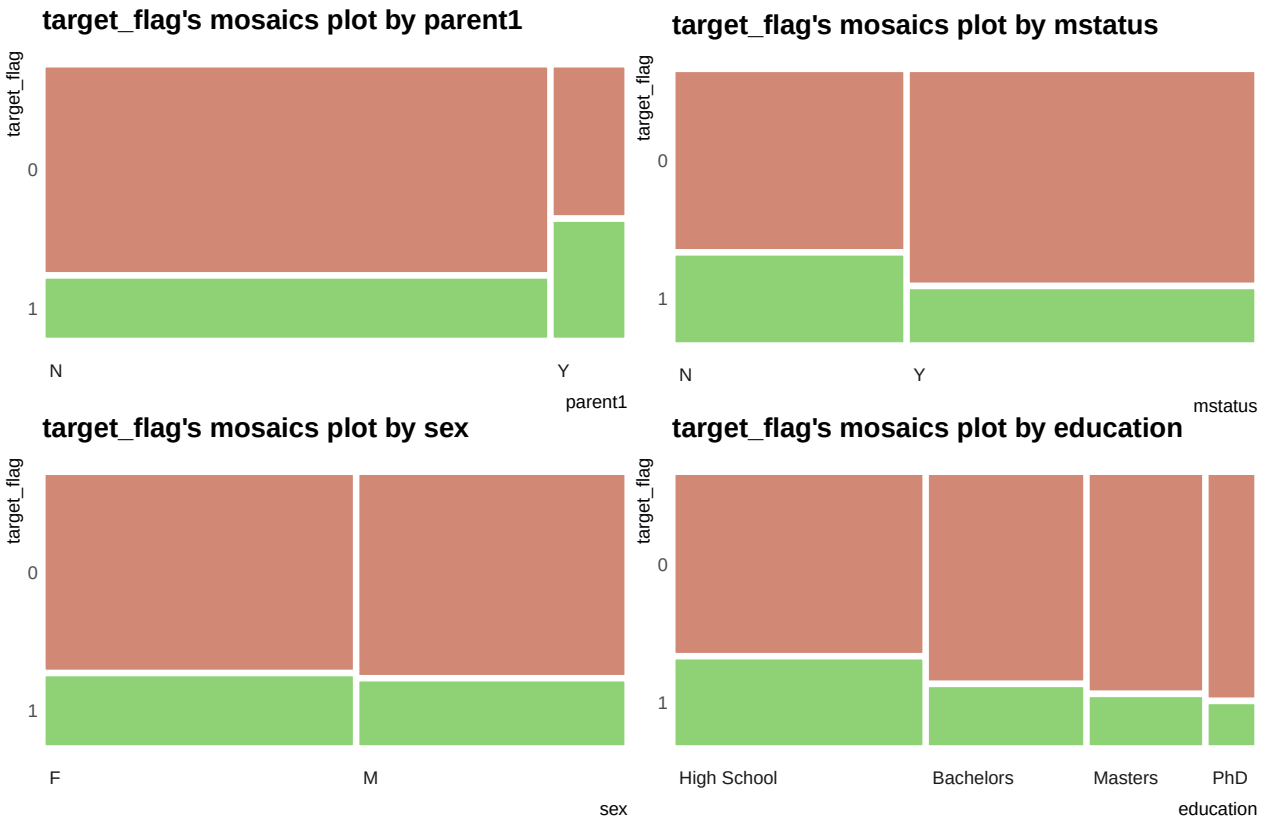
variables	outliers_cnt	outliers_ratio	outliers_mean	with_mean	without_mean
target_amt	1,620	19.851	7,039.976	1,504.325	133.318
kidsdriv	981	12.021	1.423	0.171	0.000
homekids	852	10.440	3.225	0.721	0.429

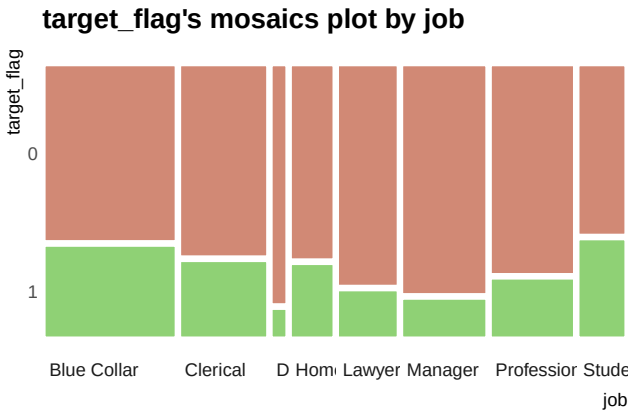
variables	outliers_cnt	outliers_ratio	outliers_mean	with_mean	without_mean
yoj	682	8.357	0.120	10.517	11.465
oldclaim	663	8.124	30,358.611	4,037.076	1,709.632
income	275	3.370	204,853.655	61,501.398	56,502.428
tif	160	1.961	17.869	5.351	5.101
mvr_pts	155	1.899	8.735	1.696	1.559
bluebook	104	1.274	42,806.442	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
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

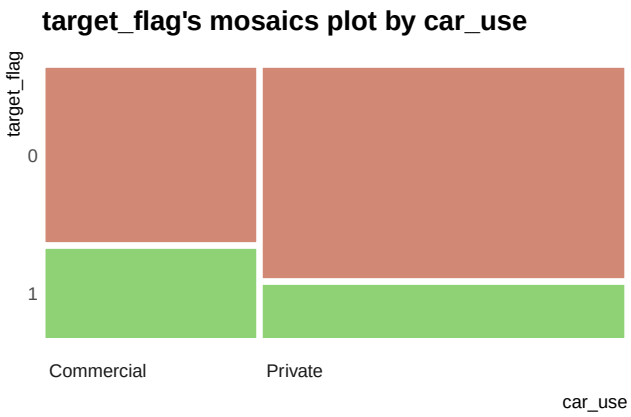
Explore relationships between response and categorical predictors

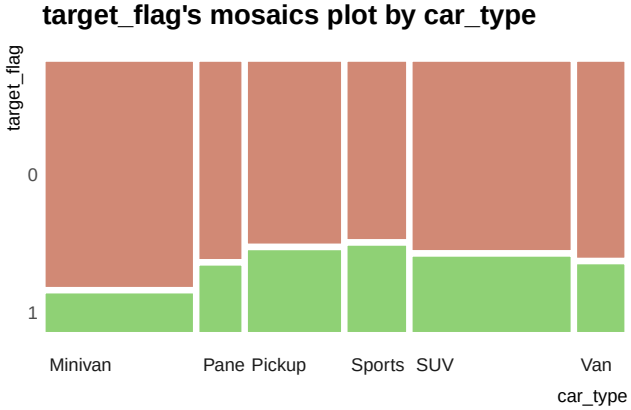
Upon reviewing the below mosaic and box plots, we can determine that the below listed variables have hardly any relationship with the response. This will be kept in mind during the variable selection phase. 'sex' 'red_car'

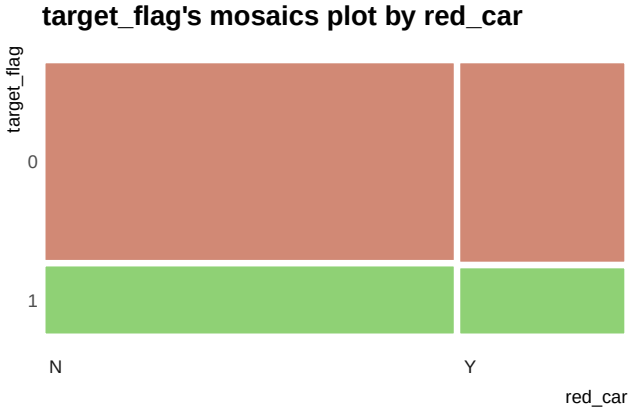
Mosaic plots

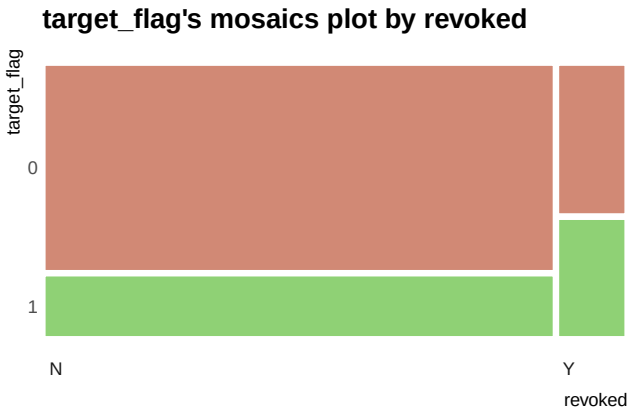


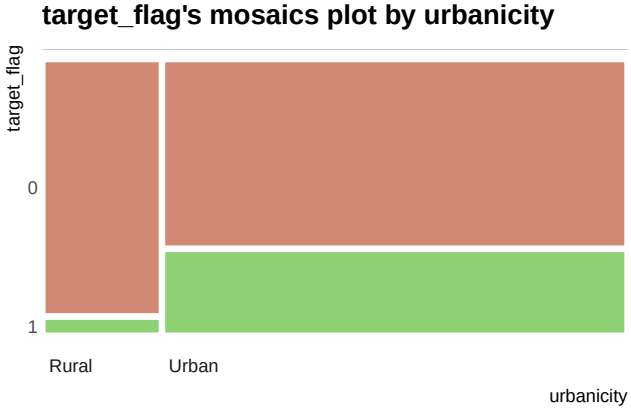












Box plot for numerical variables

$1

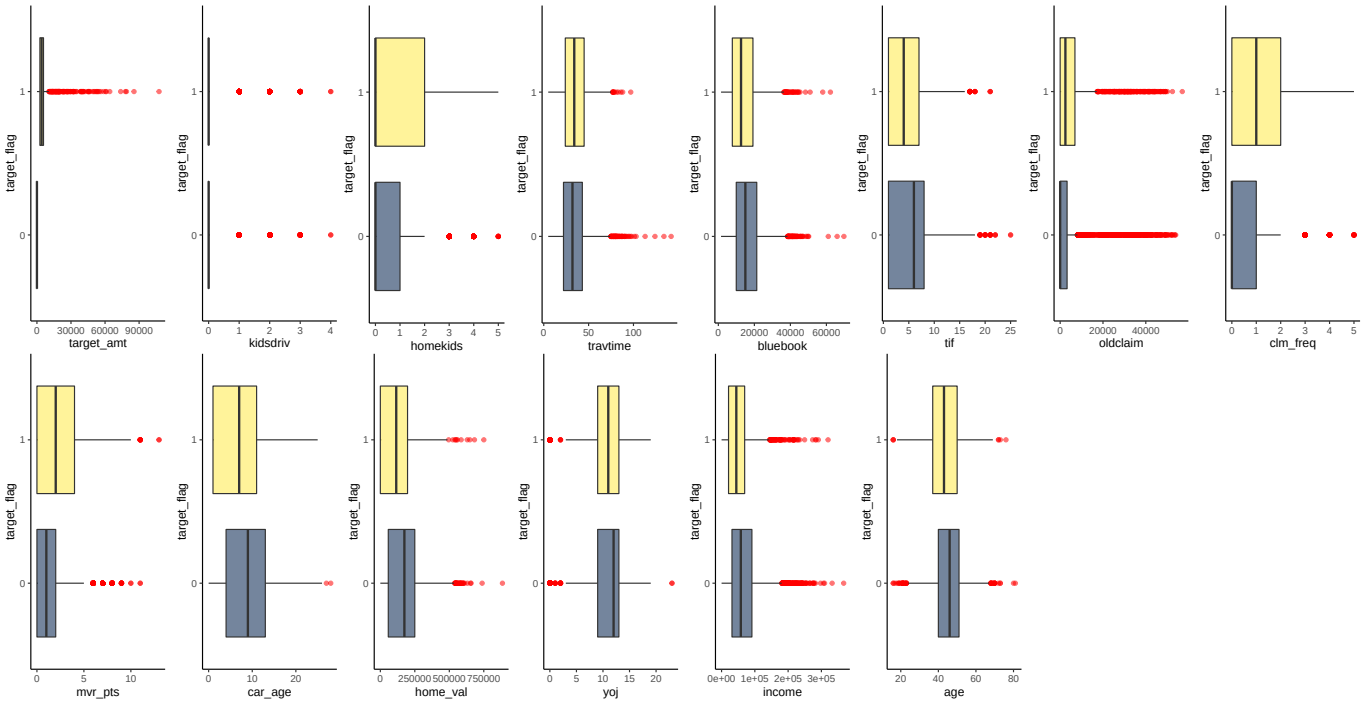
$2

$3

$4

$5

attr(,"class") [1] "list" "ggarrange"

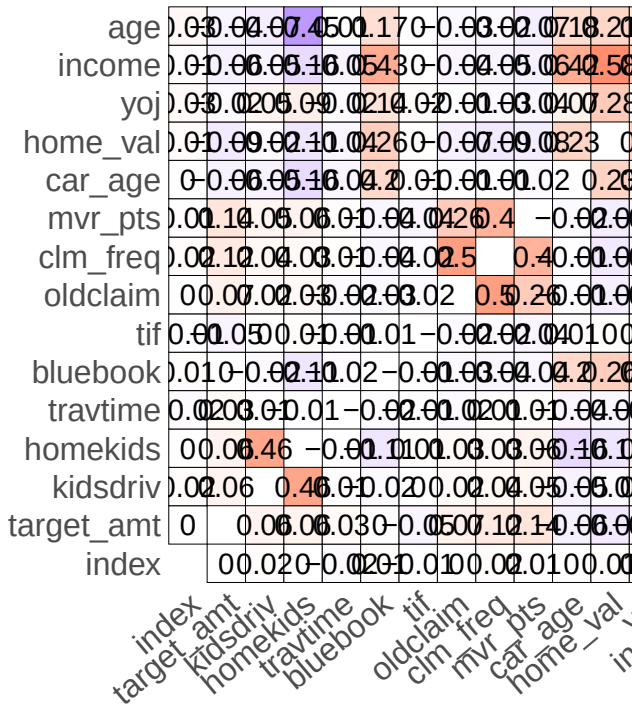


Distributions with target_flag values used for fill.

Covariance

We establish that there is only one pair of predictors that have a covariance of $>.5$. We may consider combining into an interaction term, or possible removing one from the model. We also note that correlations appear very very against the target variable; 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

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. My recommendation is to not work about class imbalance.

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.

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

target_flag	n	frequency
0	1,502	0.7362745
1	538	0.2637255

index	target_flag	kidsdriv	homekids	parent1	mstatus	sex	education	travtime	car_
1	0	N	0	N	N	M	PhD	14	Priv
2	0	N	0	N	N	M	High School	22	Cor
4	0	N	1	N	Y	F	High School	5	Priv
5	0	N	0	N	Y	M	High School	32	Priv
6	0	N	0	N	Y	F	PhD	36	Priv
7	1	N	1	Y	N	F	Bachelors	46	Cor
8	0	N	0	N	Y	F	High School	33	Priv
11	1	Y	2	N	Y	M	Bachelors	44	Cor
12	1	N	0	N	N	F	Bachelors	34	Priv
15	0	N	0	N	Y	F	Masters	36	Priv

Model 1: Base logistic model

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

Call: glm(formula = target_flag ~ kidsdriv + homekids + parent1 + mstatus + education + travtime + car_use + bluebook + tif + car_type + oldclaim + clm_freq + revoked + mvr_pts + urbanicity + home_val + yoj + income + job, family = "binomial", data = df_train)

Deviance Residuals: Min 1Q Median 3Q Max
-2.3967 -0.7180 -0.4083 0.6176 2.9852

Coefficients: Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.407e+00 2.162e-01 -11.135 < 2e-16 ***kidsdrivY 6.474e-01 1.102e-01 5.876 4.19e-09***
homekids 7.357e-02 3.828e-02 1.922 0.054586 .
parent1Y 2.962e-01 1.253e-01 2.363 0.018106 *
mstatusY -4.942e-01 9.933e-02 -4.975 6.53e-07 ***educationBachelors -5.444e-01 8.846e-02 -6.154 7.54e-10*** educationMasters -3.908e-01 1.107e-01 -3.529 0.000418 ***educationPhD -5.037e-01 1.645e-01 -3.062 0.002198*** *travtime 1.561e-02 2.147e-03 7.271 3.56e-13* **car_usePrivate -7.500e-01 9.381e-02 -7.995 1.29e-15** *bluebook -2.838e-05 5.500e-06 -5.160 2.48e-07* **tif -5.473e-02 8.446e-03 -6.479 9.21e-11** *car_typePanel Truck 6.940e-01 1.674e-01 4.146 3.38e-05* **car_typePickup 5.883e-01 1.149e-01 5.121 3.04e-07** *car_typeSports Car 1.029e+00 1.226e-01 8.392 < 2e-16* **car_typeSUV 6.847e-01 9.848e-02 6.953 3.57e-12** *car_typeVan 7.155e-01 1.380e-01 5.186 2.15e-07* **oldclaim -1.357e-05 4.531e-06 -2.994 0.002753** clm_freq 1.986e-01 3.310e-02 5.999 1.98e-09 ***revokedY 9.038e-01 1.045e-01 8.649 < 2e-16*** mvr_pts 1.189e-01 1.563e-02 7.607 2.81e-14 ***urbanicityUrban 2.318e+00 1.297e-01 17.868 < 2e-16*** home_val -1.450e-06 4.027e-07 -3.600 0.000318 * **yoj -1.432e-02 8.909e-03 -1.607 0.108074**

income -3.849e-06 1.232e-06 -3.125 0.001780 jobProfessional -1.380e-01 9.429e-02 -1.464 0.143240 — Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 7061.5  on 6119  degrees of freedom
```

Residual deviance: 5499.3 on 6094 degrees of freedom AIC: 5551.3

Number of Fisher Scoring iterations: 5

Model 1 Evaluation

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

Diagnostics

Confusion Matrix and Statistics

```
      Reference
```

Prediction 0 1 0 4158 945 1 348 669

```
           Accuracy : 0.7887
             95% CI : (0.7783, 0.7989)
No Information Rate : 0.7363
P-Value [Acc > NIR] : < 2.2e-16

              Kappa : 0.3827
```

Mcnemar’s Test P-Value : < 2.2e-16

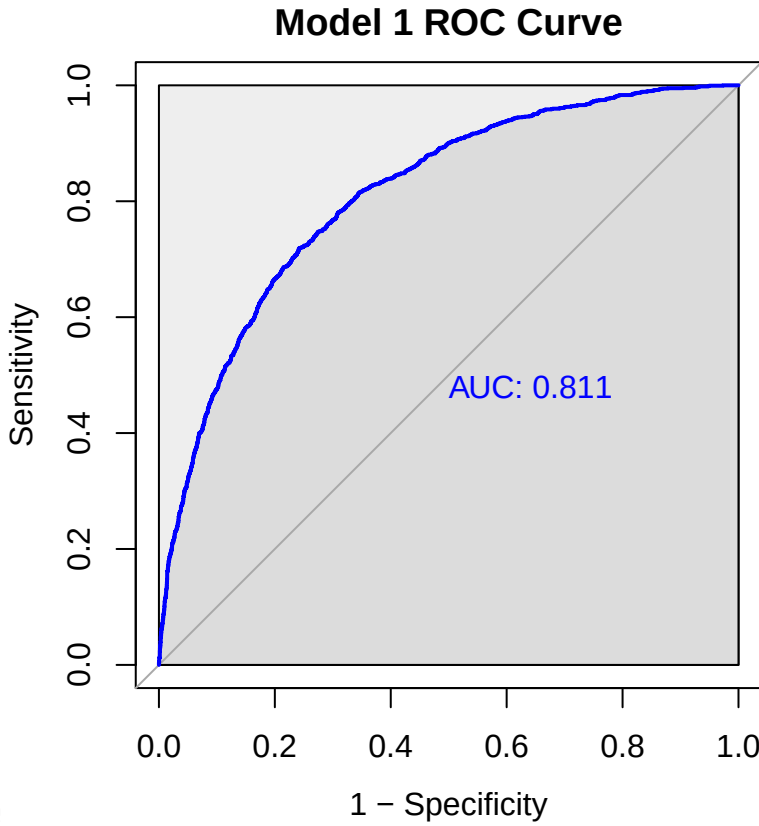
```
        Sensitivity : 0.9228
        Specificity : 0.4145
     Pos Pred Value : 0.8148
     Neg Pred Value : 0.6578
         Prevalence : 0.7363
     Detection Rate : 0.6794
```

Detection Prevalence : 0.8338
Balanced Accuracy : 0.6686

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

A tibble: 1 x 12

model predictors sensitivity specificity pos_rate neg_rate precision recall 1 Base~ 25 0.923 0.414 0.815 0.658



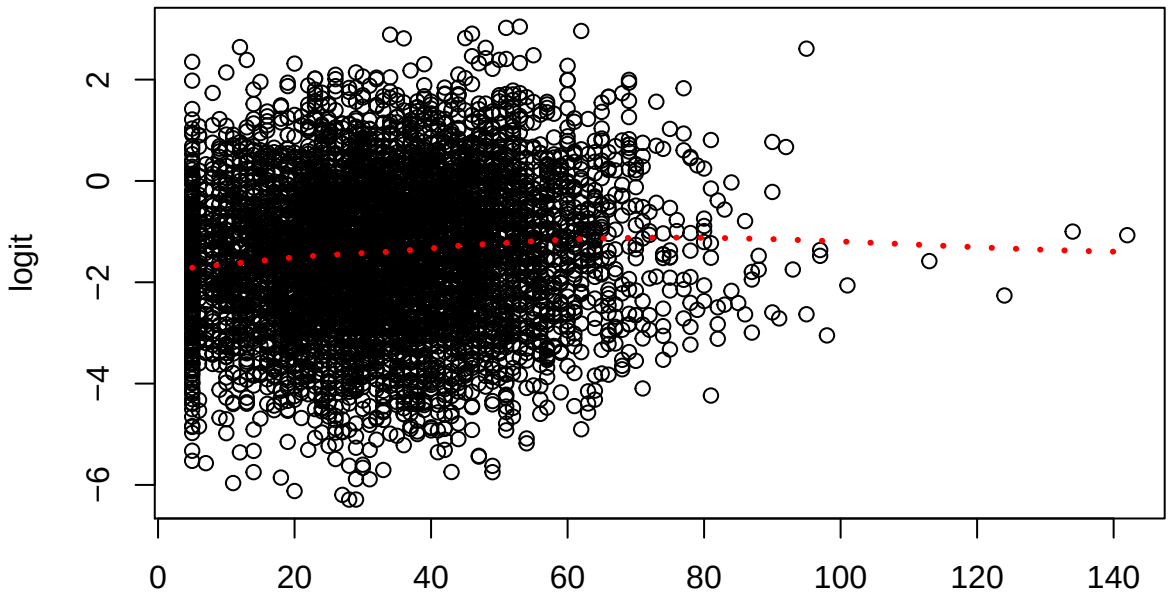
0.815 0.923 # . . . with 4 more variables: f1 , auc , AIC , BIC

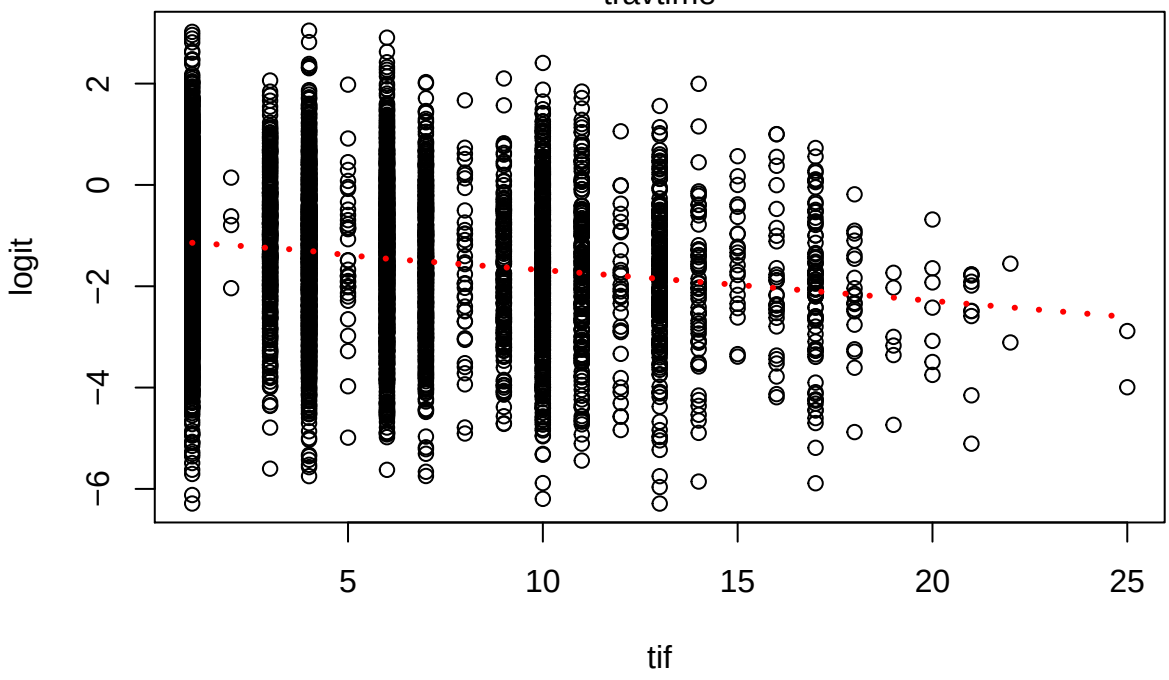
Dispersion We assess dispersion with two calculations; their results are shown below.

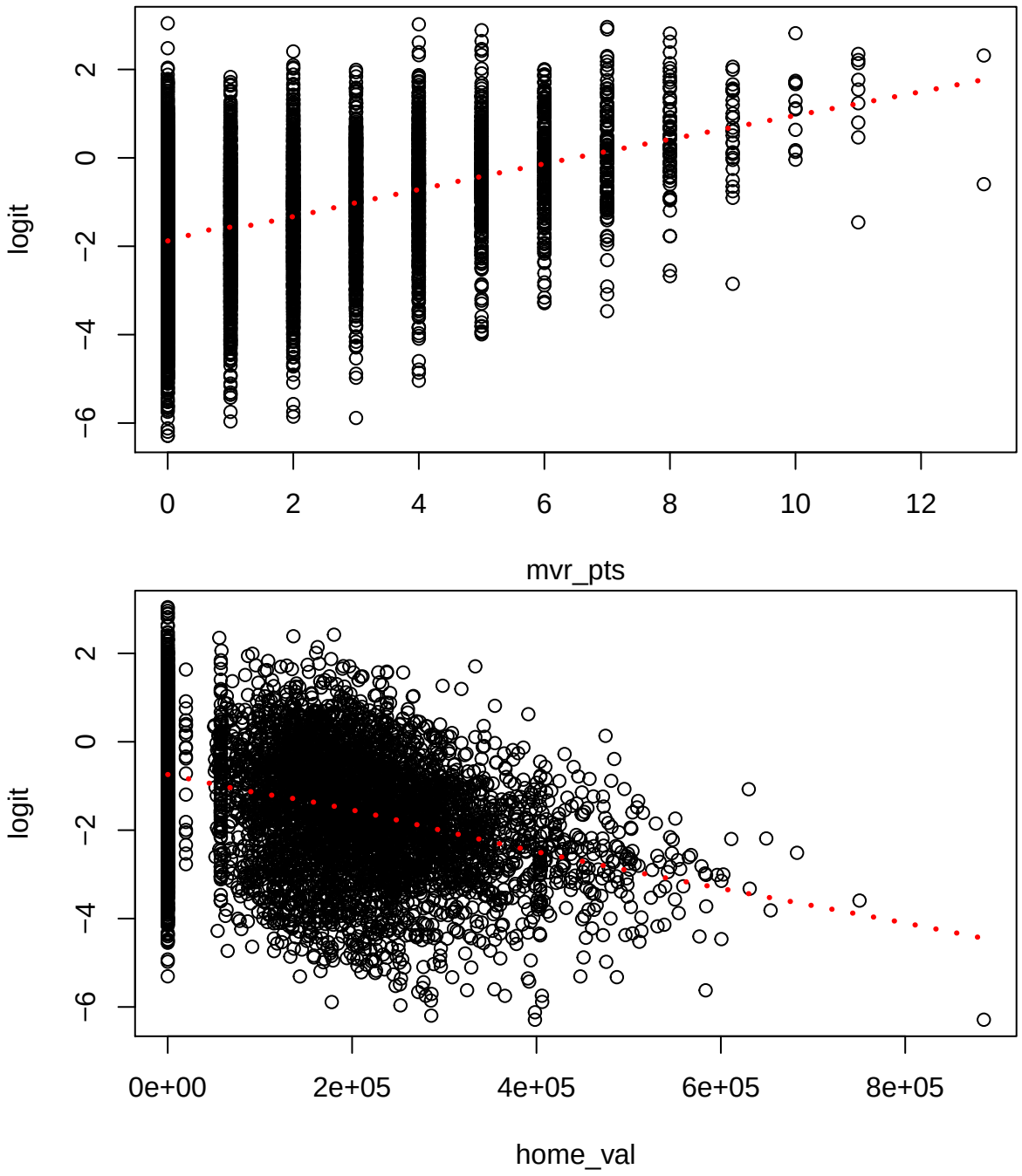
[1] "We divide the deviance by the residuals to obtaine the values 0.9024. There is no overt concern since the values is not greater than 1" [1] "Next we obtain a Pearson Chi-Squared test statistic of 0.3133 This communcates that the null hypothesis is not rejected and their are no problems with dispersion."

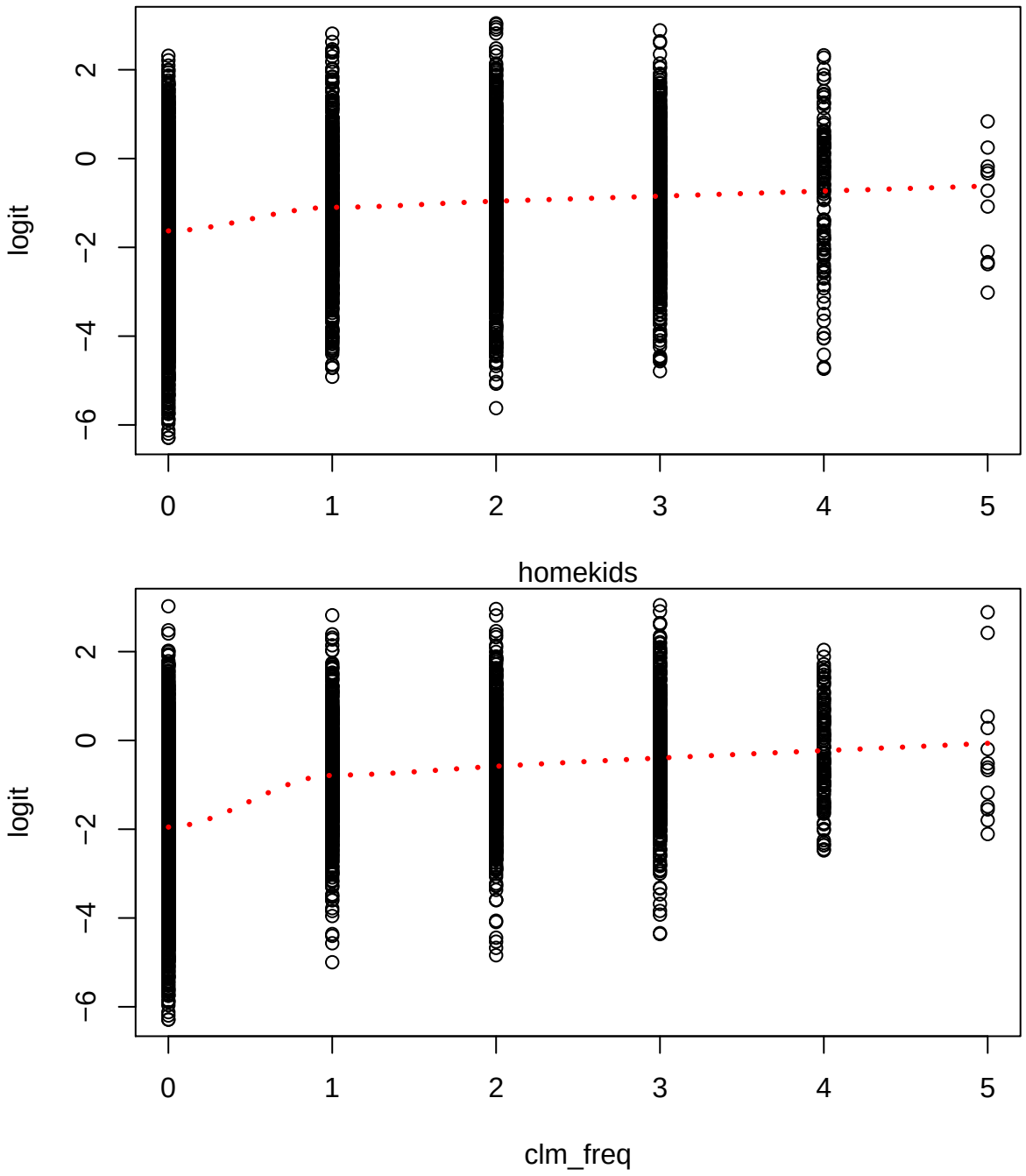
Assumption of Linearity

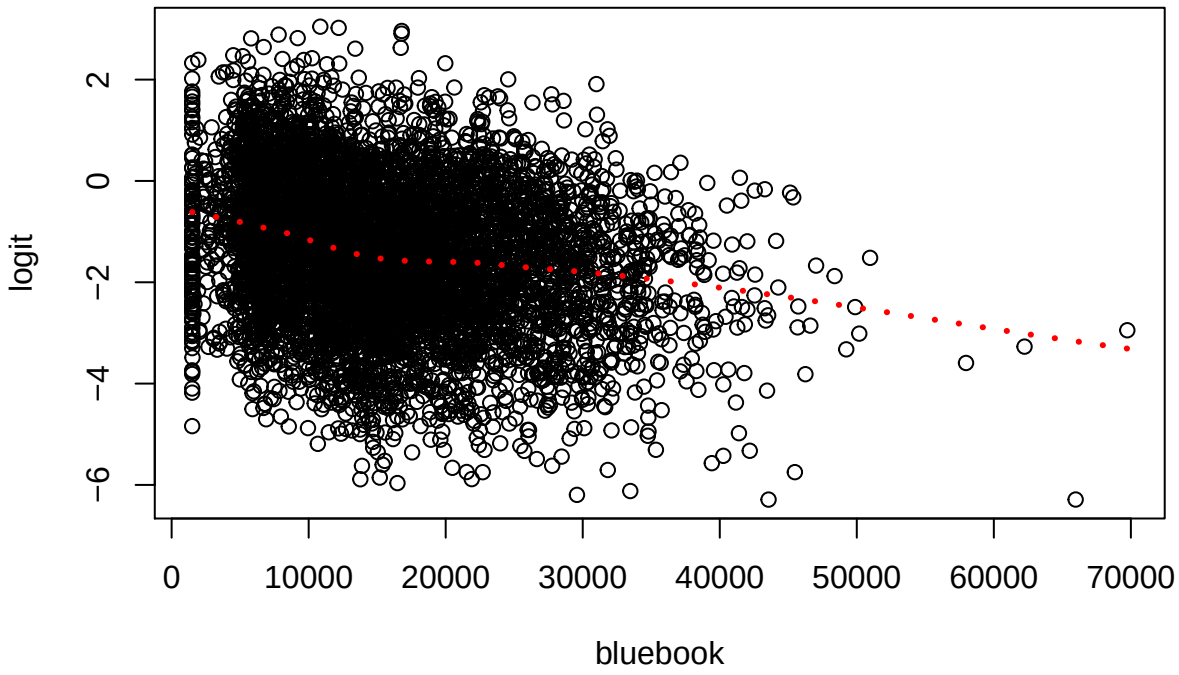
We can note that linearity is questionable for home-kids, but not convincing enough to remove at this time. yoj, oldclaim, or income are not considered in this diagnositic since they were not significant.


logit
2
0
−2
−4
−6
0
20
40
60
80
100
120
140
travtime


logit
2
0
−2
−4
−6
5
10
15
20
25
tif


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



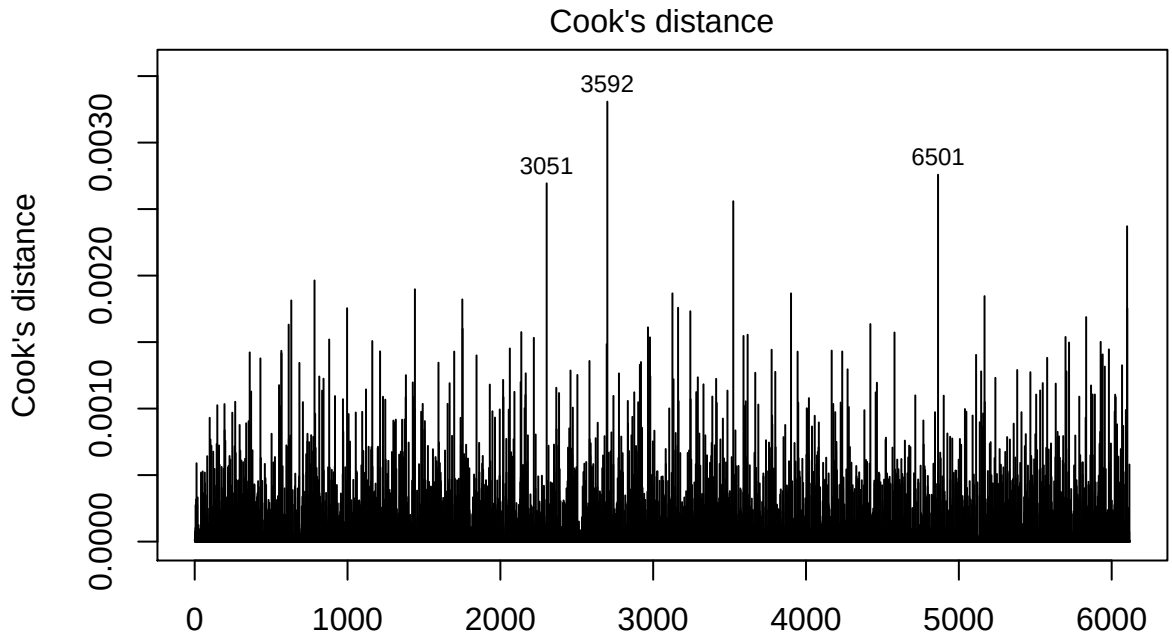
Outliers & Influenctial Points

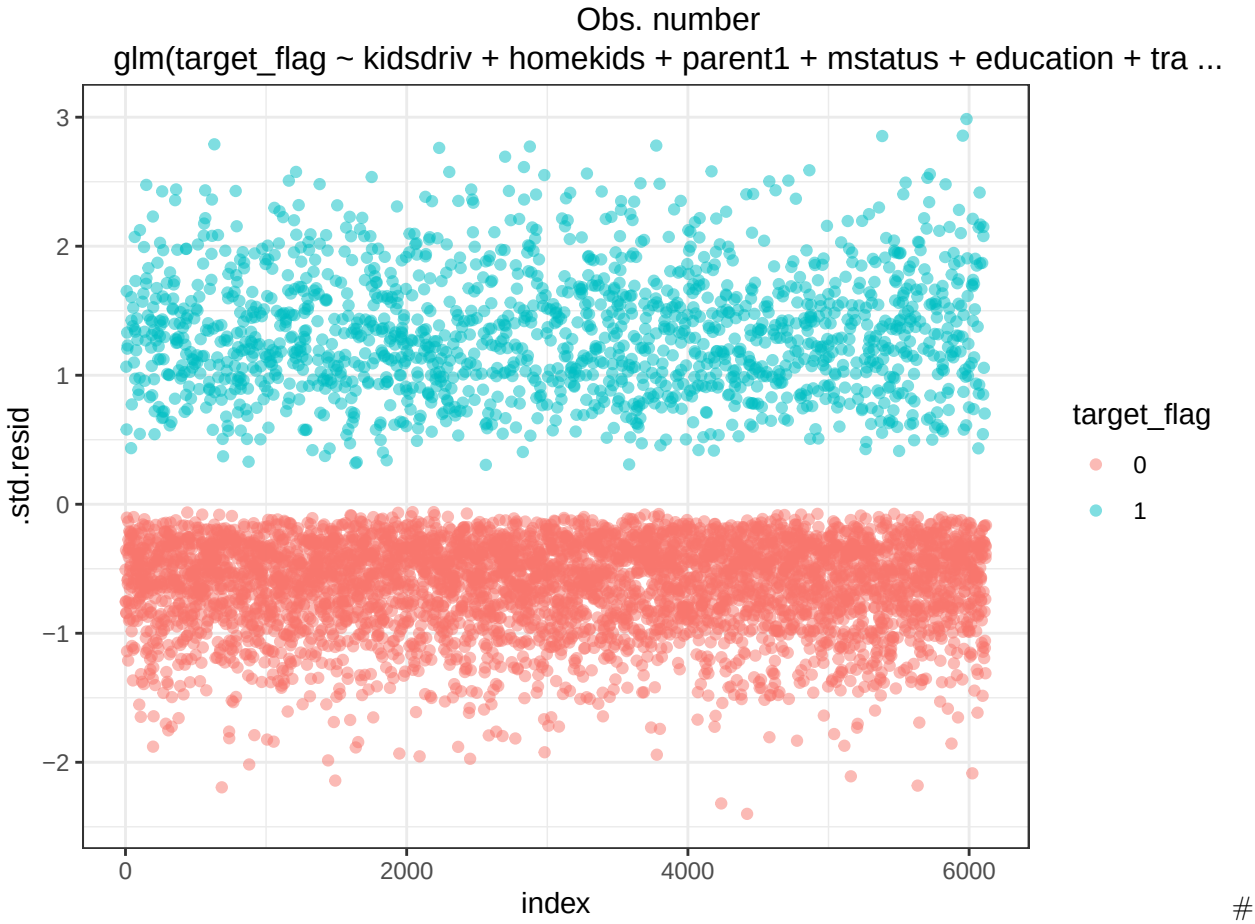
Examining the standardized residuals (.std.resid) and the Cook's distance (.cooksd) using the R function augment() [broom package] we can note the below findings. -Cooks distance indicates several standout obs (3722, 3592, 6501) but no influential points (id. D >1.0) -There are no obs with std residual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

A tibble: 10 x 28

.rownames target_flag kidsdriv homekids parent1 mstatus education travtime 1 1030 1 N 0 N Y High Sch~ 40 2 1917 0 N 4 Y N High Sch~ 23 3 3051 1 N 0 N N PhD 41 4 3592 1 N 0 N Y PhD 19 5 4170 1 N 2 N Y High Sch~ 19 6 4690 1 Y 2 N Y Masters 41 7 5202 1 N 4 N Y High Sch~ 17 8 6501 1 N 0 N N High Sch~ 27 9 6911 1 Y 2 N Y PhD 53 10 8134 1 N 0 N N PhD 32 # ... with 20 more variables: car_use , bluebook , tif , # car_type , oldclaim , clm_freq , revoked , # mvr_pts , urbanicity , home_val , yoj , income , # job , .fitted , .resid , .std.resid , .hat , # .sigma , .cooksd , index

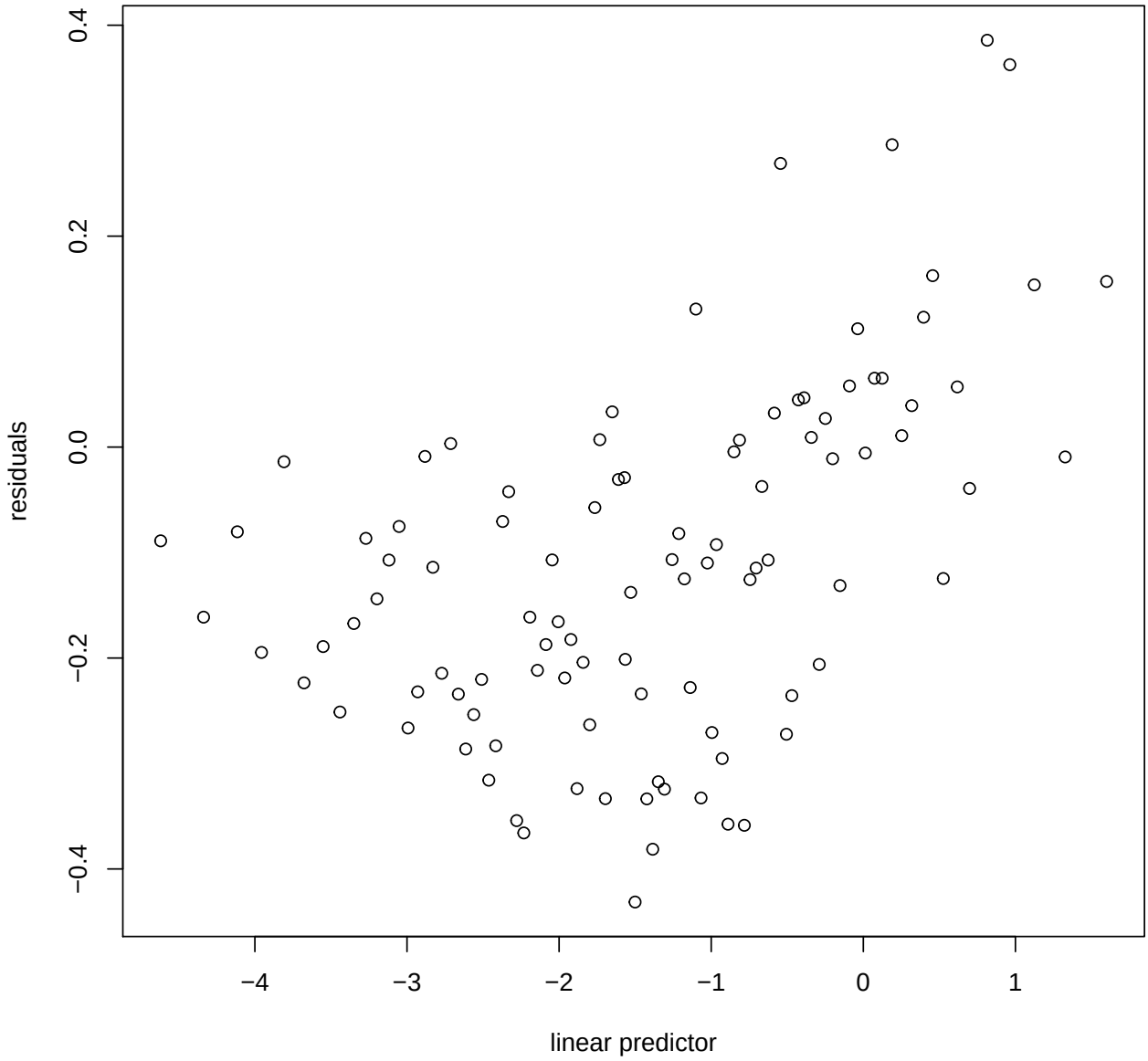




A tibble: 0 x 28 # ... with 28 variables: .rownames , target_flag , kidsdriv , # homekids , parent1 , mstatus , education , # travtime , car_use , bluebook , tif , car_type , # oldclaim , clm_freq , revoked , mvr_pts , # urbanicity , home_val , yoj , income , job , # .fitted , .resid , .std.resid , .hat , .sigma , # .cooksd , index

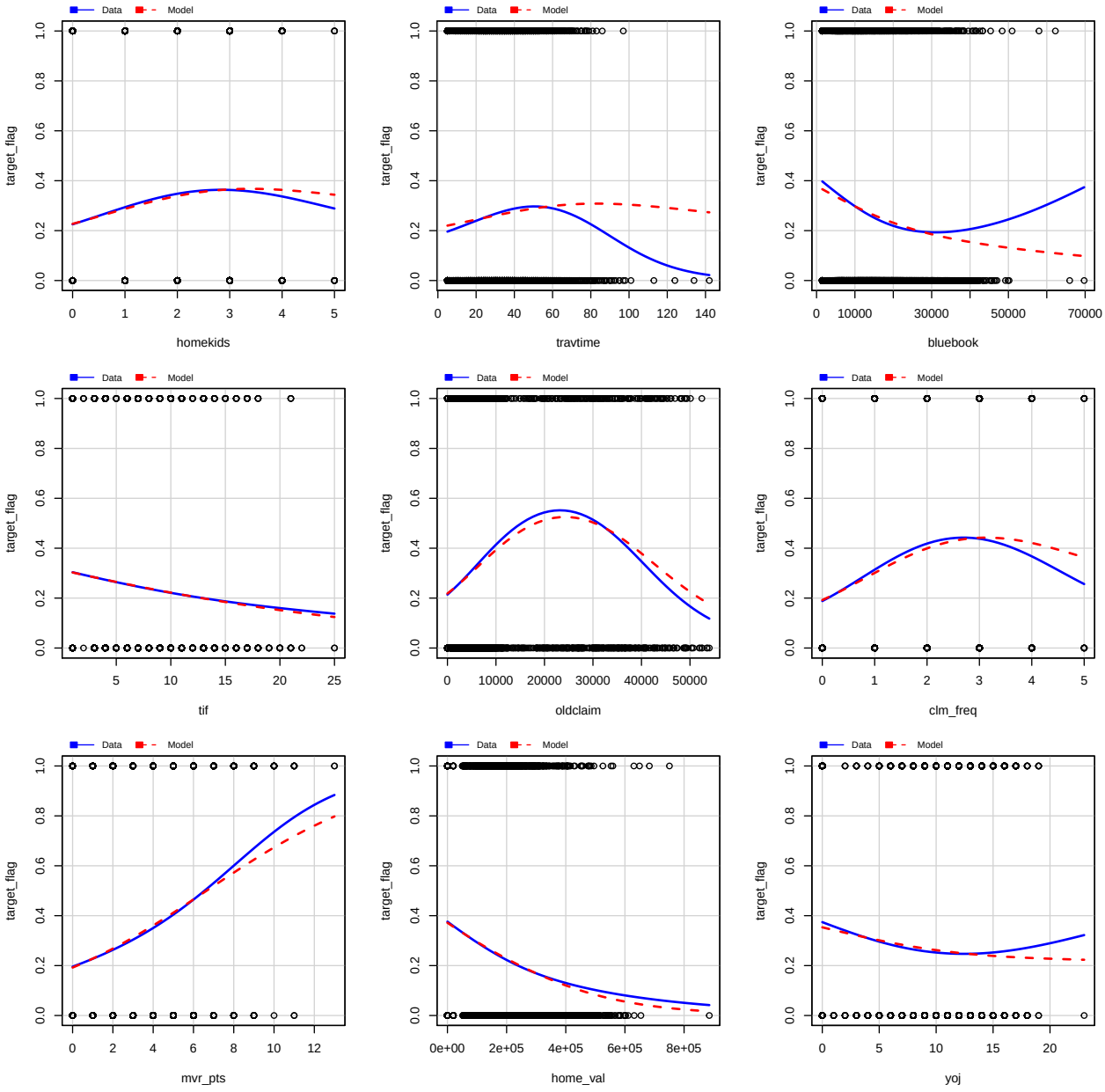
Check for Independence

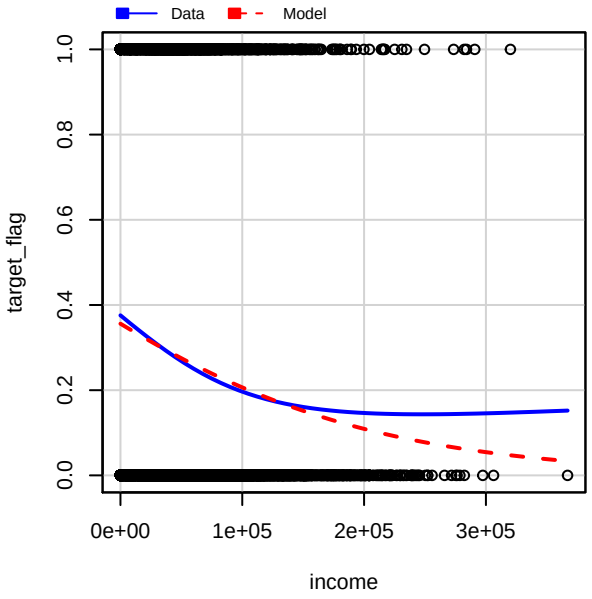
Each point on the below plot represents an aggregation of the prediction & residual values for each percentile bin of the predictions. We observe that the higher percentiles have a higher average residual. We see this as definite pattern which suggests that the model may be misclassified.

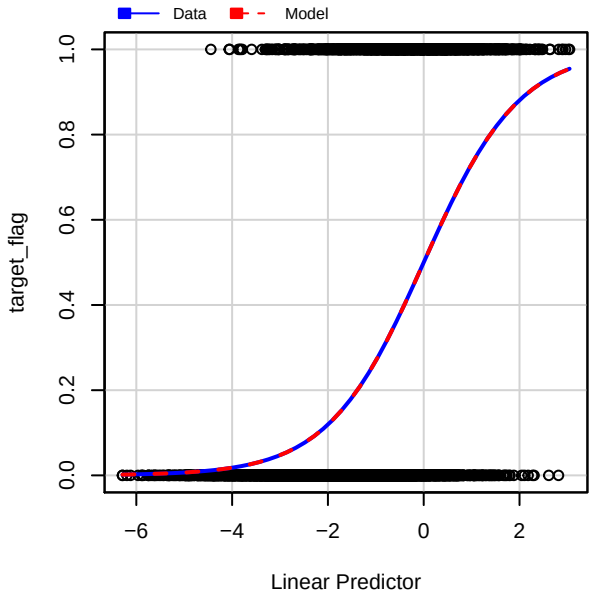


Goodness of Fit - marginal plots

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





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

Call: glm(formula = 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, family = "binomial", data = trans_df)

Deviance Residuals: Min 1Q Median 3Q Max
-2.4800 -0.7193 -0.4074 0.6057 2.9317

Coefficients: Estimate Std. Error z value Pr(>|z|)
(Intercept) 4.427e-01 6.157e-01 0.719 0.472111
kidsdrivY 7.315e-01 1.032e-01 7.089 1.35e-12 ***parent1Y 3.989e-01 1.091e-01 3.655 0.000257*** mstatusY -4.864e-01 9.440e-02 -5.153 2.57e-07 ***educationBachelors -5.248e-01 8.856e-02 -5.925 3.11e-09*** educationMasters -3.694e-01 1.097e-01 -3.368 0.000756 ***educationPhD -5.315e-01 1.550e-01 -3.429 0.000606*** car_usePrivate -7.462e-01 9.363e-02 -7.969 1.60e-15 ***tif -5.377e-02 8.463e-03 -6.353 2.11e-10*** car_typePanel Truck 5.955e-01 1.596e-01 3.731 0.000191 ***car_typePickup 6.010e-01 1.146e-01 5.244 1.57e-07*** car_typeSports Car 1.016e+00 1.230e-01 8.256 < 2e-16 ***car_typeSUV 6.986e-01 9.828e-02 7.108 1.18e-12*** car_typeVan 7.386e-01 1.381e-01 5.347 8.92e-08 ***oldclaim -1.367e-05 4.543e-06 -3.009 0.002620*** *revokedY 9.140e-01 1.046e-01 8.738 < 2e-16* **urbanicityUrban 2.312e+00 1.300e-01 17.787 < 2e-16** *home_val -1.346e-06 3.983e-07 -3.380 0.000726* **jobProfessional -1.850e-01 9.578e-02 -1.932 0.053379 .**
travtime 3.706e-02 7.554e-03 4.906 9.28e-07 *I(travtimeˆ2) -2.890e-04 9.886e-05 -2.923 0.003462* ***mvr_pts 1.206e-01 1.567e-02 7.696 1.40e-14*** clm_freq 1.994e-01 3.312e-02 6.022 1.72e-09 ***log_bluebook -3.663e-01 6.325e-02 -5.791 7.01e-09*** sqrt_income -2.203e-03 4.836e-04 -4.555 5.24e-06 *** — Signif. codes: 0 ' ' ***0.001*** '' *0.01* '' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 7061.5  on 6119  degrees of freedom
```

Residual deviance: 5480.0 on 6095 degrees of freedom AIC: 5530

Number of Fisher Scoring iterations: 5

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

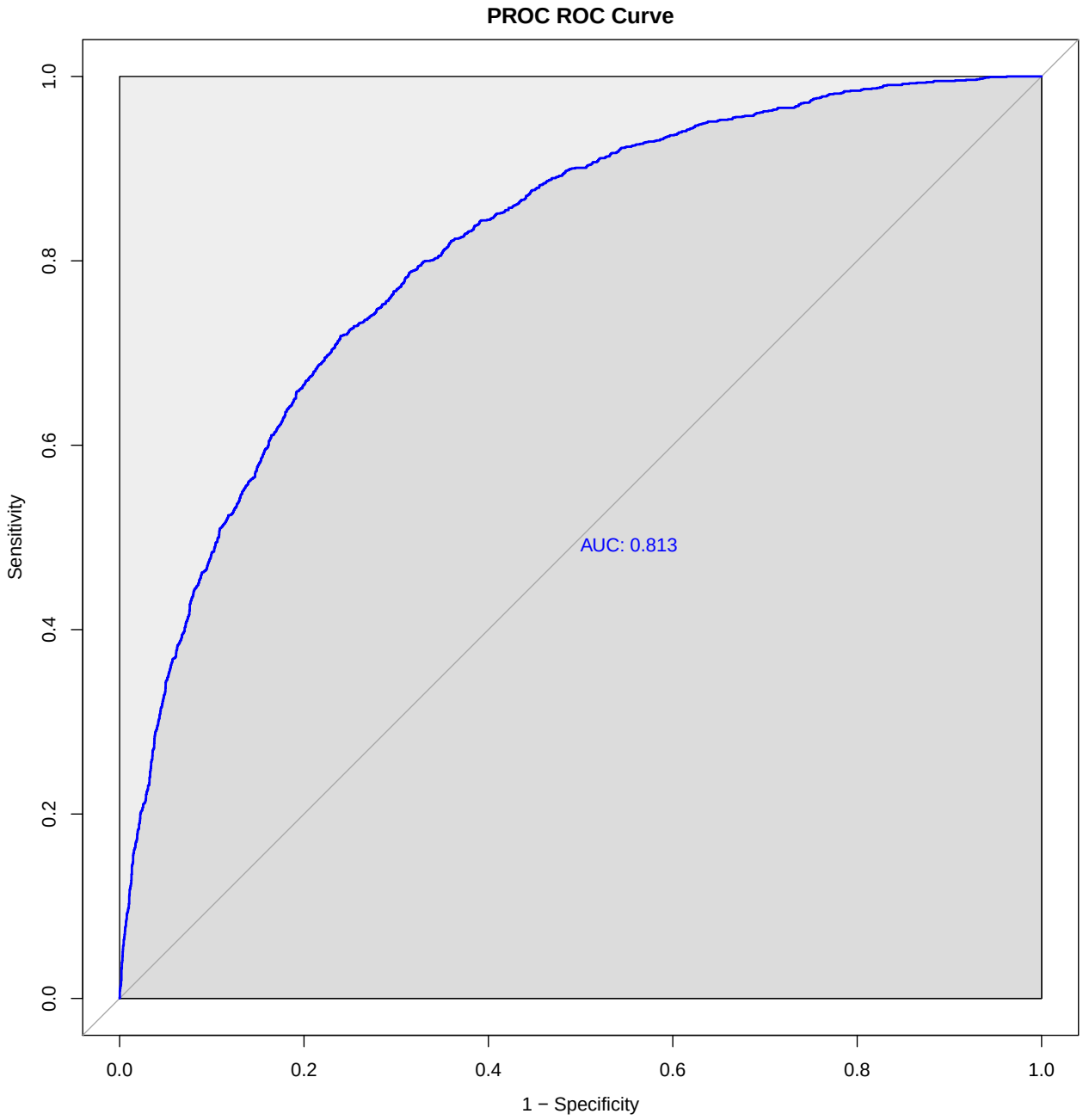
```
        Sensitivity : 0.9239
        Specificity : 0.4250
     Pos Pred Value : 0.8177
     Neg Pred Value : 0.6667
         Prevalence : 0.7363
     Detection Rate : 0.6802
```

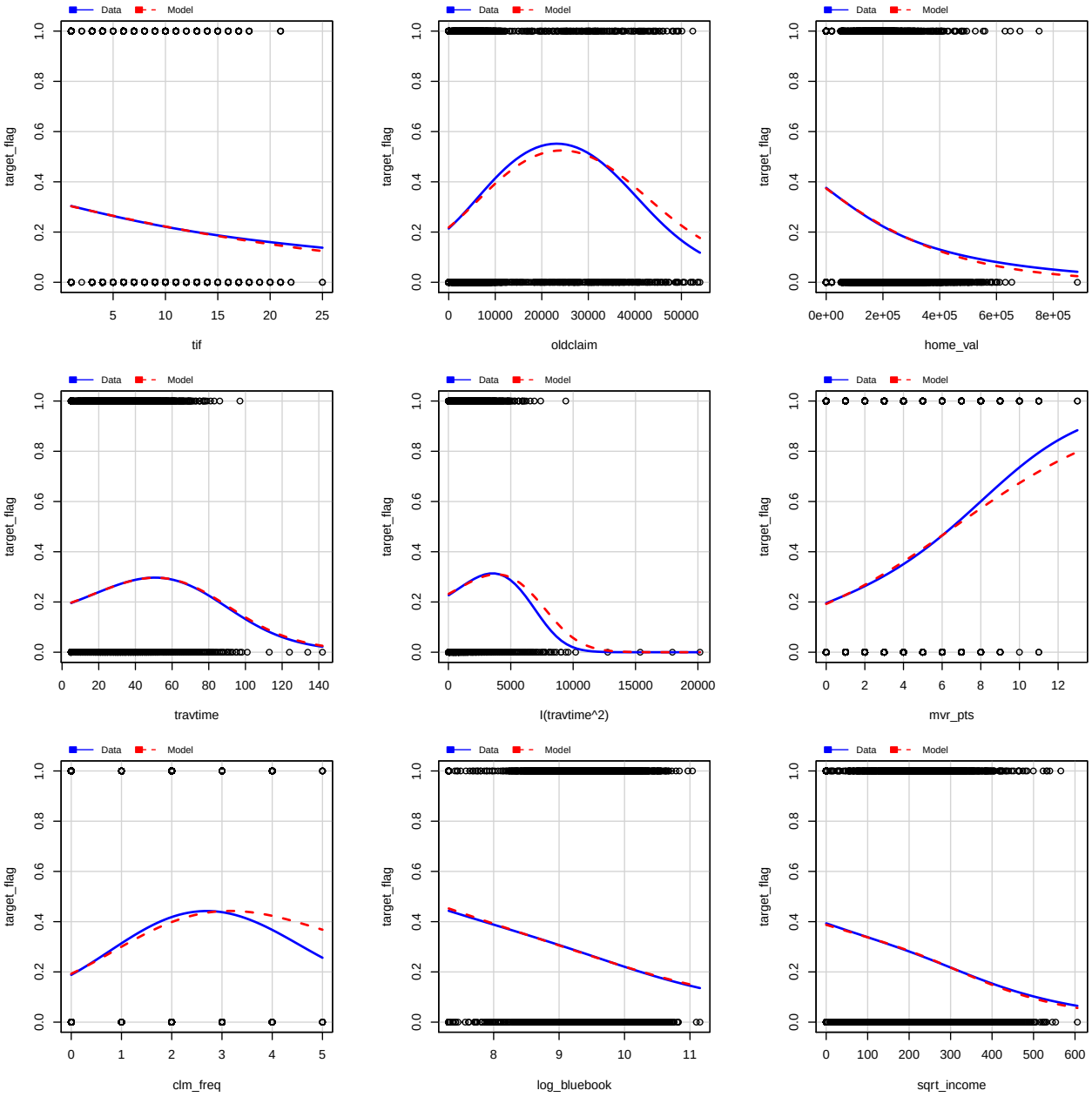
Detection Prevalence : 0.8319
Balanced Accuracy : 0.6745

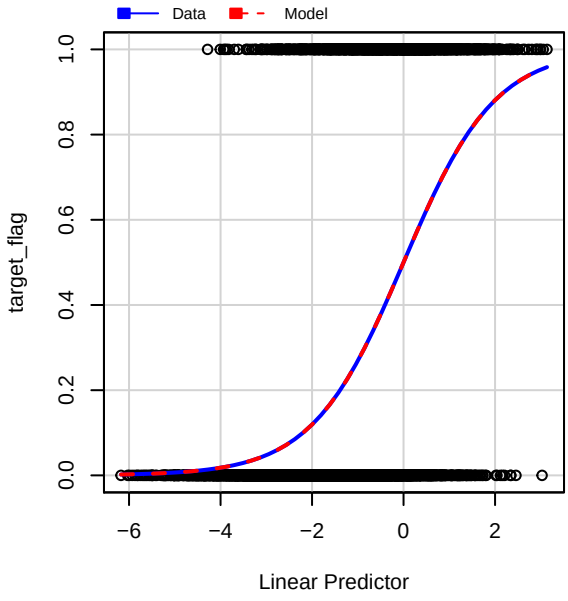
```
   'Positive' Class : 0
```

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

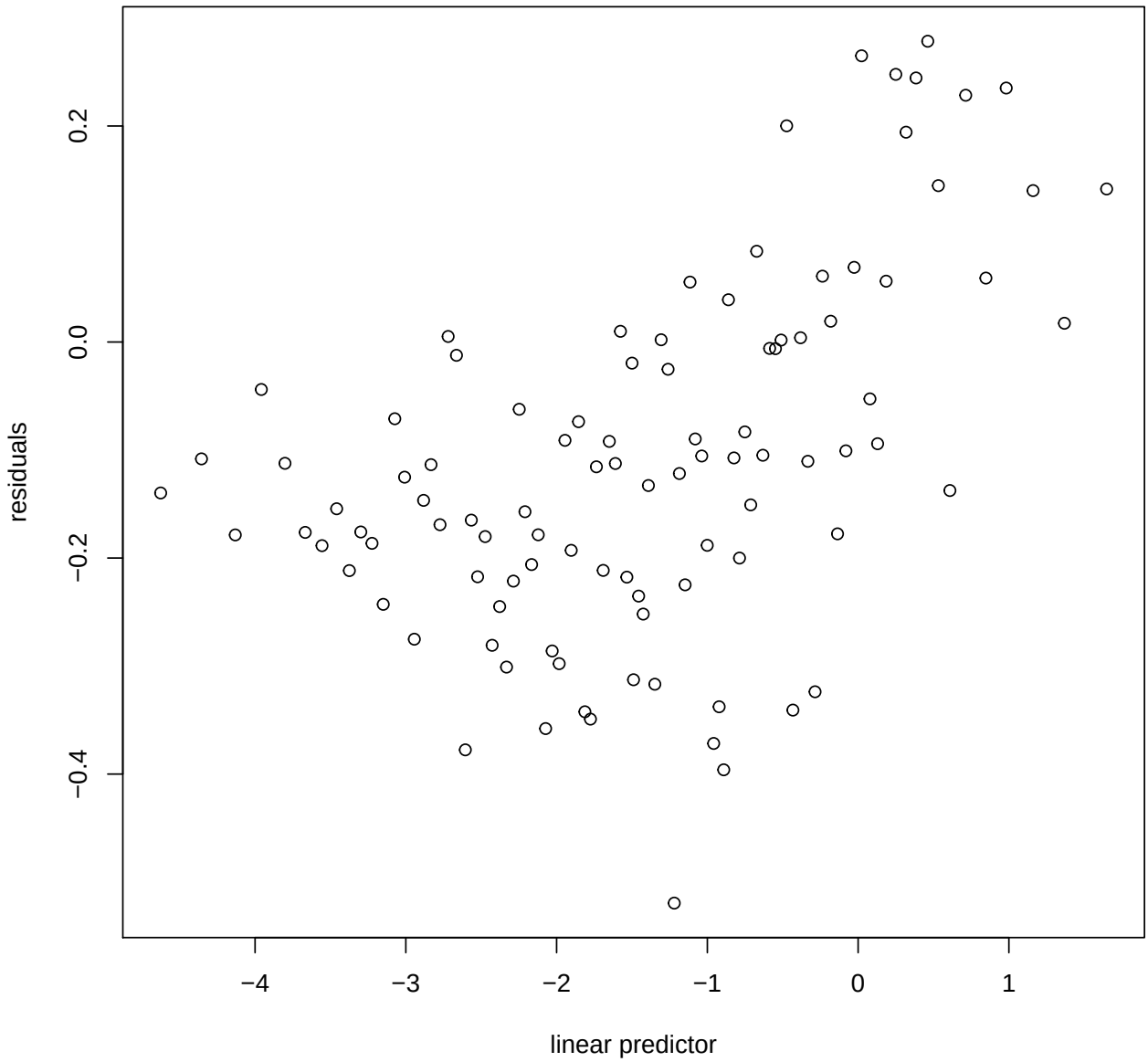

PROC ROC Curve
Sensitivity
1 − Specificity
0.0
0.2
0.4
0.6
0.8
1.0
AUC: 0.813


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



Independence

Still seeing pattern - possible misspecification

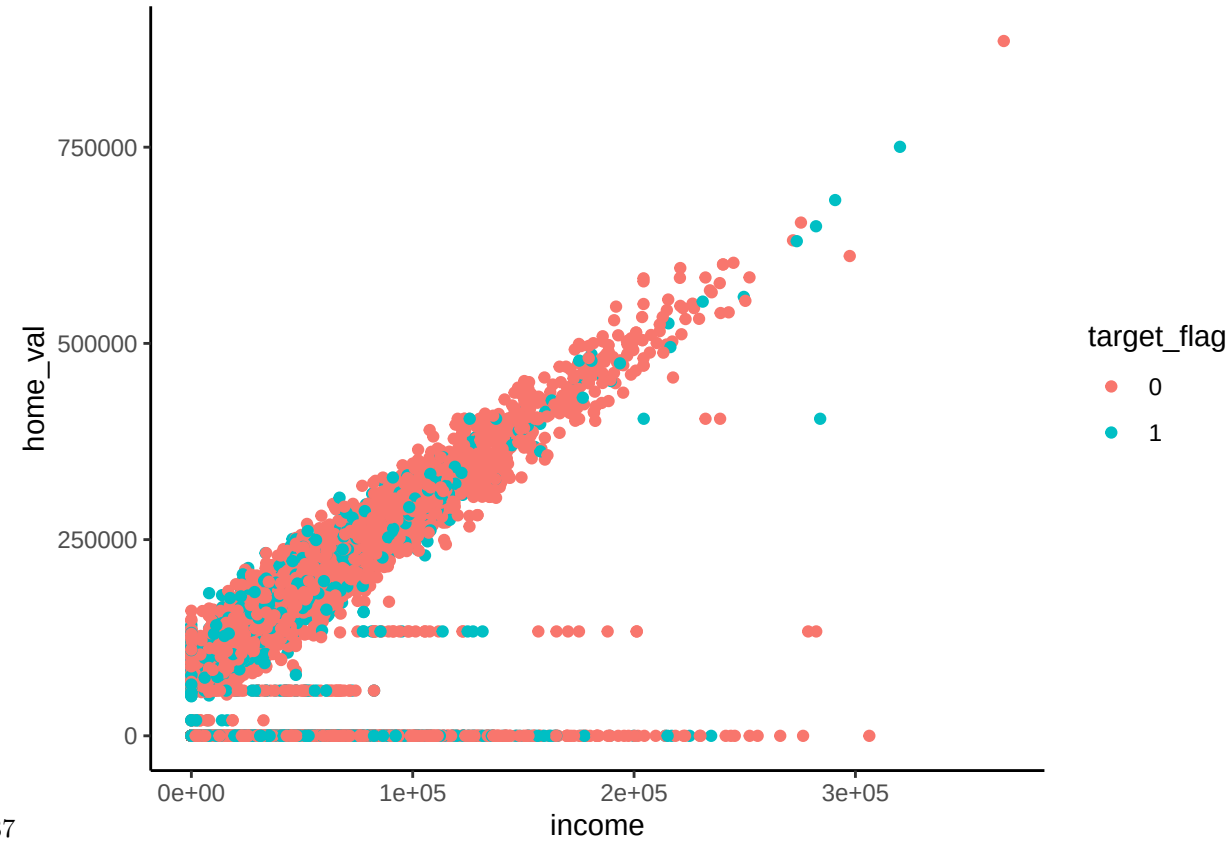


Model 3 - Feature engineering and Interactions among predictor variables

Establish new variable of liquidity = (home_val+1)/(income+1) to account for the covariance 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")

bluebook=log(bluebook) additional interaction terms car_use*car_type # A tibble: 4 x 3 var1 var2 coef_corr 1 clm_freq oldclaim 0.503 2 oldclaim clm_freq 0.503 3 income home_val 0.587 4 home_val income



0.587
Cramer V 0.3306 Cramer V 0.54 Cramer V 0.04474

Call: glm(formula = urbanicity ~ travtime, family = binomial, data = int_df)

Deviance Residuals: Min 1Q Median 3Q Max
-2.1260 0.4869 0.6032 0.7009 1.2667

Coefficients: Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.277898 0.080733 28.21 <2e-16 ***travtime -0.025623 0.001994 -12.85 <2e-16*** — Signif. codes: 0 '' ***0.001*** '' *0.01* '' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 6163.5  on 6119  degrees of freedom
```

Residual deviance: 5993.9 on 6118 degrees of freedom AIC: 5997.9

Number of Fisher Scoring iterations: 4

Call: glm(formula = revoked ~ mvr_pts, family = binomial, data = int_df)

Deviance Residuals: Min 1Q Median 3Q Max
-0.6869 -0.5108 -0.4773 -0.4773 2.1112

Coefficients: Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.11476 0.05143 -41.116 < 2e-16 ***mvr_pts 0.07188 0.01700 4.229 2.35e-05*** — Signif. codes: 0 '' ***0.001*** '' *0.01* '' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 4513.4  on 6119  degrees of freedom
```

Residual deviance: 4496.3 on 6118 degrees of freedom AIC: 4500.3

Number of Fisher Scoring iterations: 4

Call: glm(formula = kidsdriv ~ clm_freq, family = binomial, data = int_df)

Deviance Residuals: Min 1Q Median 3Q Max
-0.6022 -0.4989 -0.4756 -0.4756 2.1145

Coefficients: Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.12242 0.04977 -42.642 < 2e-16 * **clm_freq 0.10141 0.03303 3.071 0.00214** — Signif. codes: 0 ‘***’ ***0.001*** ’**’ *0.01* ’*’ 0.05 ‘.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 4376.7  on 6119  degrees of freedom
```

Residual deviance: 4367.6 on 6118 degrees of freedom AIC: 4371.6

Number of Fisher Scoring iterations: 4

Call: glm(formula = car_type ~ clm_freq, family = binomial, data = int_df)

Deviance Residuals: Min 1Q Median 3Q Max
-1.8822 -1.5841 0.7740 0.8194 0.8194

Coefficients: Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.91897 0.03447 26.657 < 2e-16 ***clm_freq 0.13317 0.02648 5.029 4.94e-07*** — Signif. codes: 0 ‘***’ ***0.001*** ’**’ *0.01* ’*’ 0.05 ‘.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 7081.9  on 6119  degrees of freedom
```

Residual deviance: 7055.6 on 6118 degrees of freedom AIC: 7059.6

Number of Fisher Scoring iterations: 4

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

Call: glm(formula = target_flag ~ kidsdriv + parent1 + mstatus + education + travtime + car_use + I(log(bluebook)) + tif + car_type + oldclaim + clm_freq + revoked + mvr_pts + urbanicity + liquidity + car_use:car_type + travtime:urbanicity, family = binomial, data = int_df)

Deviance Residuals: Min 1Q Median 3Q Max
-2.4174 -0.7181 -0.4147 0.6510 2.9042

Coefficients: (1 not defined because of singularities) Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.399e+00 6.511e-01 2.149 0.031619 *
kidsdrivY 6.948e-01 1.025e-01 6.781 1.19e-11 ***parent1Y 4.194e-01 1.085e-01 3.867 0.000110*** mstatusY -4.270e-01 9.355e-02 -4.564 5.02e-06 ***educationBachelors -7.119e-01 8.322e-02 -8.554 < 2e-16*** educationMasters -6.984e-01 9.587e-02 -7.286 3.20e-13 ***educationPhD -1.033e+00 1.373e-01 -7.526 5.24e-14*** travtime 2.904e-03 6.295e-03 0.461 0.644520
car_usePrivate -5.645e-01 1.715e-01 -3.292 0.000994 ***I(log(bluebook)) -4.528e-01 6.093e-02 -7.431 1.08e-13*** tifmoderate -3.640e-01 7.248e-02 -5.022 5.12e-07 ***tifhigh -4.273e-01 1.167e-01 -3.662 0.000251*** car_typePanel Truck 6.714e-01 1.898e-01 3.537 0.000405 ***car_typePickup 8.005e-01 1.748e-01 4.578 4.68e-06*** car_typeSports Car 7.210e-01 2.717e-01 2.654 0.007950 ** car_typeSUV 8.957e-01 1.932e-01 4.637 3.53e-06 ***car_typeVan 9.544e-01 1.948e-01 4.899 9.63e-07*** oldclaim -2.085e-05 4.871e-06 -4.280 1.87e-05 ***clm_freqmoderate 7.121e-01 9.029e-02 7.887 3.11e-15*** clm_freqhigh 9.832e-01 1.987e-01 4.949 7.46e-07 ***revokedY 9.726e-01 1.056e-01 9.208 < 2e-16***

mvr_ptslow 2.537e-01 7.799e-02 3.253 0.001140 ** mvr_ptshigh 4.738e-01 9.345e-02 5.070 3.98e-07 ***urbanicityUrban 1.645e+00 2.862e-01 5.747 9.10e-09*** liquidityhigh -3.608e-01 8.844e-02 -4.080 4.51e-05 ** *car_usePrivate:car_typePanel Truck NA NA NA NA*
car_usePrivate:car_typePickup -3.994e-01 2.380e-01 -1.678 0.093309 .
car_usePrivate:car_typeSports Car 3.678e-01 3.022e-01 1.217 0.223527
car_usePrivate:car_typeSUV -2.290e-01 2.228e-01 -1.027 0.304191
car_usePrivate:car_typeVan -6.192e-01 2.932e-01 -2.112 0.034671
travtime:urbanicityUrban 1.498e-02 6.697e-03 2.237 0.025294 *
— Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 7061.5  on 6119  degrees of freedom
```

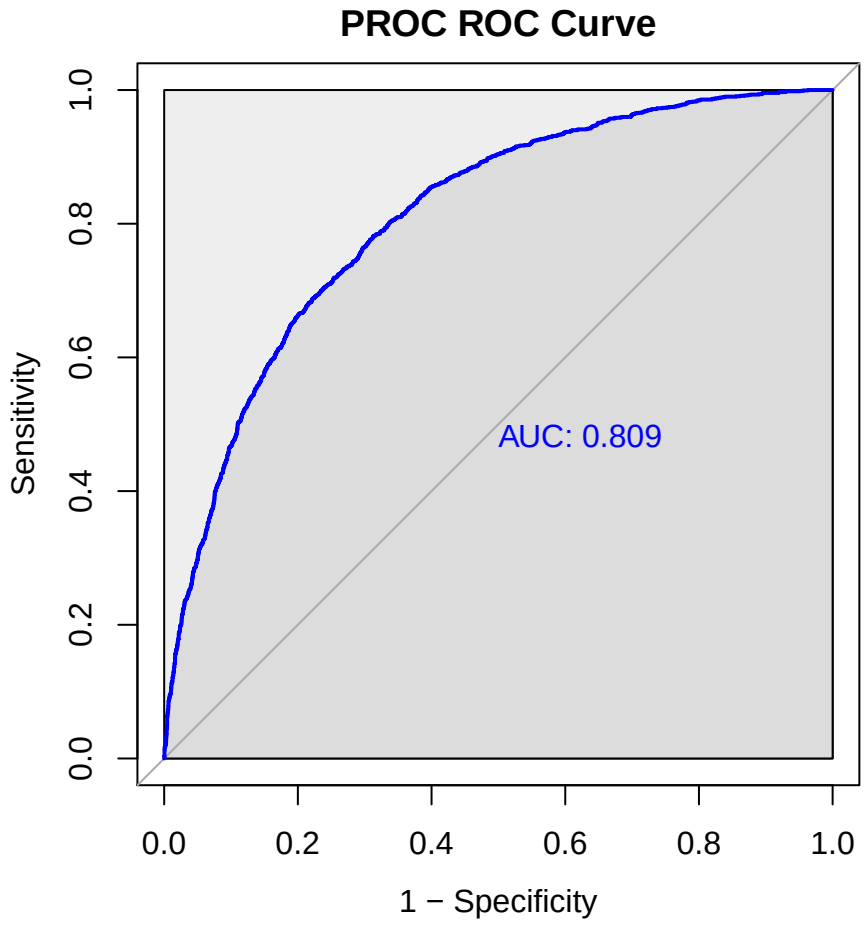
Residual deviance: 5526.3 on 6090 degrees of freedom AIC: 5586.3

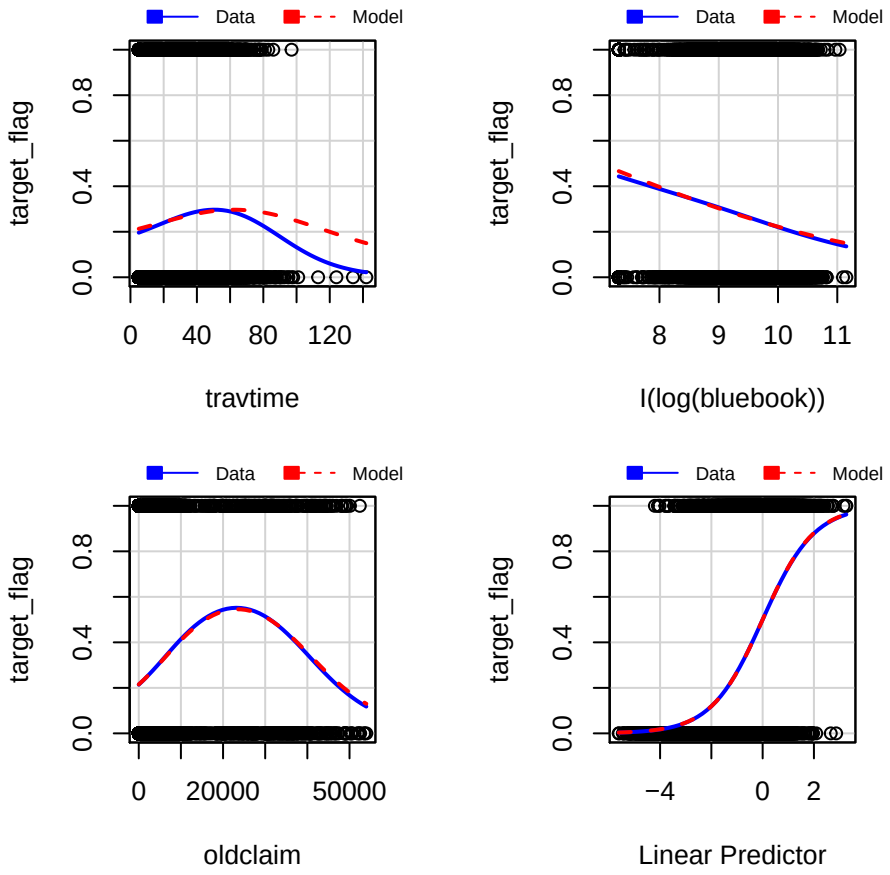
Number of Fisher Scoring iterations: 5

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





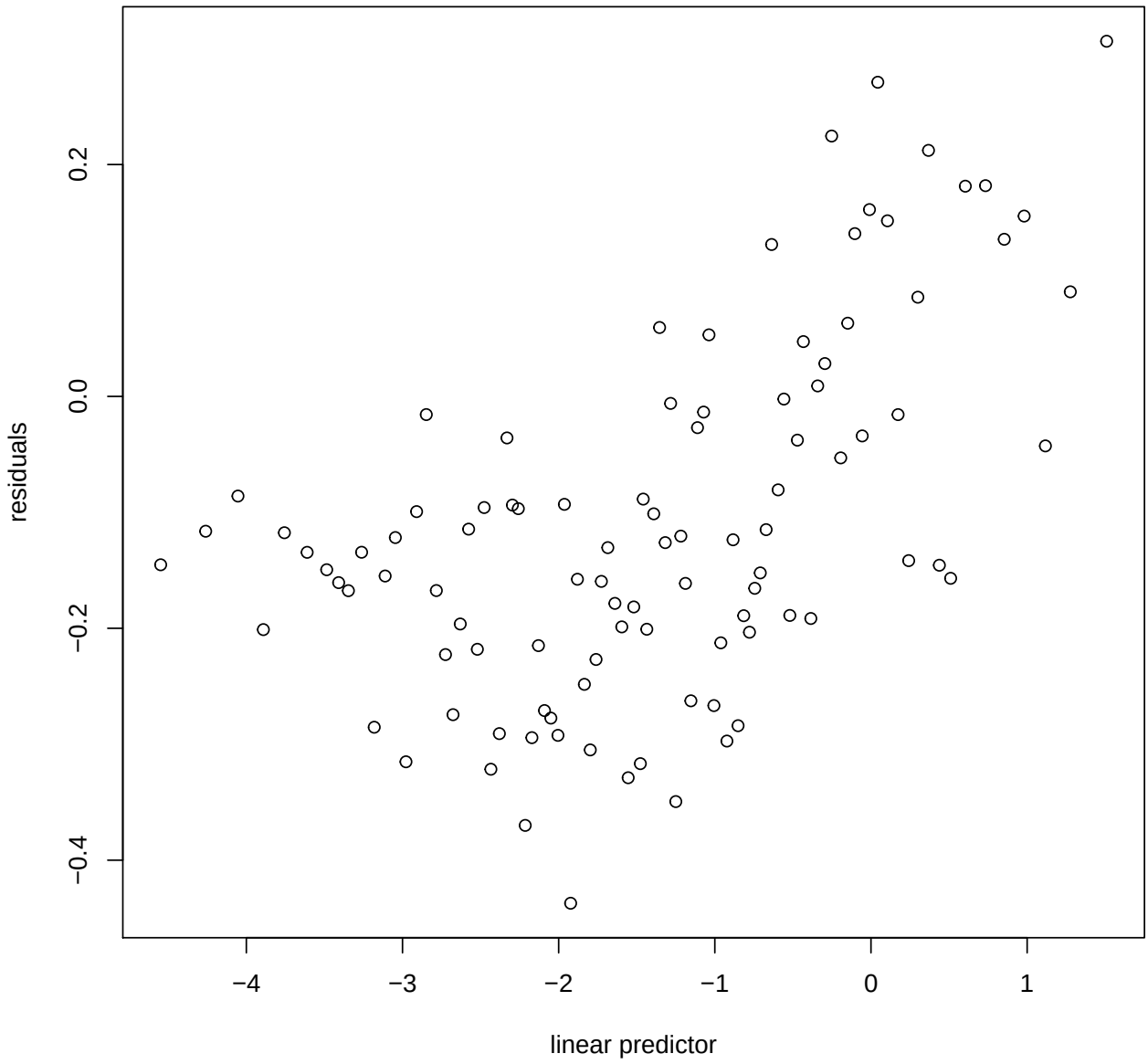
Dispersion

No evidence of significant dispersion

[1] 0.9074426 [1] 0.872883

Independence

We note that the patterns in residuals still suggest misspecification. 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.

model	predictors	sensitivity	specificity	pos_rate	neg_rate	precision	recall	f1	
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.

!!!! Cross-validate model2 - consider 0.3 threshold

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 vairables with significant p-values; and the r-squared is very low. Also the redisidual plots fail the required assumptions regarding the normal distrbution 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

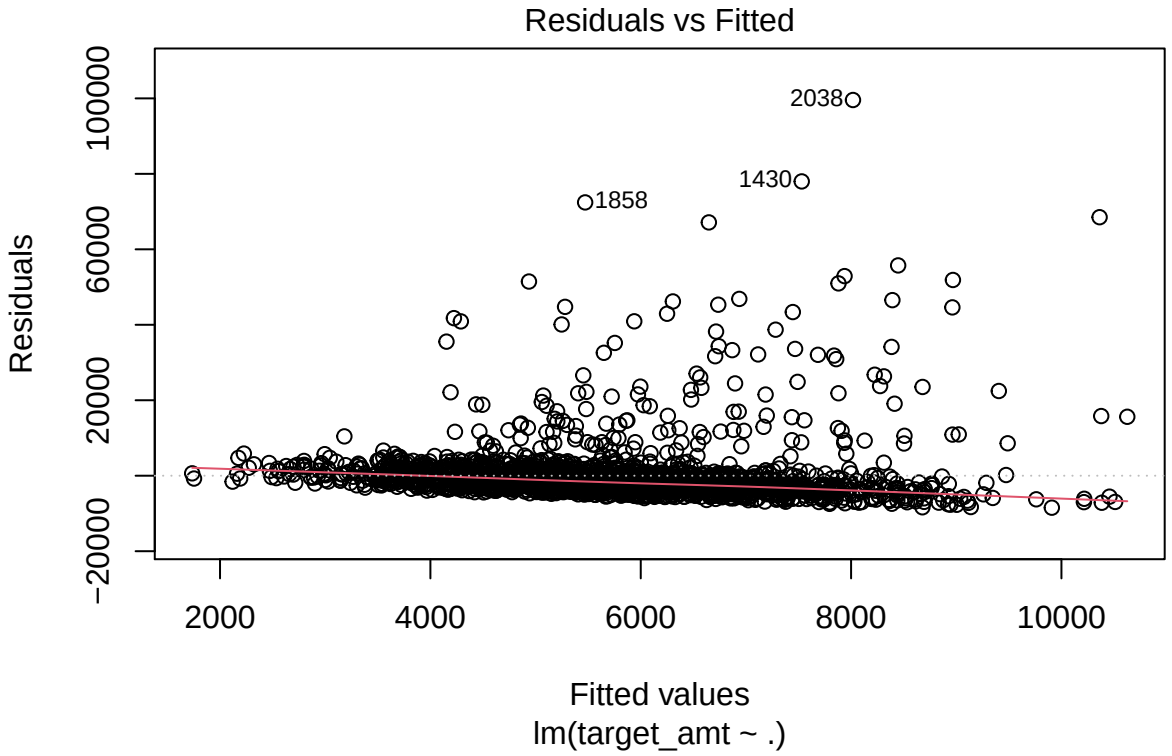
Call: lm(formula = target_amt ~ ., data = dfcrash)

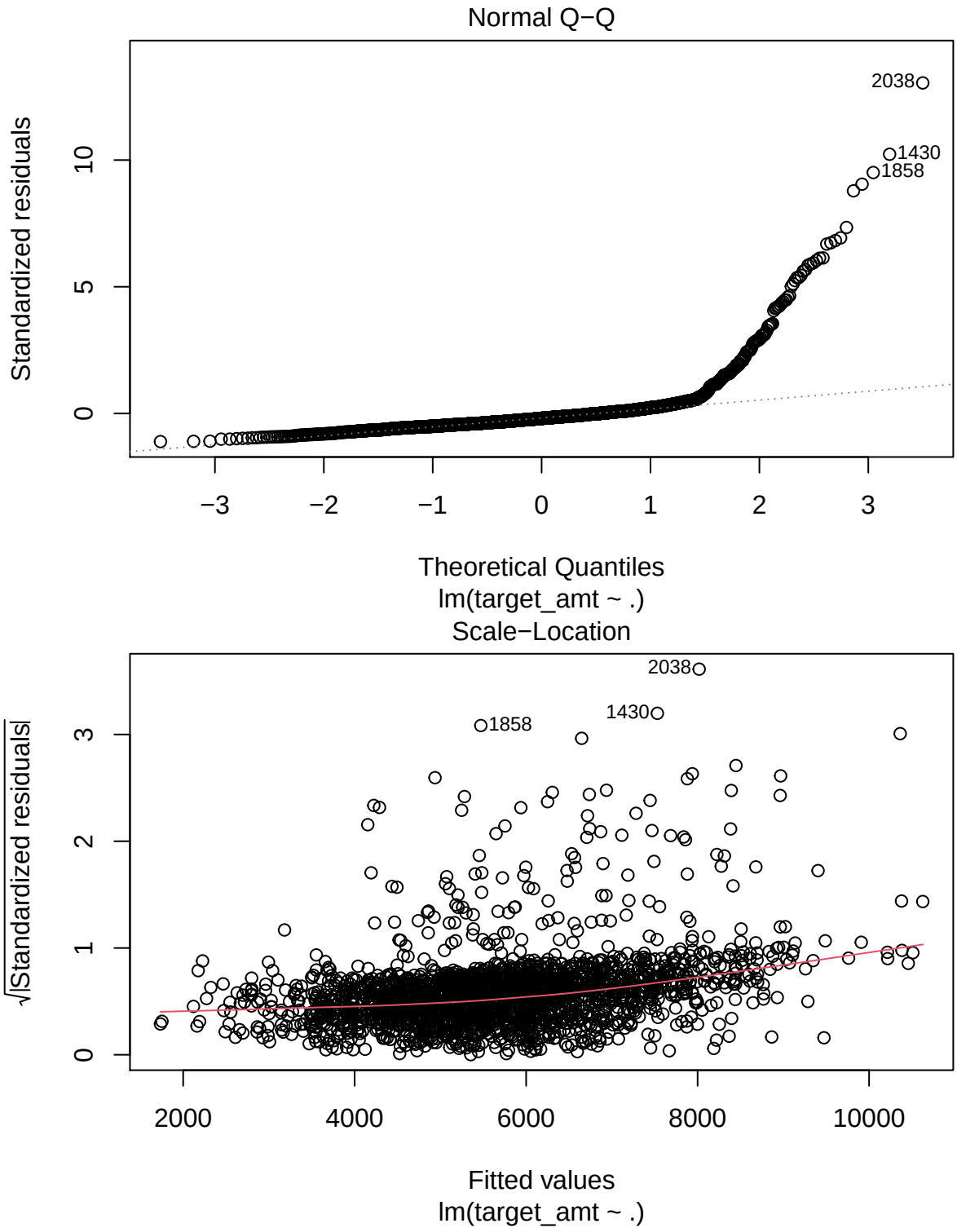
Residuals: Min 1Q Median 3Q Max -8468 -3162 -1474 460 99568

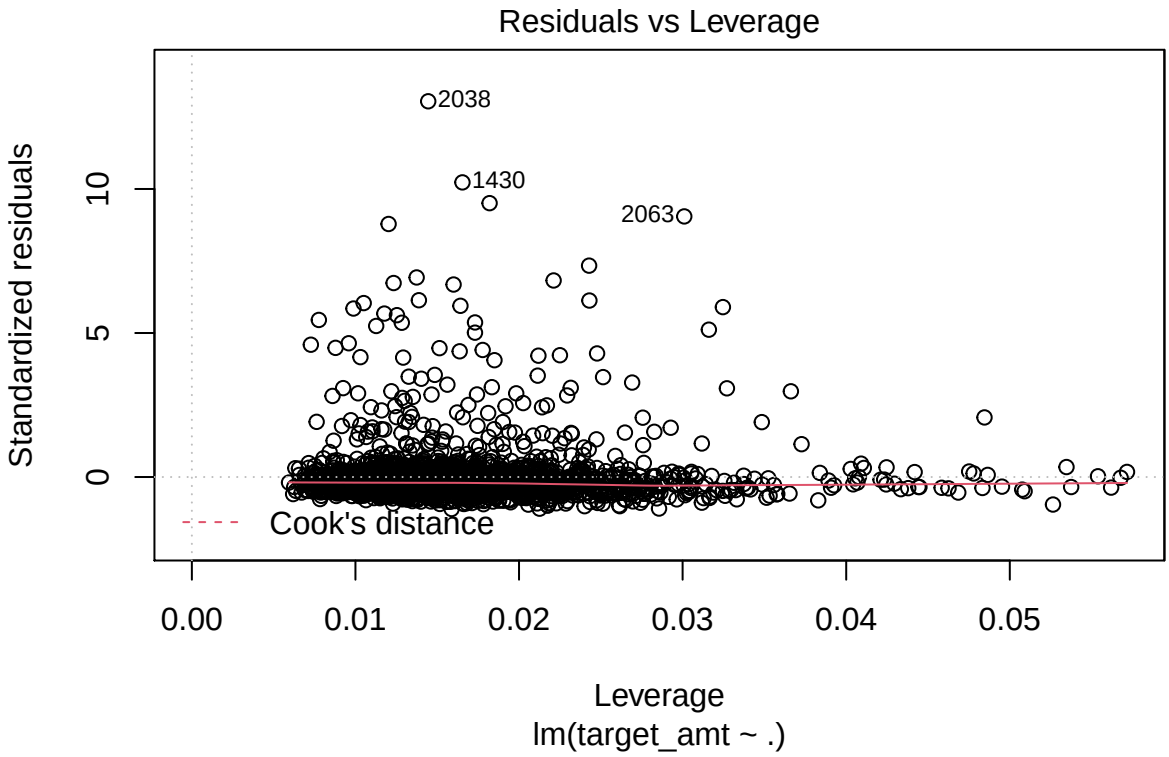
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.033e+03 1.578e+03 1.922 0.0548 .
kidsdriv -1.662e+02 3.159e+02 -0.526 0.5988
homekids 2.103e+02 2.073e+02 1.014 0.3105
parent1Y 2.505e+02 5.876e+02 0.426 0.6699
mstatusY -8.665e+02 5.069e+02 -1.710 0.0875 .
sexM 1.386e+03 6.566e+02 2.111 0.0349 *
educationBachelors 6.246e+02 5.034e+02 1.241 0.2148
educationMasters 1.230e+03 8.847e+02 1.390 0.1647
educationPhD 2.713e+03 1.148e+03 2.362 0.0183 *
travtime 1.133e-01 1.107e+01 0.010 0.9918
car_usePrivate -4.511e+02 4.904e+02 -0.920 0.3577
bluebook 1.255e-01 3.054e-02 4.109 4.12e-05 ** *tif -1.562e+01 4.251e+01 -0.367 0.7133*
car_typePanel Truck -4.797e+02 9.474e+02 -0.506 0.6127
car_typePickup -3.304e+01 5.933e+02 -0.056 0.9556
car_typeSports Car 1.027e+03 7.493e+02 1.371 0.1706
car_typeSUV 8.862e+02 6.664e+02 1.330 0.1837
car_typeVan 1.168e+02 7.640e+02 0.153 0.8785
red_carY -1.697e+02 4.967e+02 -0.342 0.7327
oldclaim 2.551e-02 2.262e-02 1.128 0.2596
clm_freq -1.127e+02 1.580e+02 -0.713 0.4759
revokedY -1.139e+03 5.163e+02 -2.205 0.0276
mvr_pts 1.106e+02 6.843e+01 1.616 0.1062
urbanicityUrban 1.036e+02 7.560e+02 0.137 0.8910
car_age -9.794e+01 4.532e+01 -2.161 0.0308 *

home_val 2.244e-03 2.096e-03 1.071 0.2844
yoj 3.080e+01 4.913e+01 0.627 0.5309
income -1.315e-02 7.039e-03 -1.868 0.0619 .
age 1.731e+01 2.124e+01 0.815 0.4152
jobClerical -2.157e+02 5.810e+02 -0.371 0.7105
jobDoctor -1.725e+03 1.728e+03 -0.998 0.3184
jobHome Maker -5.605e+02 8.658e+02 -0.647 0.5175
jobLawyer 4.534e+02 1.021e+03 0.444 0.6570
jobManager -9.318e+02 7.994e+02 -1.166 0.2439
jobProfessional 5.529e+02 6.443e+02 0.858 0.3909
jobStudent -4.728e+02 7.151e+02 -0.661 0.5086
— Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7689 on 2116 degrees of freedom Multiple R-squared: 0.03037, Adjusted R-squared: 0.01433 F-statistic: 1.893 on 35 and 2116 DF, p-value: 0.001253




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 ~ .)



Call: lm(formula = .outcome ~ ., data = dat)

Residuals: Min 1Q Median 3Q Max -8468 -3162 -1474 460 99568

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.033e+03 1.578e+03 1.922 0.0548 .
kidsdriv -1.662e+02 3.159e+02 -0.526 0.5988
homekids 2.103e+02 2.073e+02 1.014 0.3105
parent1Y 2.505e+02 5.876e+02 0.426 0.6699
mstatusY -8.665e+02 5.069e+02 -1.710 0.0875 .
sexM 1.386e+03 6.566e+02 2.111 0.0349 *
educationBachelors 6.246e+02 5.034e+02 1.241 0.2148
educationMasters 1.230e+03 8.847e+02 1.390 0.1647
educationPhD 2.713e+03 1.148e+03 2.362 0.0183 *
travtime 1.133e-01 1.107e+01 0.010 0.9918
car_usePrivate -4.511e+02 4.904e+02 -0.920 0.3577
bluebook 1.255e-01 3.054e-02 4.109 4.12e-05 ** *tif -1.562e+01 4.251e+01 -0.367 0.7133*
car_typePanel Truck -4.797e+02 9.474e+02 -0.506 0.6127
car_typePickup -3.304e+01 5.933e+02 -0.056 0.9556
car_typeSports Car 1.027e+03 7.493e+02 1.371 0.1706
car_typeSUV 8.862e+02 6.664e+02 1.330 0.1837
car_typeVan 1.168e+02 7.640e+02 0.153 0.8785
red_carY -1.697e+02 4.967e+02 -0.342 0.7327
oldclaim 2.551e-02 2.262e-02 1.128 0.2596
clm_freq -1.127e+02 1.580e+02 -0.713 0.4759
revokedY -1.139e+03 5.163e+02 -2.205 0.0276
mvr_pts 1.106e+02 6.843e+01 1.616 0.1062
urbanicityUrban 1.036e+02 7.560e+02 0.137 0.8910
car_age -9.794e+01 4.532e+01 -2.161 0.0308 *
home_val 2.244e-03 2.096e-03 1.071 0.2844
yoj 3.080e+01 4.913e+01 0.627 0.5309

income -1.315e-02 7.039e-03 -1.868 0.0619 .
age 1.731e+01 2.124e+01 0.815 0.4152
jobClerical -2.157e+02 5.810e+02 -0.371 0.7105
jobDoctor -1.725e+03 1.728e+03 -0.998 0.3184
`jobHome Maker` -5.605e+02 8.658e+02 -0.647 0.5175
jobLawyer 4.534e+02 1.021e+03 0.444 0.6570
jobManager -9.318e+02 7.994e+02 -1.166 0.2439
jobProfessional 5.529e+02 6.443e+02 0.858 0.3909
jobStudent -4.728e+02 7.151e+02 -0.661 0.5086
— Signif. codes: 0 ‘***’ **0.001** ‘**’ *0.01* ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7689 on 2116 degrees of freedom Multiple R-squared: 0.03037, Adjusted R-squared: 0.01433 F-statistic: 1.893 on 35 and 2116 DF, p-value: 0.001253 ### Cost Model 2: Feature Reduction

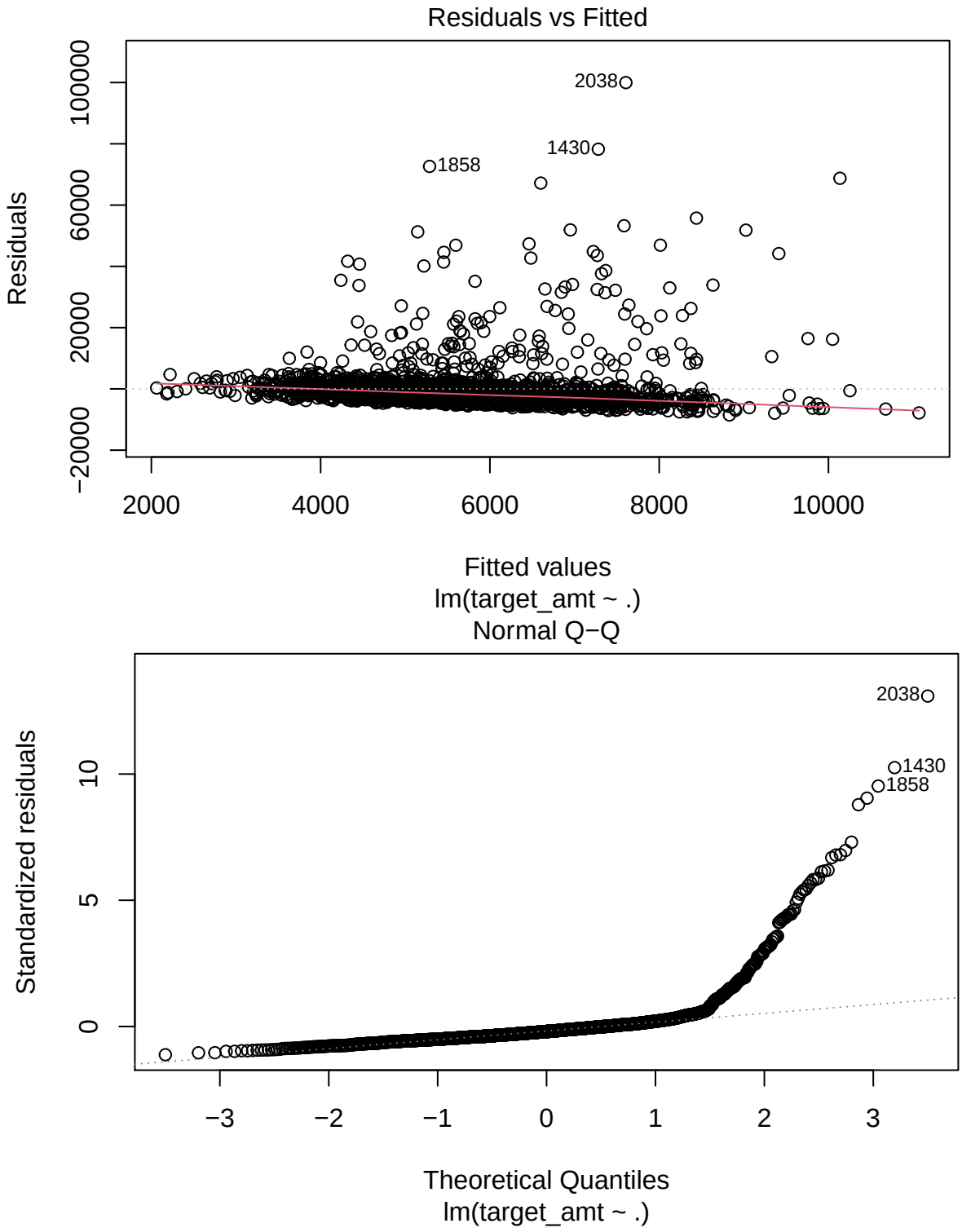
By removing some of the variables we earmarked earlier in the analysis, we can see a reduction in the Residual Standard Error. -parent1, -age, -homekids, - kidsdriv, -red_car, -urbanicity,-job

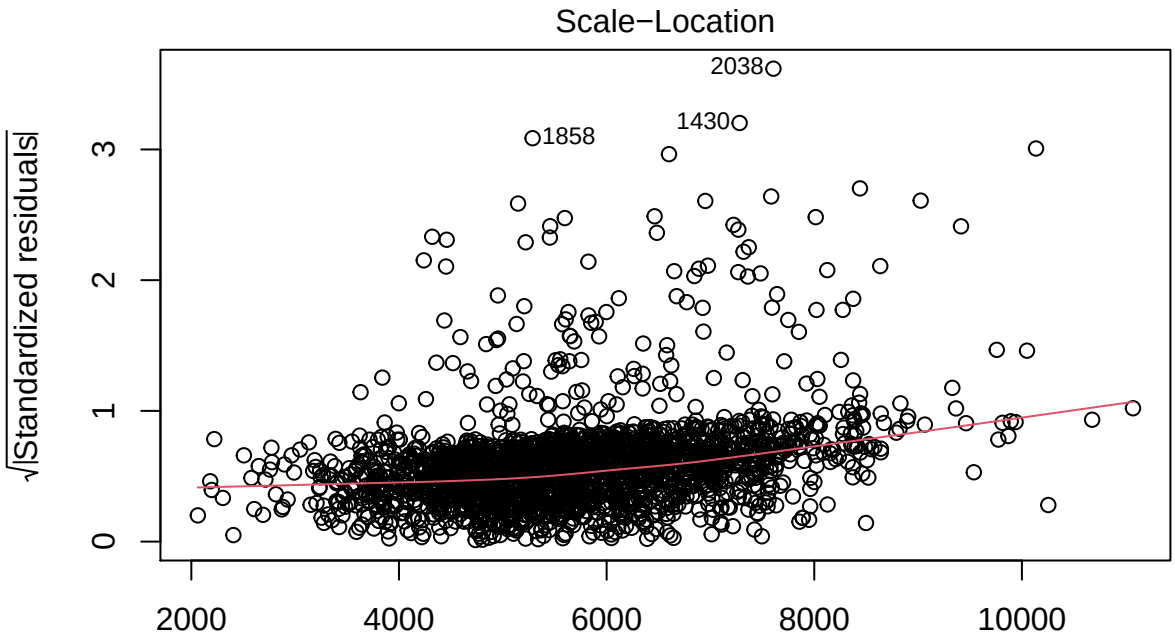
Call: lm(formula = target_amt ~ ., data = dfcrashm2)

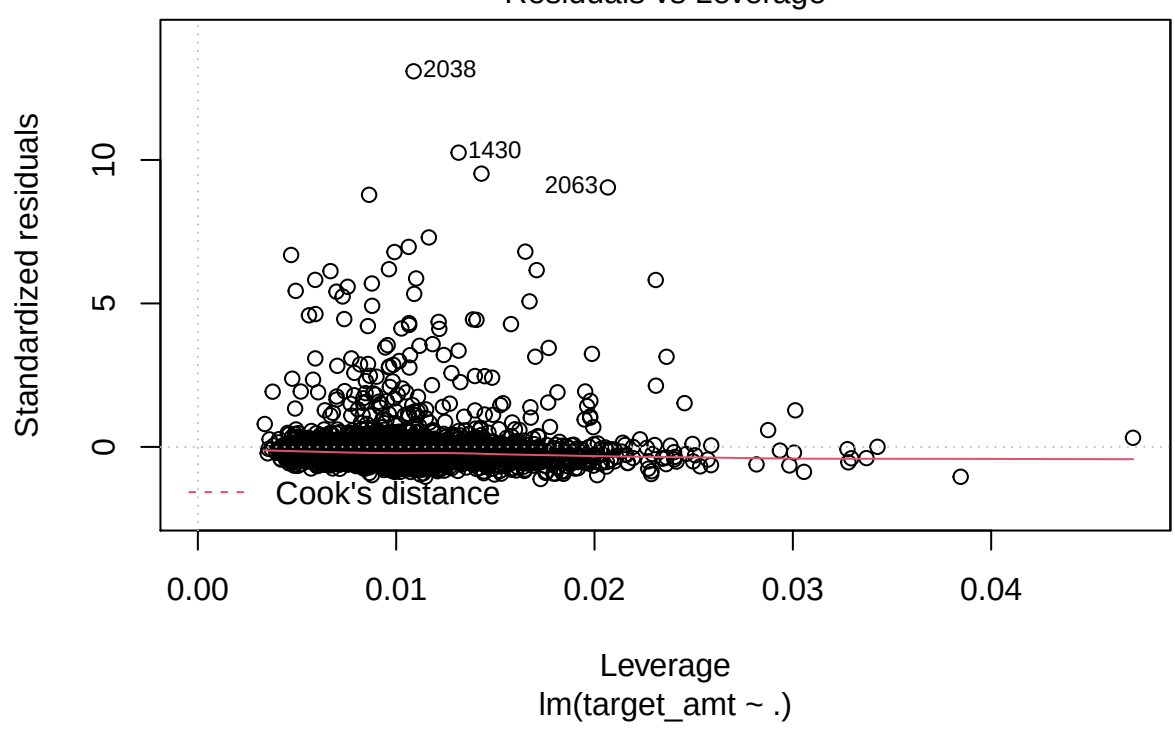
Residuals: Min 1Q Median 3Q Max -8514 -3149 -1511 447 99979

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.571e+03 1.038e+03 3.440 0.000592 ***
***mstatusY -9.389e+02 4.173e+02 -2.250 0.024549 ***
sexM 1.262e+03 5.817e+02 2.169 0.030167 *
educationBachelors 7.134e+02 4.784e+02 1.491 0.136088
educationMasters 1.310e+03 7.148e+02 1.833 0.066881 .
educationPhD 2.060e+03 9.907e+02 2.079 0.037699 *
travtime 8.560e-01 1.099e+01 0.078 0.937916
car_usePrivate -4.474e+02 4.106e+02 -1.090 0.275940
bluebook 1.284e-01 3.004e-02 4.273 2.01e-05 *tif -1.242e+01 4.234e+01 -0.293 0.769321*
car_typePanel Truck -5.741e+02 9.147e+02 -0.628 0.530293
car_typePickup -9.923e+01 5.838e+02 -0.170 0.865038
car_typeSports Car 1.004e+03 7.413e+02 1.354 0.175889
car_typeSUV 8.580e+02 6.577e+02 1.305 0.192195
car_typeVan 5.554e+01 7.533e+02 0.074 0.941238
oldclaim 2.257e-02 2.250e-02 1.003 0.315858
clm_freq -1.155e+02 1.566e+02 -0.737 0.460900
revokedY -1.019e+03 5.115e+02 -1.993 0.046437
mvr_pts 1.228e+02 6.793e+01 1.808 0.070749 .
car_age -9.769e+01 4.518e+01 -2.162 0.030723 *
home_val 2.416e-03 2.040e-03 1.184 0.236511
yoj 5.788e+01 4.224e+01 1.370 0.170762
income -1.220e-02 6.377e-03 -1.914 0.055805 .
— Signif. codes: 0 ‘***’ **0.001** ‘**’ *0.01* ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7680 on 2129 degrees of freedom Multiple R-squared: 0.02656, Adjusted R-squared: 0.0165 F-statistic: 2.641 on 22 and 2129 DF, p-value: 5.05e-05


Residuals vs Fitted
Residuals
Fitted values
lm(target_amt ~ .)
2038
1430
1858
−20000
20000
60000
100000
2000
4000
6000
8000
10000
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(target_amt ~ .)
2038
1430
1858
0
5
10
−3
−2
−1
0
1
2
3

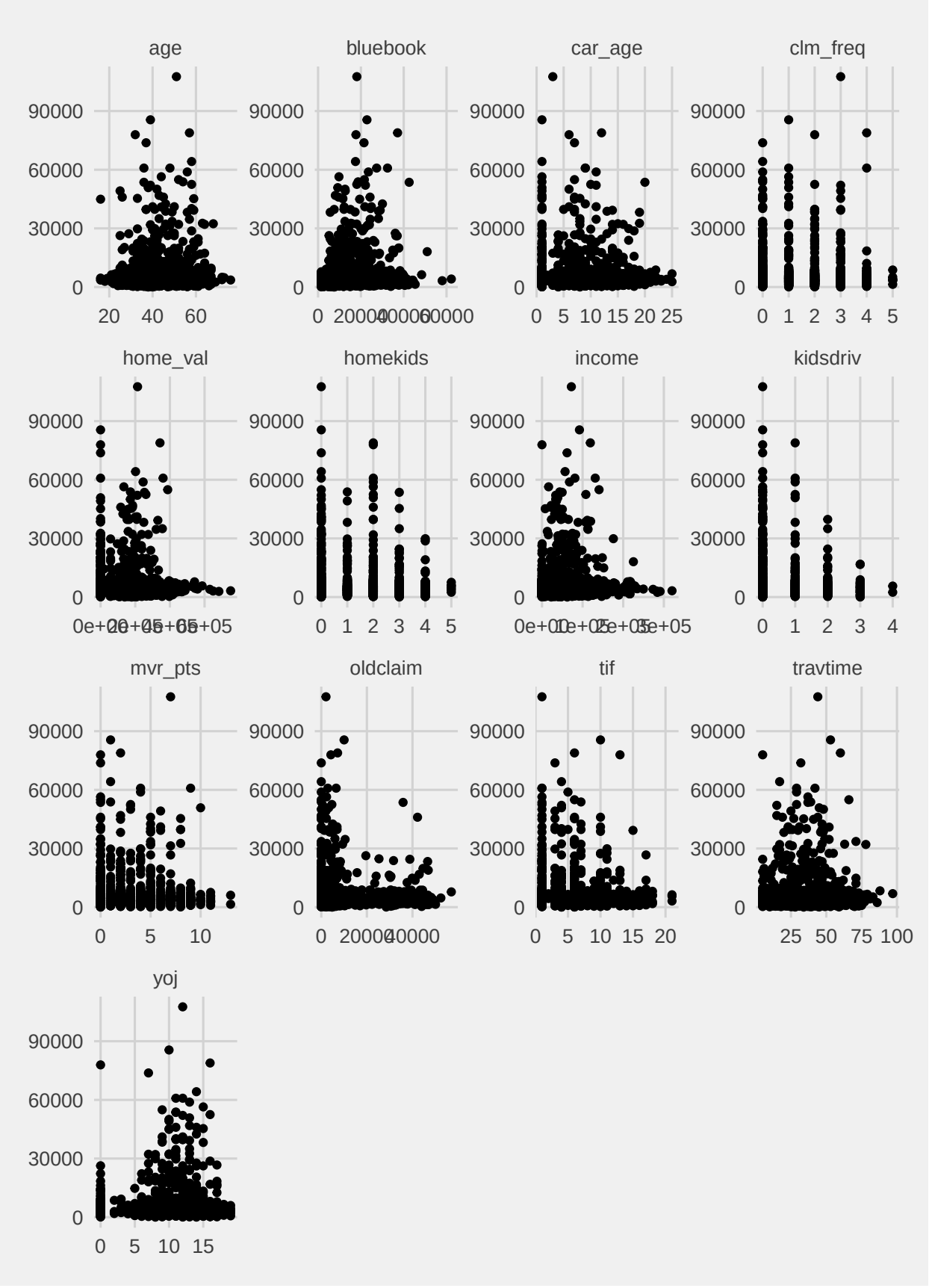




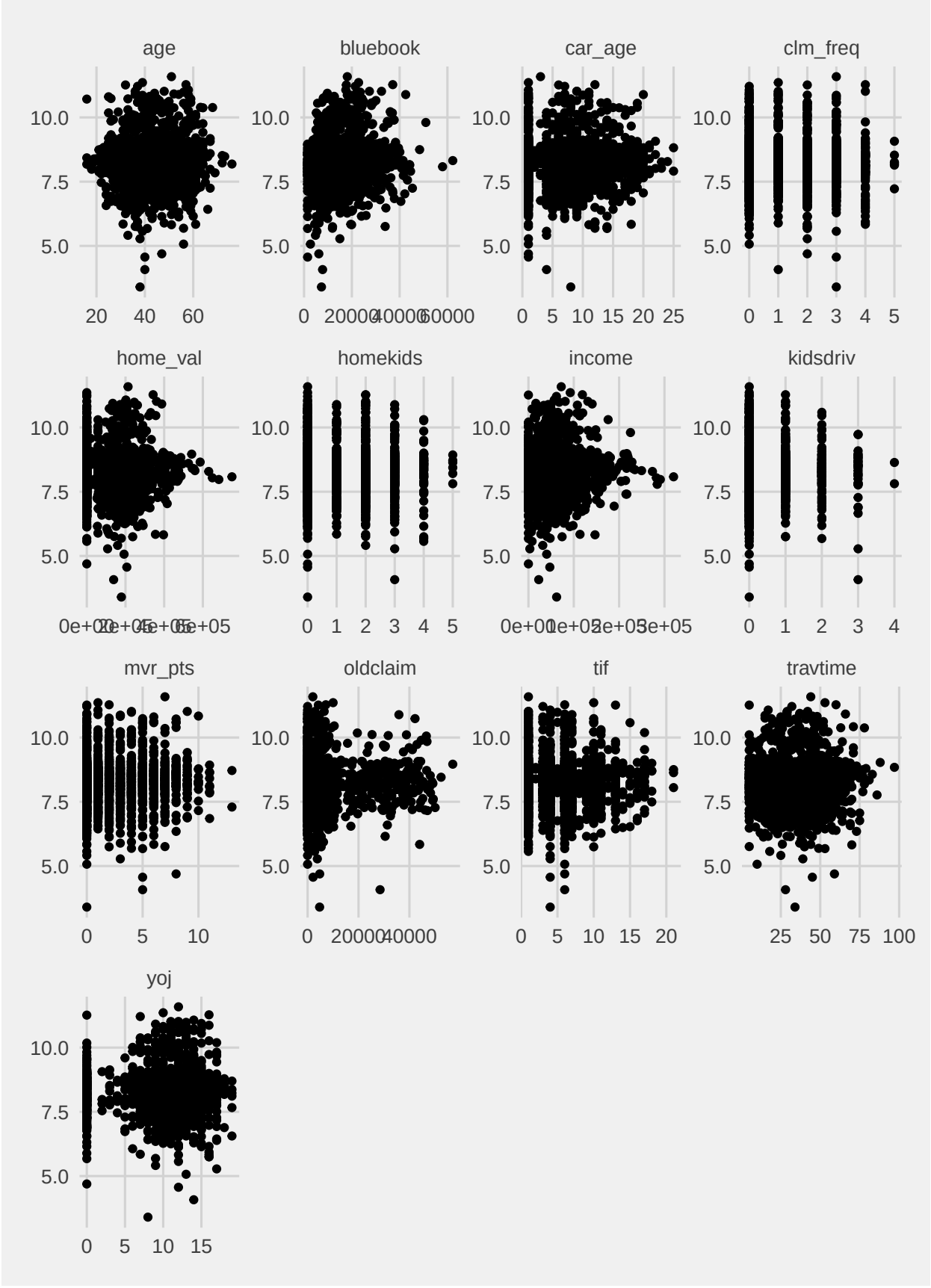
###

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.

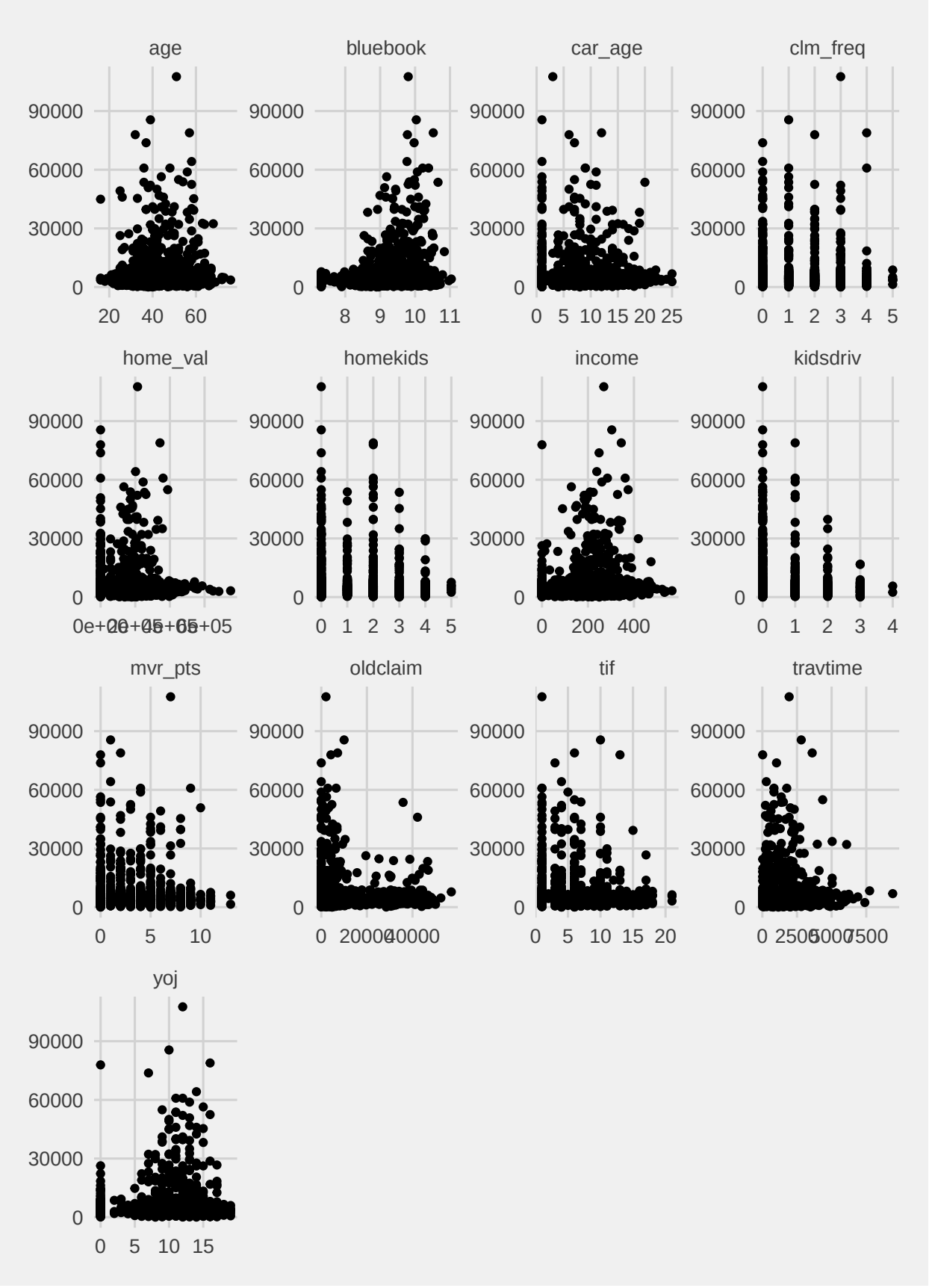

age
bluebook
car_age
clm_freq
home_val
homekids
income
kidsdriv
mvr_pts
oldclaim
tif
travtime
yoj

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

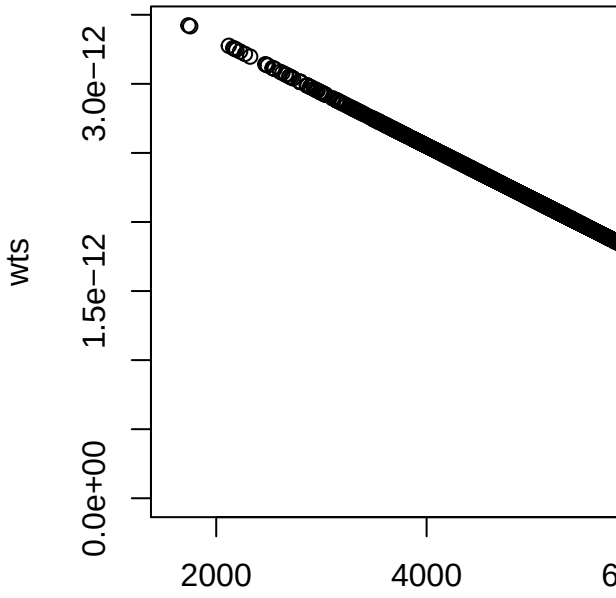

age
bluebook
car_age
clm_freq
home_val
homekids
income
kidsdriv
mvr_pts
oldclaim
tif
travtime
yoj

ticed above.

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


age
bluebook
car_age
clm_freq
home_val
homekids
income
kidsdriv
mvr_pts
oldclaim
tif
travtime
yoj

Apply For our third model, we move to include weights. The first step shown below is the calculation of 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. 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 behave along the scale of the response.

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

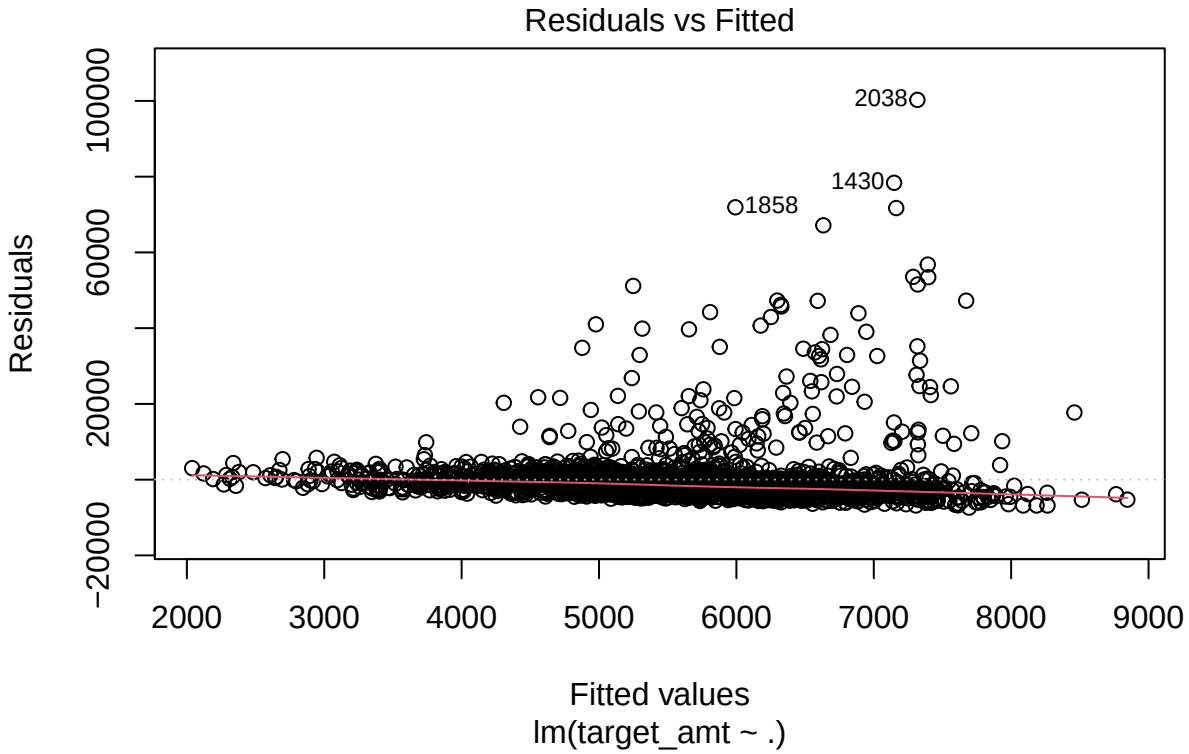
Call: lm(formula = target_amt ~ ., data = dfcrashm3, weights = wts)

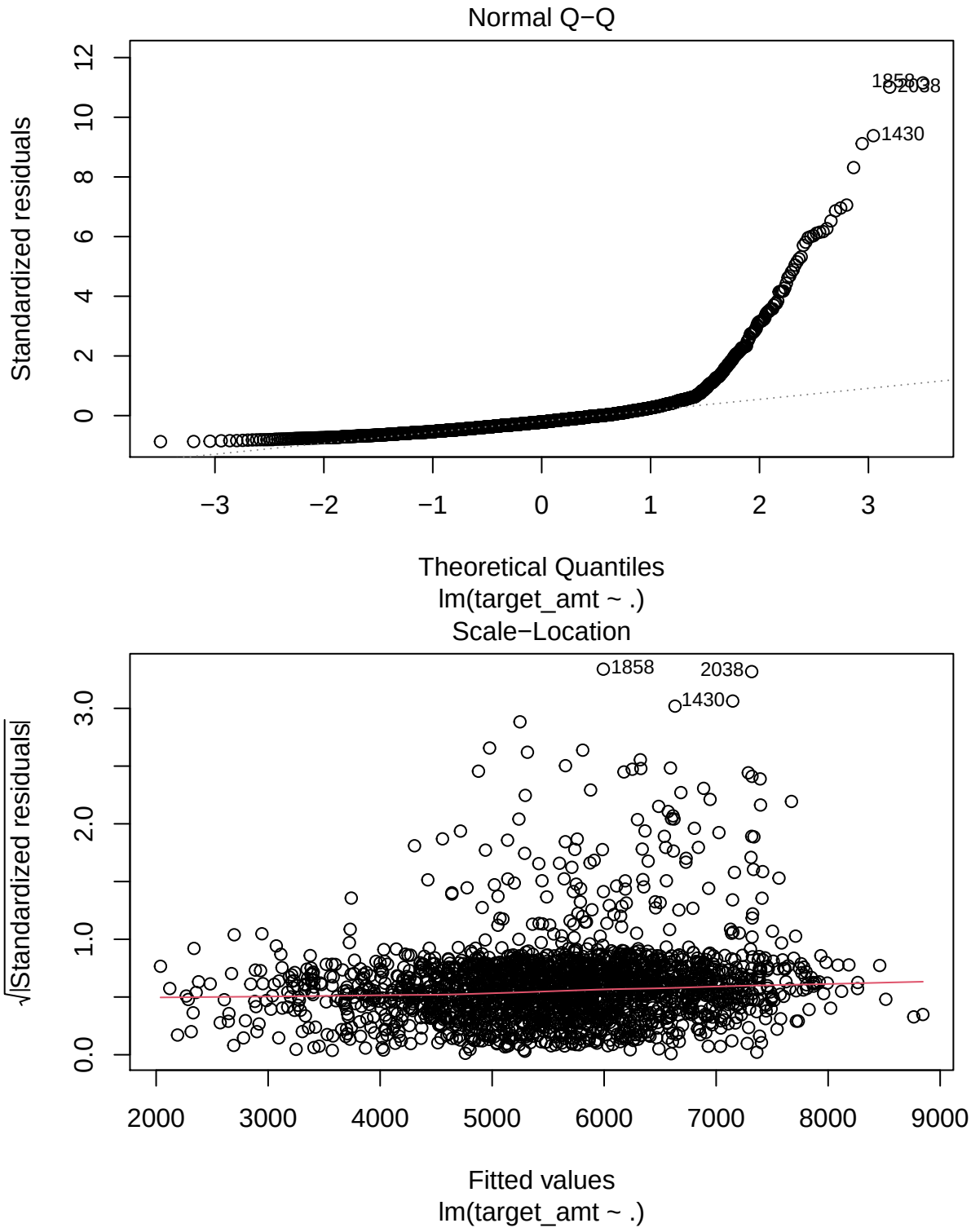
Weighted Residuals: Min 1Q Median 3Q Max -0.007905 -0.003983 -0.001955 0.000526 0.101315

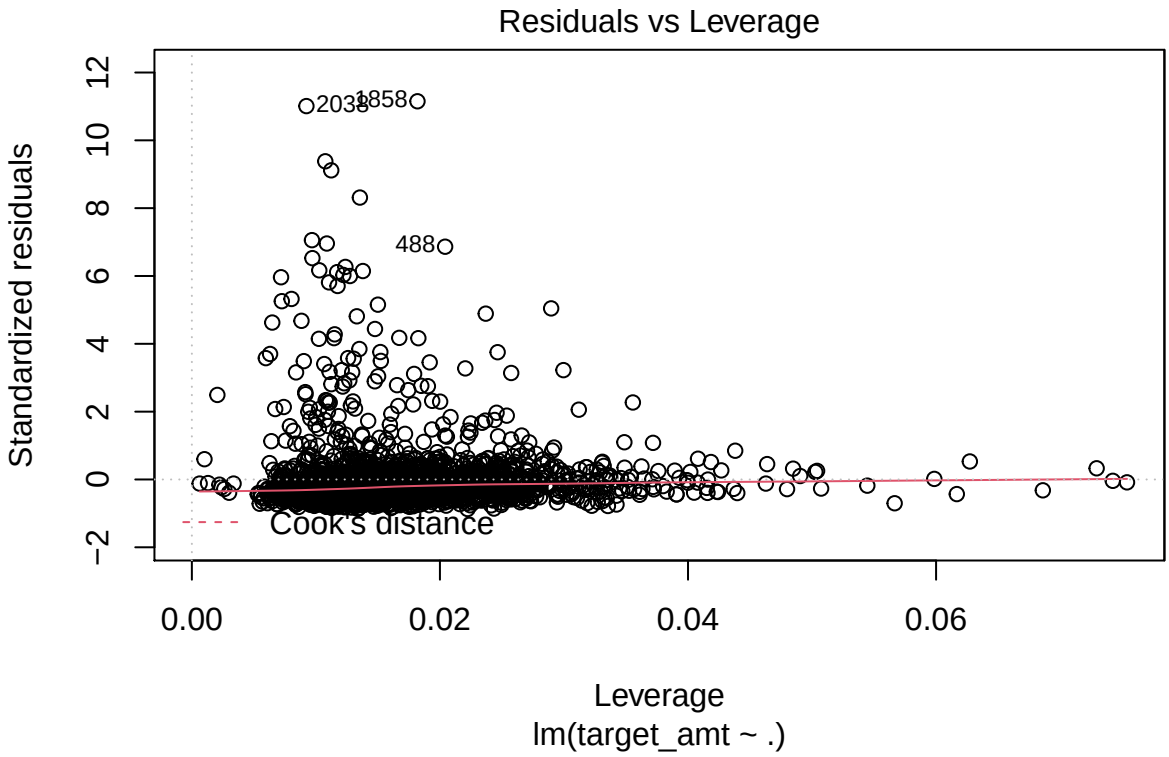
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.433e+03 2.857e+03 -2.252 0.0244 *
kidsdriv -1.457e+02 2.723e+02 -0.535 0.5927
homekids 1.173e+02 1.788e+02 0.656 0.5119
parent1Y 7.492e+01 5.166e+02 0.145 0.8847
mstatusY -4.046e+02 4.444e+02 -0.910 0.3627
sexM 8.351e+02 5.649e+02 1.478 0.1395
educationBachelors 1.875e+02 4.420e+02 0.424 0.6715
educationMasters 7.788e+02 7.957e+02 0.979 0.3278
educationPhD 1.687e+03 1.071e+03 1.575 0.1153
travtime -7.691e-02 1.285e-01 -0.598 0.5496
car_usePrivate -1.848e+02 4.208e+02 -0.439 0.6606
bluebook 1.237e+03 2.758e+02 4.484 7.72e-06 ** *tif -7.845e+00 3.657e+01 -0.215 0.8302*
car_typePanel Truck -1.891e+02 8.461e+02 -0.223 0.8232
car_typePickup -1.243e+02 5.058e+02 -0.246 0.8059
car_typeSports Car 7.211e+02 6.316e+02 1.142 0.2537
car_typeSUV 4.667e+02 5.510e+02 0.847 0.3971
car_typeVan 1.716e+02 7.133e+02 0.241 0.8100
red_carY -5.020e+01 4.477e+02 -0.112 0.9107
oldclaim 2.616e-02 1.925e-02 1.359 0.1744
clm_freq -1.899e+02 1.366e+02 -1.390 0.1646
revokedY -1.043e+03 4.262e+02 -2.448 0.0144
mvr_pts 9.614e+01 6.120e+01 1.571 0.1164
urbanicityUrban 3.184e+02 6.463e+02 0.493 0.6223

car_age -6.369e+01 4.043e+01 -1.575 0.1154
home_val 7.776e-04 1.873e-03 0.415 0.6780
yoj 3.150e+01 4.421e+01 0.712 0.4763
income -2.587e+00 3.042e+00 -0.850 0.3953
age 1.019e+01 1.841e+01 0.554 0.5798
jobClerical -1.203e+02 5.061e+02 -0.238 0.8122
jobDoctor -1.594e+03 1.524e+03 -1.046 0.2958
jobHome Maker -4.970e+02 7.969e+02 -0.624 0.5329
jobLawyer 5.135e+01 9.238e+02 0.056 0.9557
jobManager -1.222e+03 6.933e+02 -1.763 0.0781 .
jobProfessional 3.912e+02 5.950e+02 0.657 0.5110
jobStudent -2.244e+02 6.898e+02 -0.325 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



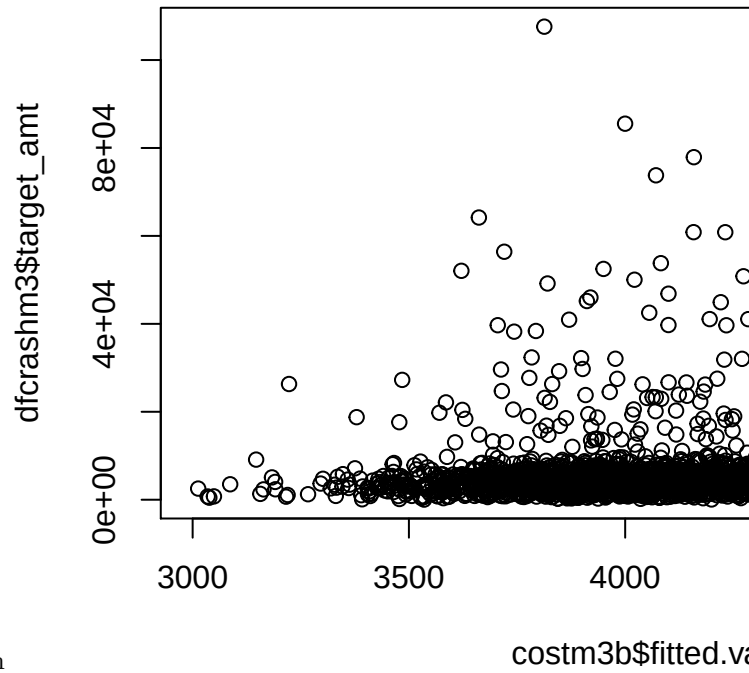

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(target_amt ~ .)
1858
2038
1430
Scale−Location
√|Standardized residuals|
Fitted values
lm(target_amt ~ .)



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 3Q Max -0.0064007 -0.0018924 0.0001064 0.0024489 0.1039949

Coefficients: Value Std. Error t value
(Intercept) 1615.9462 959.1152 1.6848 kidsdriv -89.1719 91.4198 -0.9754 homekids 104.6917 60.0427 1.7436 parent1Y -319.9879 173.4188 -1.8452 mstatusY -320.6395 149.1890 -2.1492 sexM -28.3388 189.6396 -0.1494 educationBachelors -283.5063 148.3931 -1.9105 educationMasters 103.1816 267.1437 0.3862 educationPhD 192.9405 359.5514 0.5366 travtime 0.0074 0.0431 0.1722 car_usePrivate 54.6428 141.2733 0.3868 bluebook 230.5048 92.5909 2.4895 tif 4.9114 12.2767 0.4001 car_typePanel Truck 171.8550 284.0551 0.6050 car_typePickup 126.1302 169.8115 0.7428 car_typeSports Car 51.7434 212.0460 0.2440 car_typeSUV 35.3362 184.9845 0.1910 car_typeVan 10.3613 239.4623 0.0433 red_carY -10.4762 150.3058 -0.0697 oldclaim 0.0044 0.0065 0.6759 clm_freq -83.0662 45.8690 -1.8109 revokedY 18.1152 143.0876 0.1266 mvr_pts 55.3398 20.5470 2.6933 urbanicityUrban 287.3834 216.9791 1.3245 car_age 7.8953 13.5732 0.5817 home_val 0.0007 0.0006 1.1831 yoj -2.8462 14.8421 -0.1918 income -1.1106 1.0214 -1.0873 age 2.8436 6.1797 0.4602 jobClerical -13.6442 169.9129 -0.0803 jobDoctor -237.7411 511.7751 -0.4645 jobHome Maker -269.2748 267.5484 -1.0065 jobLawyer -239.6016 310.1383 -0.7726 jobManager 32.2350 232.7371 0.1385 jobProfessional 121.3529 199.7532 0.6075 jobStudent 85.7391 231.5941 0.3702



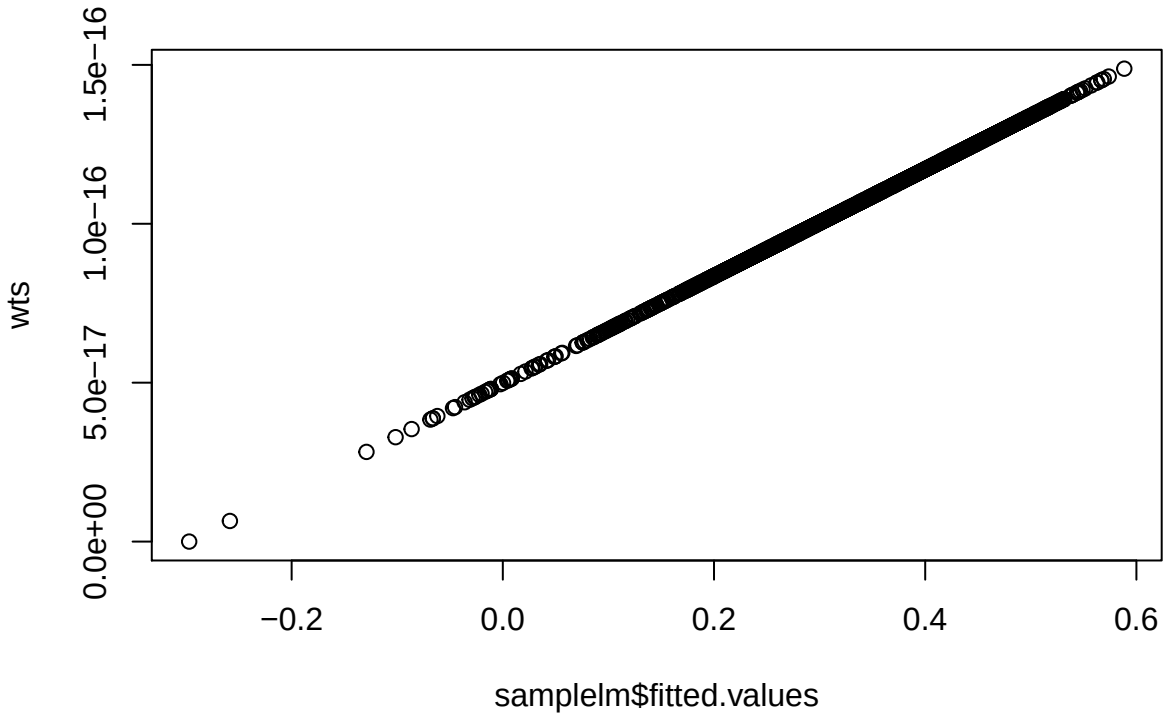
Residual standard error: 0.003109 on 2116 degrees of freedom

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 usless, 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.



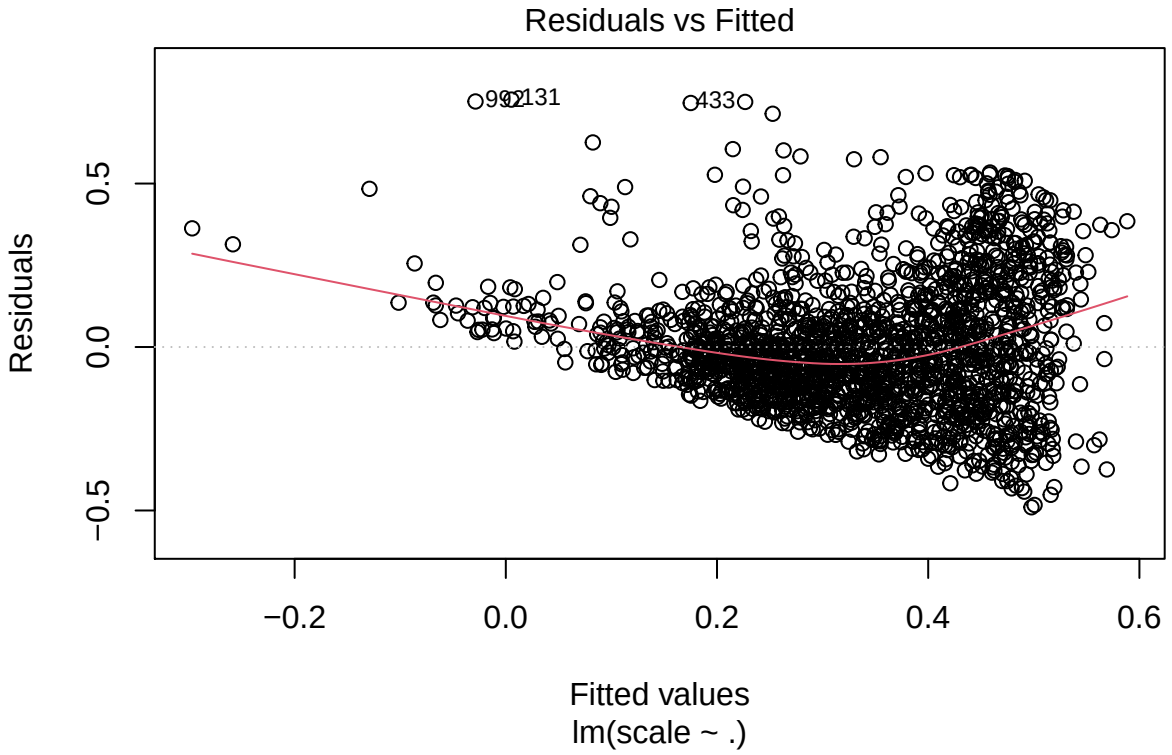
Call: lm(formula = scale ~ ., data = dfcrashm5)

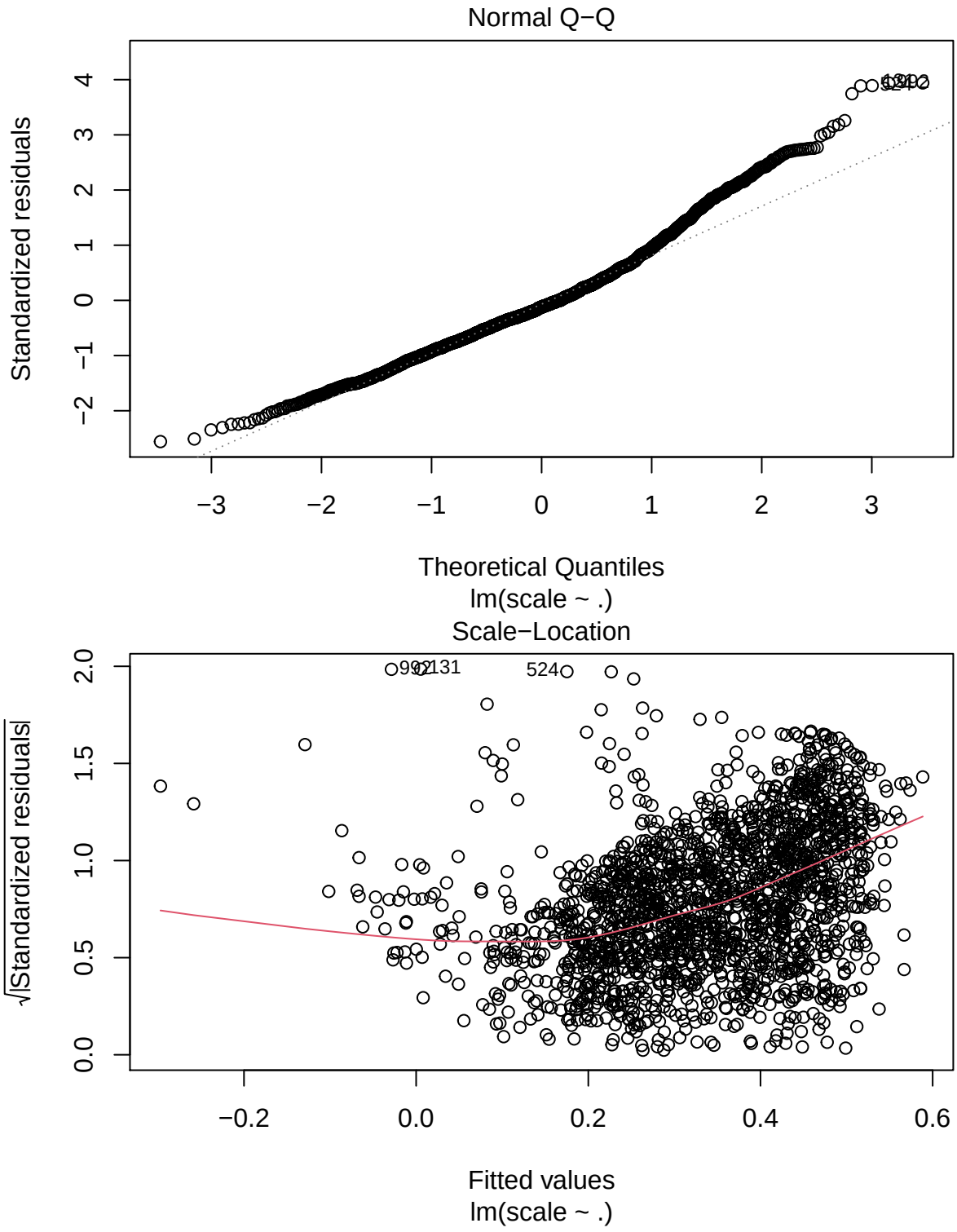
Residuals: Min 1Q Median 3Q Max -0.49020 -0.12789 -0.02344 0.10167 0.75679

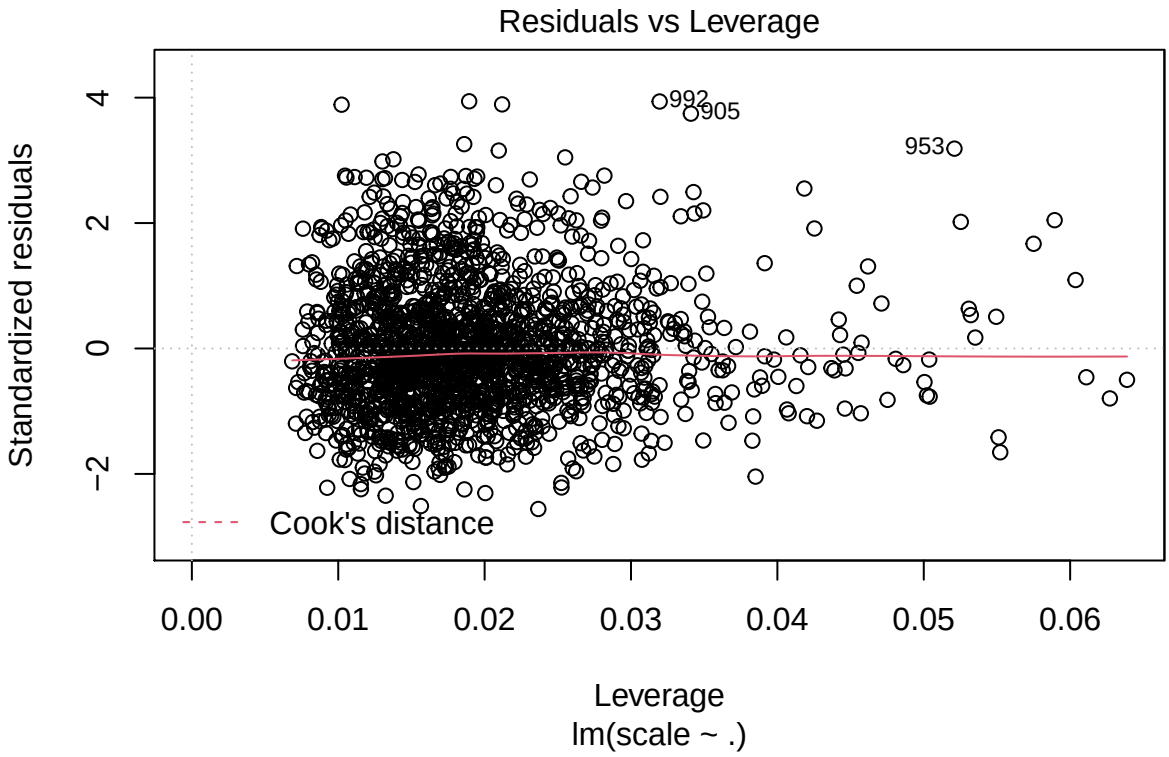
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.493e-01 4.293e-02 12.795 < 2e-16 ***kidsdriv -3.070e-04 8.504e-03 -0.036 0.97121***
homekids 3.677e-03 5.628e-03 0.653 0.51365
parent1Y -3.524e-03 1.596e-02 -0.221 0.82529
mstatusY -2.042e-02 1.374e-02 -1.486 0.13747
sexM -9.401e-03 1.745e-02 -0.539 0.59024
educationBachelors -1.367e-02 1.370e-02 -0.998 0.31854
educationMasters 1.373e-02 2.376e-02 0.578 0.56352
educationPhD 5.414e-02 3.101e-02 1.746 0.08103 .
travtime 7.277e-05 3.004e-04 0.242 0.80860
car_usePrivate 1.096e-02 1.337e-02 0.820 0.41248
bluebook -1.524e-05 8.539e-07 -17.844 < 2e-16 tif -8.546e-04 1.160e-03 -0.737 0.46135
car_typePanel Truck 9.714e-02 2.507e-02 3.875 0.00011 ***car_typePickup 4.277e-02 1.610e-02 2.656 0.00796*** *car_typeSports Car 2.194e-02 2.025e-02 1.084 0.27863*
car_typeSUV 1.705e-02 1.784e-02 0.956 0.33919
car_typeVan -7.033e-03 2.019e-02 -0.348 0.72767
red_carY 8.610e-03 1.343e-02 0.641 0.52140
oldclaim 9.930e-07 6.142e-07 1.617 0.10613
clm_freq -9.222e-03 4.272e-03 -2.159 0.03099
revokedY -6.134e-03 1.381e-02 -0.444 0.65694
mvr_pts 1.832e-03 1.864e-03 0.983 0.32575
urbanicityUrban 1.639e-02 2.075e-02 0.790 0.42967
car_age -2.827e-05 1.219e-03 -0.023 0.98151
home_val 4.538e-08 5.673e-08 0.800 0.42378
yoj 1.597e-03 1.347e-03 1.185 0.23613
income -2.812e-07 1.886e-07 -1.491 0.13601
age -6.759e-04 5.777e-04 -1.170 0.24217
jobClerical 1.384e-02 1.575e-02 0.879 0.37971

jobDoctor 1.535e-02 4.572e-02 0.336 0.73719
jobHome Maker 6.375e-03 2.344e-02 0.272 0.78573
jobLawyer -2.341e-02 2.763e-02 -0.847 0.39697
jobManager 6.398e-04 2.122e-02 0.030 0.97595
jobProfessional 1.555e-02 1.718e-02 0.905 0.36542
jobStudent 1.929e-02 1.955e-02 0.987 0.32385
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1

Residual standard error: 0.1938 on 1837 degrees of freedom Multiple R-squared: 0.2913, Adjusted R-squared: 0.2778 F-statistic: 21.58 on 35 and 1837 DF, p-value: $<$ 2.2e-16




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 Std. Error z value Pr(>|z|)
(Intercept) 4.484e-01 1.812e-01 2.475 0.0133 *
age -4.135e-03 2.352e-03 -1.758 0.0787 .
sexM -4.872e-02 7.807e-02 -0.624 0.5326
mstatusY -7.526e-02 5.678e-02 -1.325 0.1850
tif -1.092e-03 5.116e-03 -0.213 0.8310
red_carY 5.826e-02 5.931e-02 0.982 0.3260
car_age 2.444e-03 5.401e-03 0.453 0.6509
home_val 1.336e-07 2.536e-07 0.527 0.5985
parent1Y -1.468e-02 6.017e-02 -0.244 0.8073
educationBachelors -8.074e-02 6.051e-02 -1.334 0.1821
educationMasters 6.031e-02 1.057e-01 0.570 0.5684
educationPhD 1.661e-01 1.385e-01 1.199 0.2304
car_usePrivate 5.756e-02 5.843e-02 0.985 0.3245
car_typePanel Truck 4.565e-01 1.137e-01 4.016 5.91e-05 ***car_typePickup 1.698e-01 7.086e-02 2.396 0.0166***
car_typeSports Car 1.068e-01 8.998e-02 1.186 0.2354
car_typeSUV 7.940e-02 7.933e-02 1.001 0.3169
car_typeVan 2.911e-02 9.140e-02 0.318 0.7501
oldclaim 3.898e-06 2.700e-06 1.443 0.1489
revokedY -3.773e-02 6.074e-02 -0.621 0.5345
urbanicityUrban 7.898e-02 9.066e-02 0.871 0.3837
jobClerical 3.634e-02 6.912e-02 0.526 0.5991
jobDoctor 1.597e-01 2.032e-01 0.786 0.4319
jobHome Maker 1.304e-02 9.598e-02 0.136 0.8920

jobLawyer -1.103e-01 1.229e-01 -0.897 0.3695
jobManager 1.361e-02 9.481e-02 0.144 0.8858
jobProfessional 8.667e-02 7.619e-02 1.138 0.2553
jobStudent 1.376e-02 8.016e-02 0.172 0.8637
travtime -1.815e-04 1.324e-03 -0.137 0.8909
bluebook -6.904e-05 3.895e-06 -17.726 < 2e-16 * mvr_pts 9.131e-03 8.215e-03 1.112 0.2664
clm_freq -3.487e-02 1.886e-02 -1.849 0.0644 .
income -1.331e-06 8.394e-07 -1.585 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