Report-Statistican Report

Zuojun Gong

December 16, 2016

```r
# Read in the data file
init.data = read.csv("damages.csv")
```

Analysis Goal:

The relationship between the total damages and the amount of damages demanded by the plaintiff

The group hypothesizes that the following ar predictive of a higher probability that damages are awarded, of a higher amount: a. A more recent trail b. Defendent is a corporation c. Increased number of plaintiffs d. Interaction between bodily injury, total damages, and amount demanded

Data

Our Data Set contains the following Variables: TOTDAM: Total amount of damages awarded to plaintiff

DEMANDED: total amount of damages requested from the court by plaintiff (in $)

TRIDAYS: how many days the trial lasted

BODINJ: whether or not a bodily injury was part of the claim (1 - Yes; 0 - No)

DECORP: whether or not the defendant was a corporation (1 - Yes; 0 - No)

DEGOVT: whether or not the defendant was the government (1 - Yes; 0 - No)

YEAR: year the civil lawsuit was filed - categorized as follows: 1: pre-1997; 2: 1997; 3: 1998; 4: 1999; 5: 2000; 6: 2001

CLAIMTYPE: type of claim the plaintiff made 1: motor vehicle; 2: premises liability; 3: malpractice; 4: fraud; 5: rental/lease ; 6: other

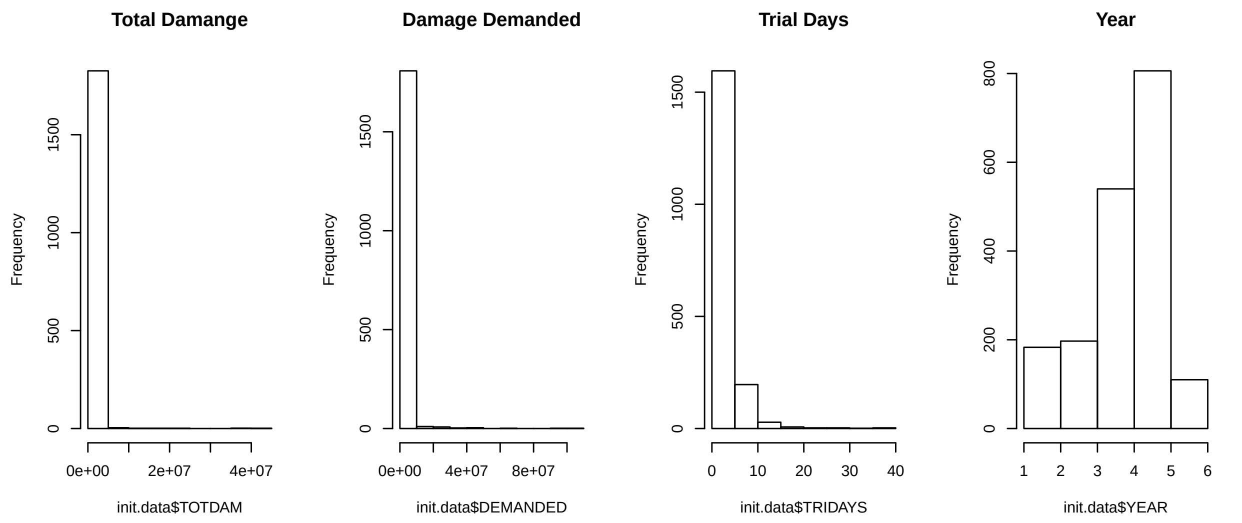
TOTALNOPL: total number of plaintiffs - categorized as follows: 1: if one plaintiff ; 2: if two plaintiffs ; 3: if >= 3 plaintiffs

TOTALNODE: total number of defendants - categorized as follows: 1: if one defendant; 2: if two defendants; 3: if >= 3 defendants

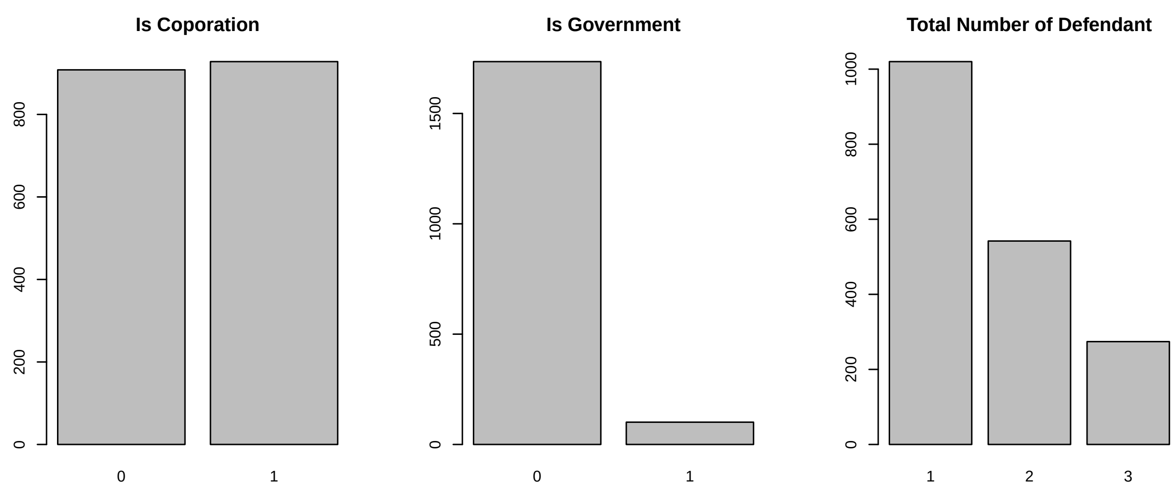
EDA

Univariate

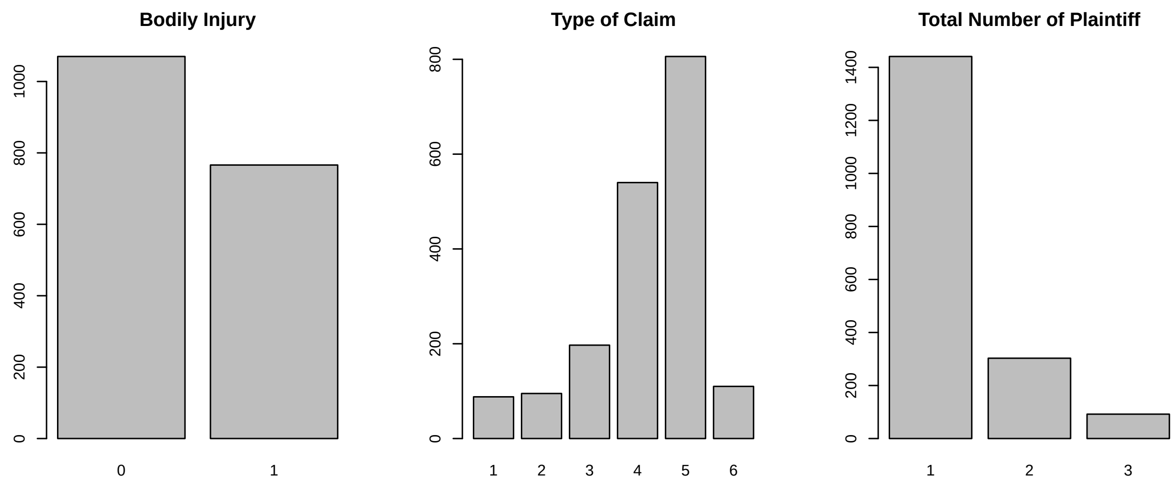
```r
# kable(summary(init.data))

# Plot EDA for the entire dataset
par(mfrow = c(1, 4))
hist(init.data$TOTDAM, main = "Total Damange")
hist(init.data$DEMANDED, main = "Damage Demanded")
hist(init.data$TRIDAYS, main = "Trial Days")
hist(init.data$YEAR, main = "Year", breaks = 6)
```




```r
par(mfrow = c(1, 3))
barplot(table(init.data$DECORP), main = "Is Coporation")
barplot(table(init.data$DEGOVT), main = "Is Government")
barplot(table(init.data$TOTALNODE), main = "Total Number of Defendant")
```




```r
par(mfrow = c(1, 3))
barplot(table(init.data$BODINJ), main = "Bodily Injury")
barplot(table(init.data$YEAR), main = "Type of Claim")
barplot(table(init.data$TOTALNOPL), main = "Total Number of Plaintiff")
```




```
# Compute Precentages table(init.data$DEGOVT) /
# sum(table(init.data$DEGOVT))
```

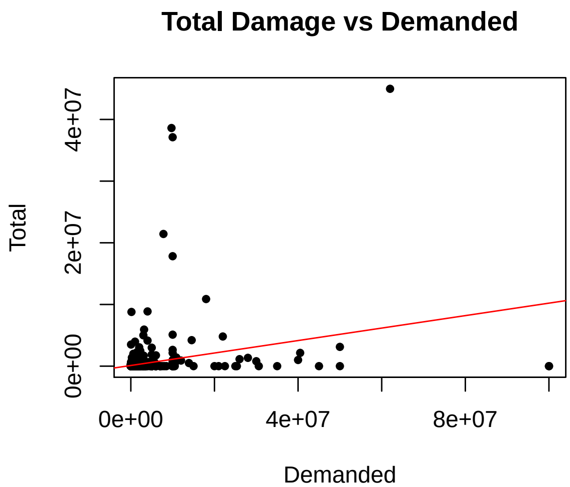
From initial look at the general statistics of the data, it appears that total damage rewarded, damage demanded, and Trial days are very right skewed. This is potentially caused by two issues: Some of the cases had were rewarded with zero award (661), and there are significant outliers in all three variables.

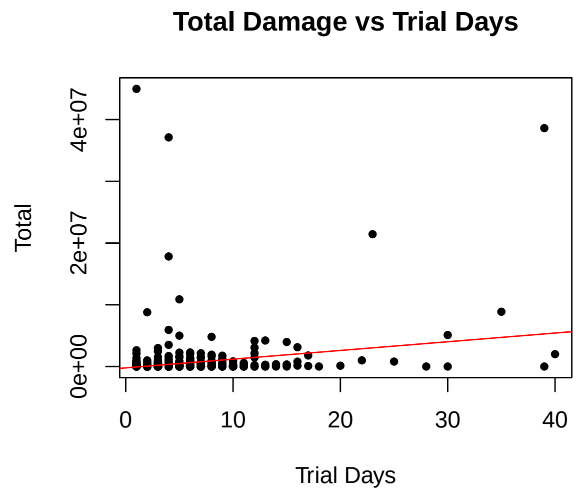
In terms of the type of defendants, it appears that approximately half of them are corporations, and only 5% of the cases involves government as defendants. About 55.5% of the cases involved one defendant.

When it comes to year when the law suit is filed, it appears to be normally distributed with a left skew and a right cutoff, possibly caused by data collection (fewer cases in the most recent year collected). We want to look into the possible relationship between year and response variables to see if regrouping is a possibility. Around 42% of the plaintiffs received have bodily injury as part of the claim, and we can see that 78.5% of the cases had only one plaintiff.

Basic Mutivariate

```
par(mfrow = c(1, 2))
plot(init.data$DEMANDED, init.data$TOTDAM, xlab = "Demanded",
    ylab = "Total", main = "Total Damage vs Demanded", pch = 16,
    cex = 0.8)
abline(lm(init.data$TOTDAM ~ init.data$DEMANDED), col = 2)

plot(init.data$TRIDAYS, init.data$TOTDAM, xlab = "Trial Days",
    ylab = "Total", main = "Total Damage vs Trial Days", pch = 16,
    cex = 0.8)
abline(lm(init.data$TOTDAM ~ init.data$TRIDAYS), col = 2)
```

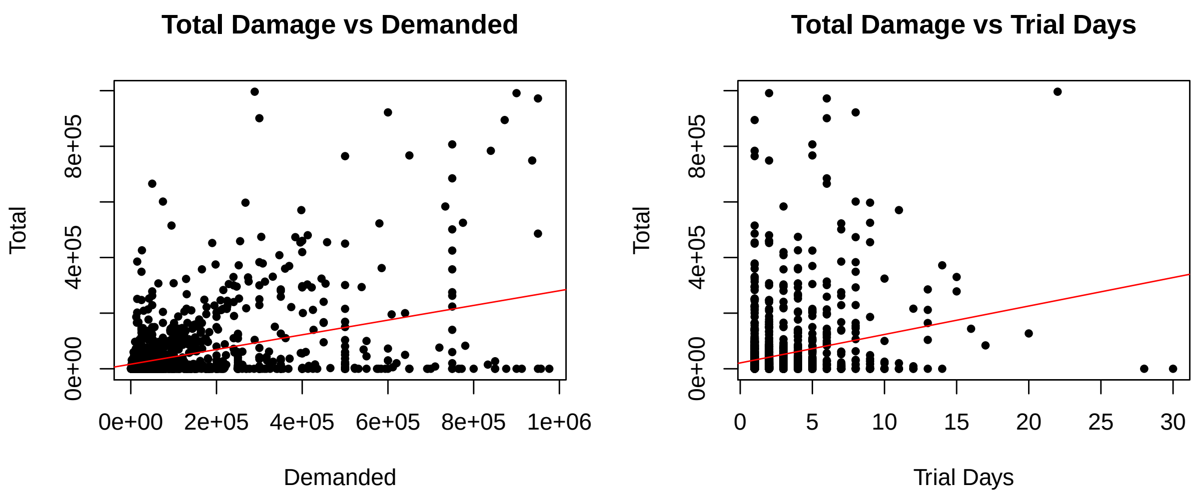





```
init.data[which(init.data$TOTDAM > 1e+06), ]
init.data[which(init.data$TOTDAM > 1e+06), ]

init.data.subset <- init.data[which(init.data$TOTDAM < 1e+06 &
    init.data$DEMANDED < 1e+06), ]
```

We can see that there are several cases with very high demanded and awarded values that departs significantly from the rest of the data. Upon examining the cases, we do not have sufficient to rule them as outliers just yet. We do observe that there are higher portion of these cases have corporates as defendants. Nonetheless, we will keep them during modeling, but exclude them for now to have a better understanding of visually of the relationships between our variables.

```
par(mfrow = c(1, 2))
plot(init.data.subset$DEMANDED, init.data.subset$TOTDAM, xlab = "Demanded",
    ylab = "Total", main = "Total Damage vs Demanded", pch = 16,
    cex = 0.8)
abline(lm(init.data.subset$TOTDAM ~ init.data.subset$DEMANDED),
    col = 2)

plot(init.data.subset$TRIDAYS, init.data.subset$TOTDAM, xlab = "Trial Days",
    ylab = "Total", main = "Total Damage vs Trial Days", pch = 16,
    cex = 0.8)
abline(lm(init.data.subset$TOTDAM ~ init.data.subset$TRIDAYS),
    col = 2)
```



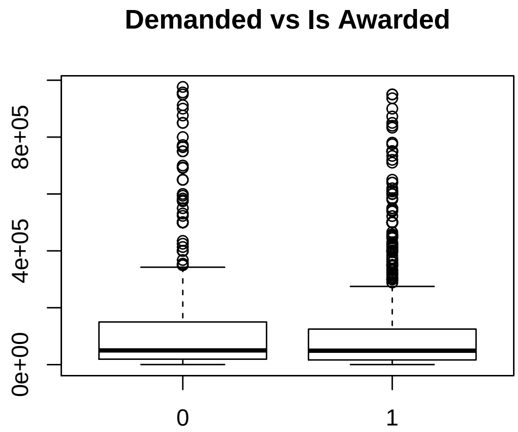
With the subset of data that contains all claims awards under $100K, we can see that there is a positive correlation between the amount demanded by the plaintiff and the amount awarded. There also appear to be a positive correlation between damage awarded and the trial days.

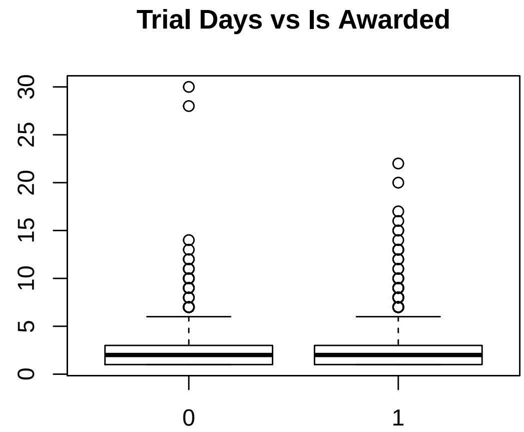
Of all the cases in our dataset, 36% did not award plaintiff with any damages, and therefore we will treat them separately. We will construct two datasets with different response variables: whether the plaintiff is awarded with damage, and if so, how much is the plaintiff awarded. Then we want to fit regression models separately to evaluate the relationship between the two response variables and the other covariates.

```
# Create dummy variable of whether the defendant is awarded
# with damages
df.isaward = init.data
df.isaward[, "is.Awarded"] = ifelse(df.isaward$TOTDAM > 0, 1,
    0)

# Subset of the origional data that contails only cases that
# are awareded with damages
df.awardsize = init.data[which(init.data$TOTDAM > 0), ]
```

Binary Award Dataset Mutivariate

```
df.isaward.sub <- df.isaward[which(df.isaward$TOTDAM < 1e+06 &
    df.isaward$DEMANDED < 1e+06), ]

## Continuous Demanded vs Awarded
par(mfrow = c(1, 2))
boxplot(df.isaward.sub$DEMANDED ~ df.isaward.sub$is.Awarded,
    main = "Demanded vs Is Awarded")
# Trial Days vs Awarded
boxplot(df.isaward.sub$TRIDAYS ~ df.isaward.sub$is.Awarded, main = "Trial Days vs Is Awarded")
```





First we look at the binary data set with whether the plaintiff received the damage award as the predictor variable, and we want to see the relationship between the amount of damage demanded and the number of days the trial lasted to whether the case was awarded. Here we only selected the sub set with award under 500K to help us visualize the majority of the cases better on grpah. We can see that there does not appear to be a relationship between Demand and is.awarded or Trial Days and is.awarded. Their median appears to be similiar for both variables.

```
## Discrete Body injury
eda.table = table(df.isaward$BODINJ, df.isaward$is.Awarded)
colnames(eda.table) = c("AWard", "NoAward")
rownames(eda.table) = c("NotInjured", "Injured")
eda.table
```

```
##
##              AWard NoAward
##   NotInjured   306     764
##   Injured      355     411
```

```
# Crop
eda.table = table(df.isaward$DECORP, df.isaward$is.Awarded)
colnames(eda.table) = c("AWard", "NoAward")
rownames(eda.table) = c("NotCrop", "Crop")
eda.table
```

```
##
##           AWard NoAward
##   NotCrop   360     548
##   Crop      301     627
```

```
# Govt
eda.table = table(df.isaward$DEGOVT, df.isaward$is.Awarded)
colnames(eda.table) = c("AWard", "NoAward")
rownames(eda.table) = c("NotGov", "Gov")
eda.table
```

```
##
##           AWard NoAward
##   NotGov    604    1131
##   Gov        57      44
```

```r
# Year
eda.table = table(df.isaward$YEAR, df.isaward$is.Awarded)
colnames(eda.table) = c("AWard", "NoAward")
eda.table
```

```
##
##     AWard NoAward
##   1    44      44
##   2    53      42
##   3    95     102
##   4   190     350
##   5   248     558
##   6    31      79
```

```r
# Claim Type
eda.table = table(df.isaward$CLAIMTYPE, df.isaward$is.Awarded)
colnames(eda.table) = c("AWard", "NoAward")
eda.table
```

```
##
##     AWard NoAward
##   1   167     267
##   2    82      76
##   3    71      36
##   4    47      81
##   5    25      67
##   6   269     648
```

```r
# Plaintiff
eda.table = table(df.isaward$TOTALNOPL, df.isaward$is.Awarded)
colnames(eda.table) = c("AWard", "NoAward")
eda.table
```

```
##
##     AWard NoAward
##   1   515     926
##   2   122     181
##   3    24      68
```

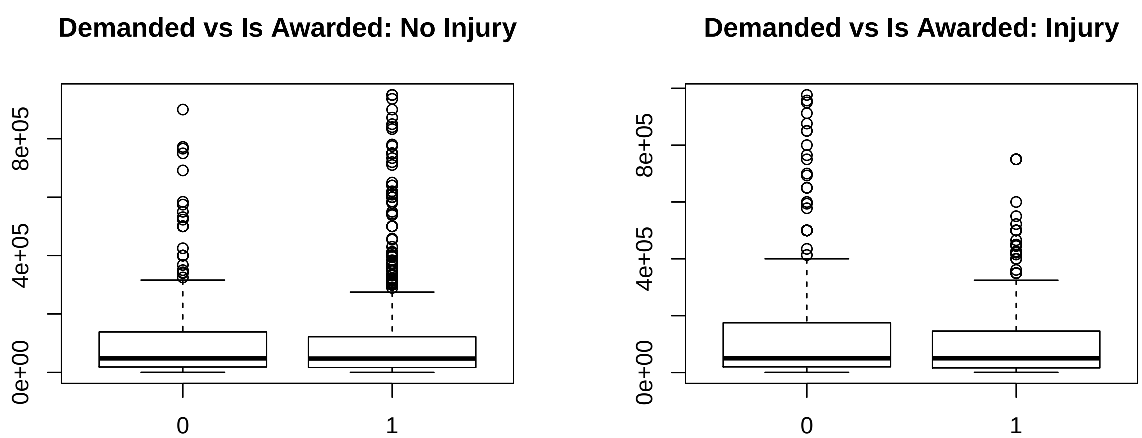
```r
# Defendant
eda.table = table(df.isaward$TOTALNODE, df.isaward$is.Awarded)
colnames(eda.table) = c("AWard", "NoAward")
eda.table
```

```
##
##     AWard NoAward
##   1   378     642
##   2   179     363
##   3   104     170
```

```
# Dispaly the precentage of sum of row So the precentage in
# each category that is awarded vs not awarded
# round(eda.table / rowSums(eda.table), 2)
```

Now we examine the discrete variables and their relationship with the response variable. In terms of whether the case involves a bodily injury, only 29% of the non-injured plaintiffs were awarded damage whereas 46% of the injured plaintiffs were awarded. This difference is unlikely caused by the sample size as the awarded portion is roughly equal and large. When the defendant was a corporation, 32% of the cases were awarded with damages, but this number is 40% in non-corporation defendants. In terms of year, it appears that in the recent years, the portion of cases that were awarded with damage shrieked. 1999-2001 had similar rates while prior to 1999 the rates were higher. This suggests that we could possibly regroup Year during modeling. In terms of the relationship with number of plaintiffs, our result suggests that portion of cases that received damages was the smallest when there's three or more plaintiffs.

```
# Interaction between bodily injury, total damages, and
# amount demanded is.award vs amount demanded by bodily
# injury
par(mfrow = c(1, 2))
boxplot(df.isaward.sub$DEMANDED ~ df.isaward.sub$is.Awarded,
    subset = df.isaward.sub$BODINJ == 0, main = "Demanded vs Is Awarded: No Injury")

boxplot(df.isaward.sub$DEMANDED ~ df.isaward.sub$is.Awarded,
    subset = df.isaward.sub$BODINJ == 1, main = "Demanded vs Is Awarded: Injury")
```

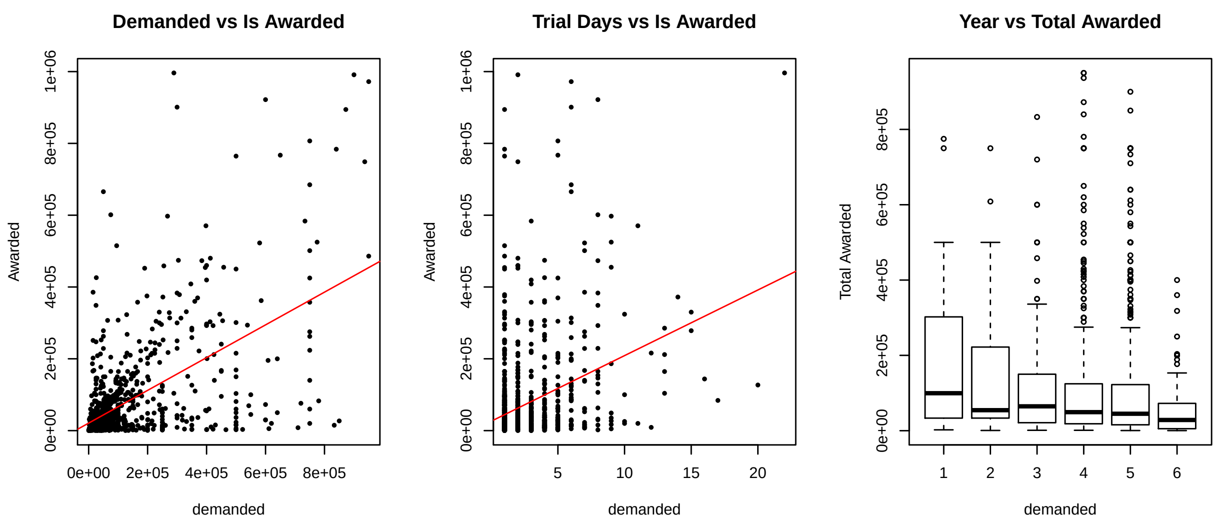


From the plots of demand vs awarded by whether the case involves bodily injury, we do not see an interaction between injury and demand with regard to whether the case was awarded with damage. Both boxes appeared to be similar.

Award Size Dataset Mutivariate

```
df.awardsize.sub <- df.awardsize[which(df.awardsize$TOTDAM <
    1e+06 & df.awardsize$DEMANDED < 1e+06), ]

## Continuous Demanded vs Awarded
```

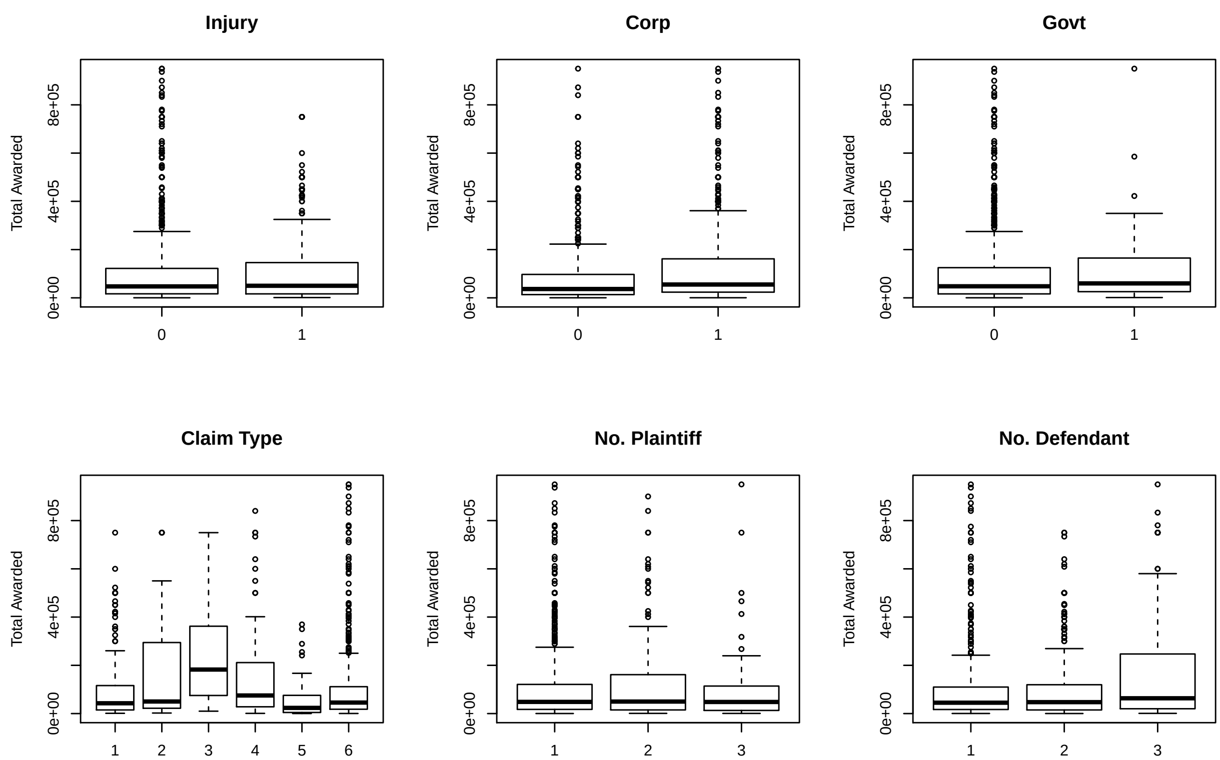
```r
par(mfrow = c(1, 3))
plot(df.awardsize.sub$DEMANDED, df.awardsize.sub$TOTDAM, main = "Demanded vs Is Awarded",
    pch = 16, cex = 0.7, xlab = "demanded", ylab = "Awarded")
abline(lm(TOTDAM ~ DEMANDED, data = df.awardsize.sub), col = 2)
# Trial Days vs Awarded
plot(df.awardsize.sub$TRIDAYS, df.awardsize.sub$TOTDAM, main = "Trial Days vs Is Awarded",
    pch = 16, cex = 0.7, xlab = "demanded", ylab = "Awarded")
abline(lm(TOTDAM ~ TRIDAYS, data = df.awardsize.sub), col = 2)
# Year
boxplot(df.awardsize.sub$DEMANDED ~ df.awardsize.sub$YEAR, main = "Year vs Total Awarded",
    ylab = "Total Awarded", xlab = "demanded", ylab = "Awarded")
```



In the dataset where we analyze the relationship between the covariates and the amount of damage that the plaintiffs were awarded (that is not zero). We can easily see positive relationship between money demanded and money awarded, and the same applies to the trial days' vs awarded case. In terms of year, we see a similar distribution when comparing to the binary dataset. It appears that the years leading up to 1999 shared more common ground than the cases before that, which were often award the plaintiff with larger amounts.

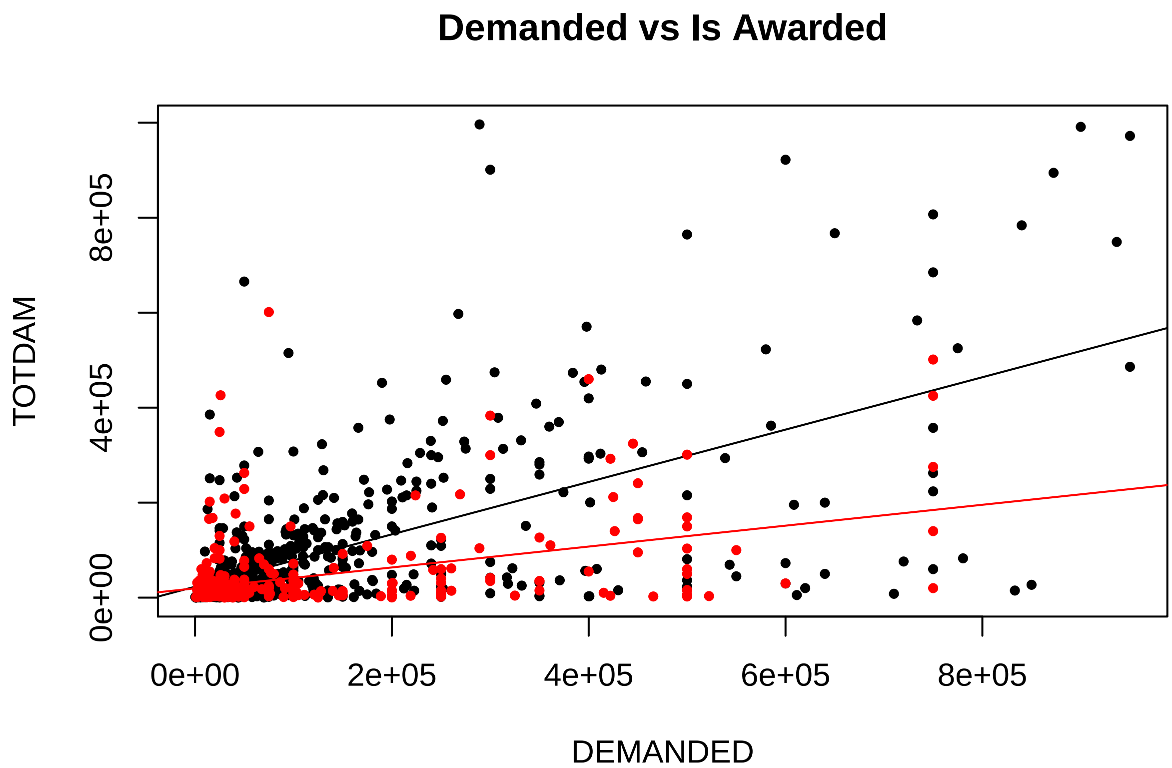
```r
## Discrete
par(mfrow = c(2, 3))
# Injury
boxplot(df.awardsize.sub$DEMANDED ~ df.awardsize.sub$BODINJ,
    main = "Injury", ylab = "Total Awarded")
# Corp
boxplot(df.awardsize.sub$DEMANDED ~ df.awardsize.sub$DECORP,
    main = "Corp", ylab = "Total Awarded")

# Govt
boxplot(df.awardsize.sub$DEMANDED ~ df.awardsize.sub$DEGOVT,
    main = "Govt", ylab = "Total Awarded")

# Claim Type
boxplot(df.awardsize.sub$DEMANDED ~ df.awardsize.sub$CLAIMTYPE,
    main = "Claim Type", ylab = "Total Awarded")
```

```r
# Plaintiffs
boxplot(df.awardsize.sub$DEMANDED ~ df.awardsize.sub$TOTALNOPL,
    main = "No. Plaintiff", ylab = "Total Awarded")

# Defendant
boxplot(df.awardsize.sub$DEMANDED ~ df.awardsize.sub$TOTALNODE,
    main = "No. Defendant", ylab = "Total Awarded")
```



The plot suggested that if the defendant is a corporation, it pays more in damages on average than non-corporation defendants. In terms of claims type, the claims involving malpractice and premises liability are the more likely to receive a higher reward. The number of defendants appears to be positively associated with the total amount of damages awarded, and the number of plaintiff appears to be the highest amongst all other categories, but the difference is not very significant.

```r
# interaction
par(mfrow = c(1, 1))
with(df.awardsize.sub[df.awardsize.sub$BODINJ == 0, ], plot(DEMANDED,
    TOTDAM, main = "Demanded vs Is Awarded", pch = 16, cex = 0.7))
with(df.awardsize.sub[df.awardsize.sub$BODINJ == 0, ], abline(lm(TOTDAM ~
    DEMANDED), col = 1))
with(df.awardsize.sub[df.awardsize.sub$BODINJ == 1, ], points(DEMANDED,
    TOTDAM, pch = 16, cex = 0.7, col = 2), abline(lm(TOTDAM ~
    DEMANDED), col = 2))
with(df.awardsize.sub[df.awardsize.sub$BODINJ == 1, ], abline(lm(TOTDAM ~
    DEMANDED), col = 2))
```



The red dots in the scatter plot above shows the plot of total amount awarded vs the total amount demanded, colored by the value of whether there was a bodily injury involved in the case (if so, colored in red). Then we fit two regression models in our plot, and it appears that there is an interaction between demanded, effects and the total damage since their temporary regression lines intercepts.

Initial Modeling

Binary Award Logistic Model

```
glm.1 <- glm(is.Awarded ~ TRIDAYS + factor(DECORP) + factor(DEGOVT) +
    factor(YEAR) + factor(CLAIMTYPE) + factor(TOTALNOPL) + factor(TOTALNODE) +
    factor(BODINJ) * DEMANDED, data = df.isaward, family = "binomial")
pander(summary(glm.1))
```

	Estimate	Std. Error	z value	Pr(>|z|)
TRIDAYS	0.03852	0.01758	2.19	0.02849
factor(DECORP)1	0.0871	0.1177	0.7398	0.4594
factor(DEGOVT)1	-0.5826	0.2378	-2.45	0.01429
factor(YEAR)2	-0.4649	0.3203	-1.451	0.1466
factor(YEAR)3	-0.1628	0.2794	-0.5826	0.5602
factor(YEAR)4	0.2548	0.256	0.9953	0.3196
factor(YEAR)5	0.3966	0.2564	1.547	0.122
factor(YEAR)6	0.4442	0.3311	1.342	0.1797
factor(CLAIMTYPE)2	-0.4086	0.2071	-1.974	0.04844
factor(CLAIMTYPE)3	-1.299	0.2493	-5.209	1.896e-07
factor(CLAIMTYPE)4	-0.8677	0.3054	-2.841	0.004499
factor(CLAIMTYPE)5	-0.4831	0.3429	-1.409	0.1589
factor(CLAIMTYPE)6	-0.4361	0.2322	-1.878	0.06039
factor(TOTALNOPL)2	-0.03103	0.139	-0.2232	0.8234
factor(TOTALNOPL)3	0.6585	0.2621	2.512	0.01199
factor(TOTALNODE)2	0.2583	0.1221	2.115	0.03442
factor(TOTALNODE)3	0.07781	0.1633	0.4764	0.6338
factor(BODINJ)1	-0.9488	0.222	-4.274	1.916e-05
DEMANDED	-3.913e-08	2.32e-08	-1.687	0.09165
factor(BODINJ)1:DEMANDED	3.352e-08	2.593e-08	1.293	0.1962
(Intercept)	0.9533	0.3574	2.667	0.007642

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	2399 on 1835 degrees of freedom
Residual deviance:	2248 on 1815 degrees of freedom

Our initial logistic regression model for predicitng whether a case will be awarded with damages is modeled as:

$logit(is.Awarded) = \beta_0 + \beta_1 * TRIDAYS + \beta_2 * DECORP + \beta_3 * DEGOVT + \beta_4 * YEAR + \beta_5 * CLAIMTYPE + \beta_6 * TOTALNOPL + \beta_7 * TOTALNODE + \beta_8 * BODINJ + \beta_9 * DEMANDED + \beta_{10} * DEMANDED{:}BODINJ$

The summary of our model suggest that parameters: Trial days, whether the defendant is the government, claim type premises liability, malpractice, fraud, and others, if there are at least three defendants, if there are two plaintiffs, whether the case involves injury, and damage demanded by the plaintiff are statistically significant. However, we are faced with the issue that the year category being not significant, and having several categories (recent years) having similar coefficients. This points towards what we have seen earlier in the EDA, and we may categorize the year in our modeling to improve fitness. Also, we know that the

DEMAND variable is not normally distributed, and it contains large values that are potential leverage points. Therefore, we will see if transformation on that model improves the goodness of fit.

From the parameters of the initial logistic model, we can see that neither of the variable of interest is statistically significant. However, we cannot make inference from this model without doing diagnostics to check model assumption and fitness.

```
df.isaward$YEAR.recat = ifelse(df.isaward$YEAR >= 4, 1, 0)
glm.1.a <- glm(is.Awarded ~ TRIDAYS + factor(DECORP) + factor(DEGOVT) +
    factor(YEAR.recat) + factor(CLAIMTYPE) + factor(TOTALNOPL) +
    factor(TOTALNODE) + factor(BODINJ) * DEMANDED, data = df.isaward,
    family = "binomial")
pander(summary(glm.1.a))
```

	Estimate	Std. Error	z value	Pr(>|z|)
TRIDAYS	0.03467	0.01737	1.997	0.04585
factor(DECORP)1	0.08316	0.1175	0.7075	0.4793
factor(DEGOVT)1	-0.5641	0.2336	-2.415	0.01575
factor(YEAR.recat)1	0.5417	0.1284	4.219	2.452e-05
factor(CLAIMTYPE)2	-0.4113	0.2062	-1.994	0.04613
factor(CLAIMTYPE)3	-1.303	0.2485	-5.243	1.577e-07
factor(CLAIMTYPE)4	-0.8594	0.304	-2.827	0.004697
factor(CLAIMTYPE)5	-0.4512	0.3339	-1.351	0.1767
factor(CLAIMTYPE)6	-0.4243	0.2305	-1.84	0.06574
factor(TOTALNOPL)2	-0.03461	0.1387	-0.2496	0.8029
factor(TOTALNOPL)3	0.6343	0.261	2.43	0.01509
factor(TOTALNODE)2	0.2613	0.1217	2.148	0.03175
factor(TOTALNODE)3	0.07657	0.1629	0.4702	0.6382
factor(BODINJ)1	-0.9493	0.2206	-4.304	1.677e-05
DEMANDED	-3.76e-08	2.305e-08	-1.631	0.1029
factor(BODINJ)1:DEMANDED	3.405e-08	2.572e-08	1.324	0.1855
(Intercept)	0.7616	0.2697	2.824	0.00474

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	2399 on 1835 degrees of freedom
Residual deviance:	2251 on 1819 degrees of freedom

```
# df.isaward$YEAR.recat = ifelse(df.isaward$YEAR >= 5, 1, 0)
# glm.1.b <- glm(is.Awarded ~ TRIDAYS + factor(DECORP) +
# factor(DEGOVT) + factor(YEAR.recat) + factor(CLAIMTYPE) +
# factor(TOTALNOPL) + factor(TOTALNODE) +
# factor(BODINJ)*DEMANDED, data = df.isaward, family =
# 'binomial') pander(summary(glm.1.b))
```

We attempted to re-categorize year into several ways, but when we put year pre-1997 - 1998 into one category, and year 1999-2001 into the other, it improves our model fitness (decreases AIC) and the new parameter is statistically significant. Therefore, we will continue with the re-categorizing of year in the logistic model.

Award Size Linear Model

```r
lm.1 <- lm(TOTDAM ~ TRIDAYS + factor(DECORP) + factor(DEGOVT) +
    factor(YEAR) + factor(CLAIMTYPE) + factor(TOTALNOPL) + factor(TOTALNODE) +
    factor(BODINJ) * DEMANDED, data = df.awardsize)
pander(summary(lm.1))
```

	Estimate	Std. Error	t value	Pr(>|t|)
TRIDAYS	168156	16282	10.33	5.614e-24
factor(DECORP)1	-97560	112377	-0.8681	0.3855
factor(DEGOVT)1	-438089	279062	-1.57	0.1167
factor(YEAR)2	-839726	367372	-2.286	0.02245
factor(YEAR)3	-1109238	311014	-3.567	0.0003766
factor(YEAR)4	-879144	278754	-3.154	0.001653
factor(YEAR)5	-867488	277333	-3.128	0.001804
factor(YEAR)6	-699035	332819	-2.1	0.03592
factor(CLAIMTYPE)2	-194320	236386	-0.822	0.4112
factor(CLAIMTYPE)3	-127569	316861	-0.4026	0.6873
factor(CLAIMTYPE)4	-545712	321390	-1.698	0.08978
factor(CLAIMTYPE)5	-487945	337326	-1.447	0.1483
factor(CLAIMTYPE)6	-476858	252571	-1.888	0.05927
factor(TOTALNOPL)2	-139675	143156	-0.9757	0.3294
factor(TOTALNOPL)3	84341	218113	0.3867	0.6991
factor(TOTALNODE)2	41286	115955	0.3561	0.7219
factor(TOTALNODE)3	49900	161105	0.3097	0.7568
factor(BODINJ)1	-303820	240647	-1.263	0.207
DEMANDED	0.6261	0.02322	26.96	1.372e-124
factor(BODINJ)1:DEMANDED	-0.5472	0.02935	-18.64	5.04e-68
(Intercept)	979864	382161	2.564	0.01047

Table 6: Fitting linear model: TOTDAM ~ TRIDAYS + factor(DECORP) + factor(DEGOVT) + factor(YEAR) + factor(CLAIMTYPE) + factor(TOTALNOPL) + factor(TOTALNODE) + factor(BODINJ) * DEMANDED

Observations	Residual Std. Error	R^2	Adjusted R^2
1175	1688838	0.4647	0.4555

```r
# summary(lm.1) mmps(lm.1)
```

Our initial linear regression model for predicting the amount that the plaintiff would be awarded in damages is modeled as:

$TOTDAM = \beta_0 + \beta_1 * TRIDAYS + \beta_2 * DECORP + \beta_3 * DEGOVT + \beta_4 * YEAR + \beta_5 * CLAIMTYPE + \beta_6 * TOTALNOPL + \beta_7 * TOTALNODE + \beta_8 * BODINJ + \beta_9 * DEMANDED + \beta_{10} * DEMANDED{:}BODINJ$

From our model summary, we can see that the trial days, all the factors in year, claim type fraud and other, and the interaction term between whether the case involves injury and the amount demanded are significant parameters. Nonetheless, while this model provides us with some insights on the relationship between the predictor variables and co-variates, we cannot make any formal inference without diagnostics. We should take note that the DEMANDED and TOTDAM contain influential possible leverage points, and transformation on these variables could improve our model fitness.

```
df.awardsize$YEAR.recat = ifelse(df.awardsize$YEAR >= 4, 1, 0)
lm.1.a <- lm(TOTDAM ~ TRIDAYS + factor(DECORP) + factor(DEGOVT) +
    factor(YEAR) + factor(CLAIMTYPE) + factor(TOTALNOPL) + factor(TOTALNODE) +
    factor(BODINJ) * DEMANDED, data = df.awardsize)
# pander(summary(lm.1.a))
```

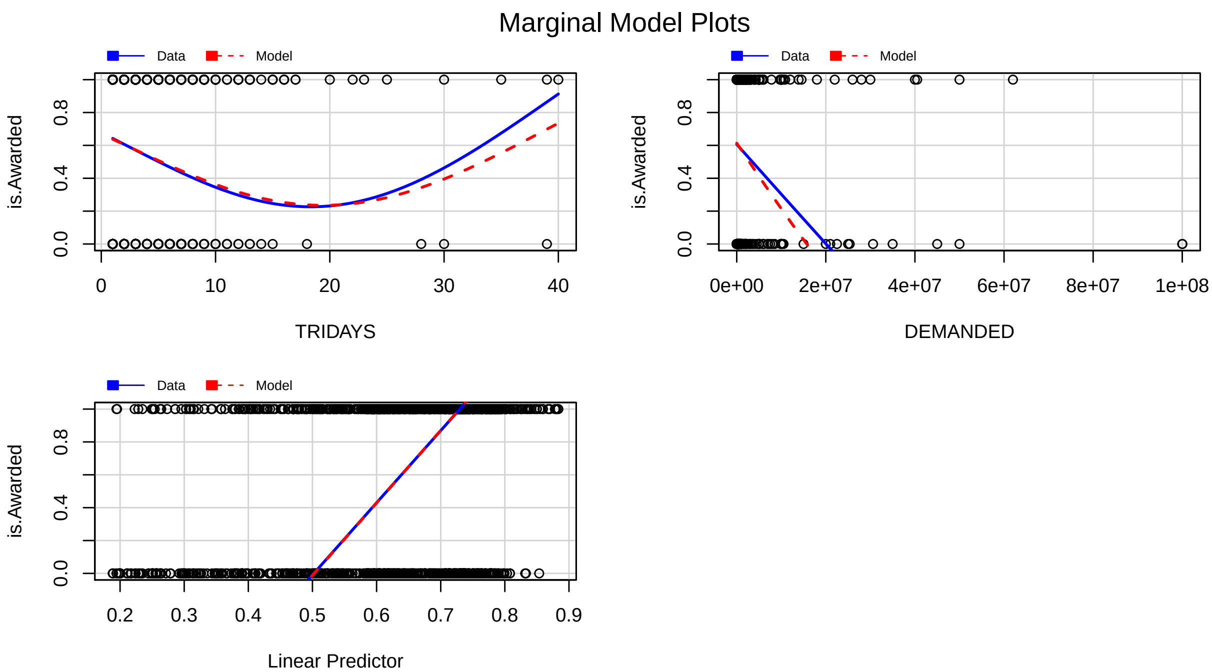
When re-categorizing year in this model, we fail to see improvment regardless of the cutoff of year, therefore we will continue to use year in our model as the factor of the origional categories.

Model Diagnostics and Transformation

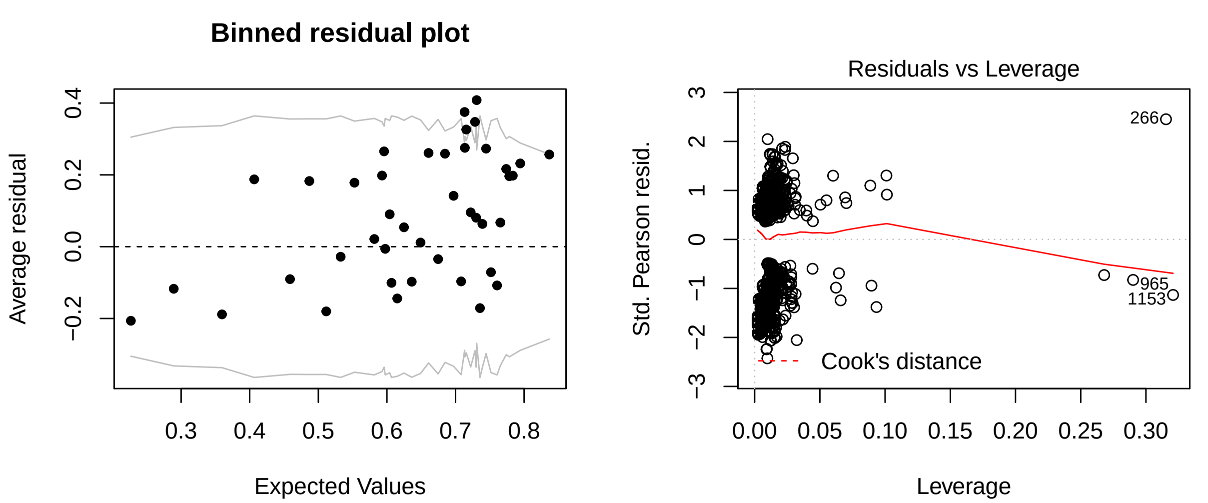
Binary Award Logistic Model

```r
mmps(glm.1.a)
```

```
## Warning in mmps(glm.1.a): Interactions and/or factors skipped
```



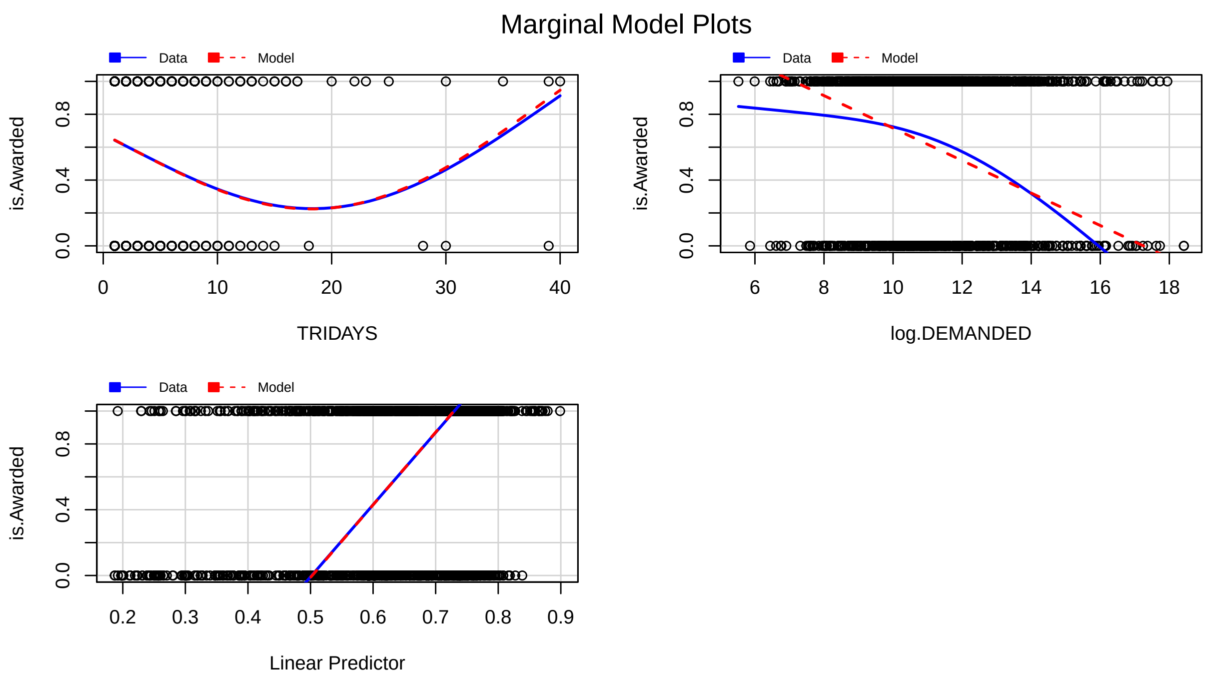

```r
par(mfrow = c(1, 2))
binnedplot(fitted(glm.1.a), resid(glm.1.a))
plot(glm.1.a, which = 5)
```



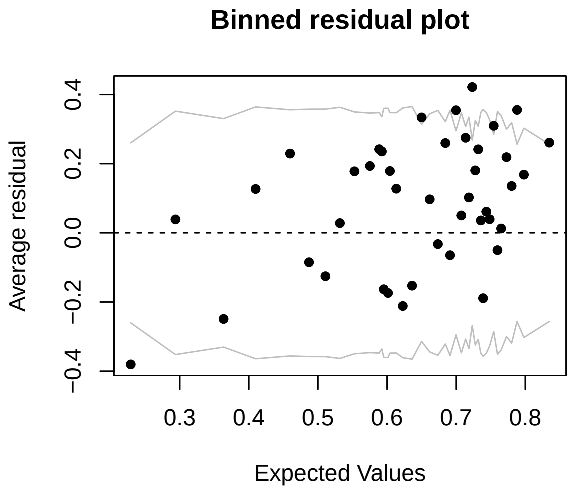
Our marginal model plot shows general good fit of our logistic model, but our residual vs leverage plot points out that there are several potential outliers: Case 266, 965, and 1153. All three cases demanded more than 50 million dollars in compensation, but we do not have enough information to rule these points as outliers. Since the are very far apart from the rest of the data, we will see if log transformation on this variable will improve the goodness of fit.

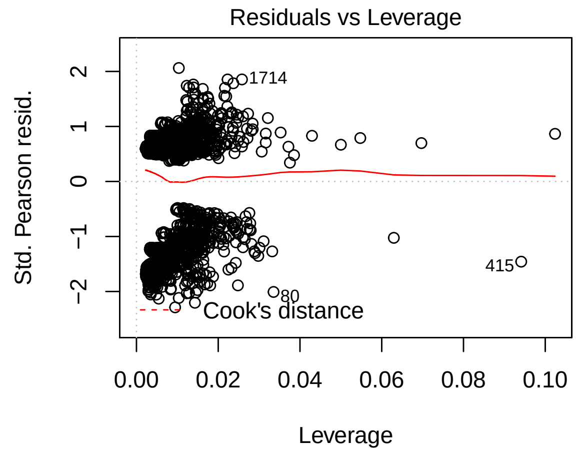
```
df.isaward$log.DEMANDED = log(df.isaward$DEMANDED)
glm.2 <- glm(is.Awarded ~ TRIDAYS + factor(DECORP) + factor(DEGOVT) +
    factor(YEAR.recat) + factor(CLAIMTYPE) + factor(TOTALNOPL) +
    factor(TOTALNODE) + factor(BODINJ) * log.DEMANDED, data = df.isaward,
    family = "binomial")
# pander(summary(glm.2))

mmps(glm.2)
```

```
## Warning in mmps(glm.2): Interactions and/or factors skipped
```




```
par(mfrow = c(1, 2))
binnedplot(fitted(glm.2), resid(glm.2))
plot(glm.2, which = 5)
```

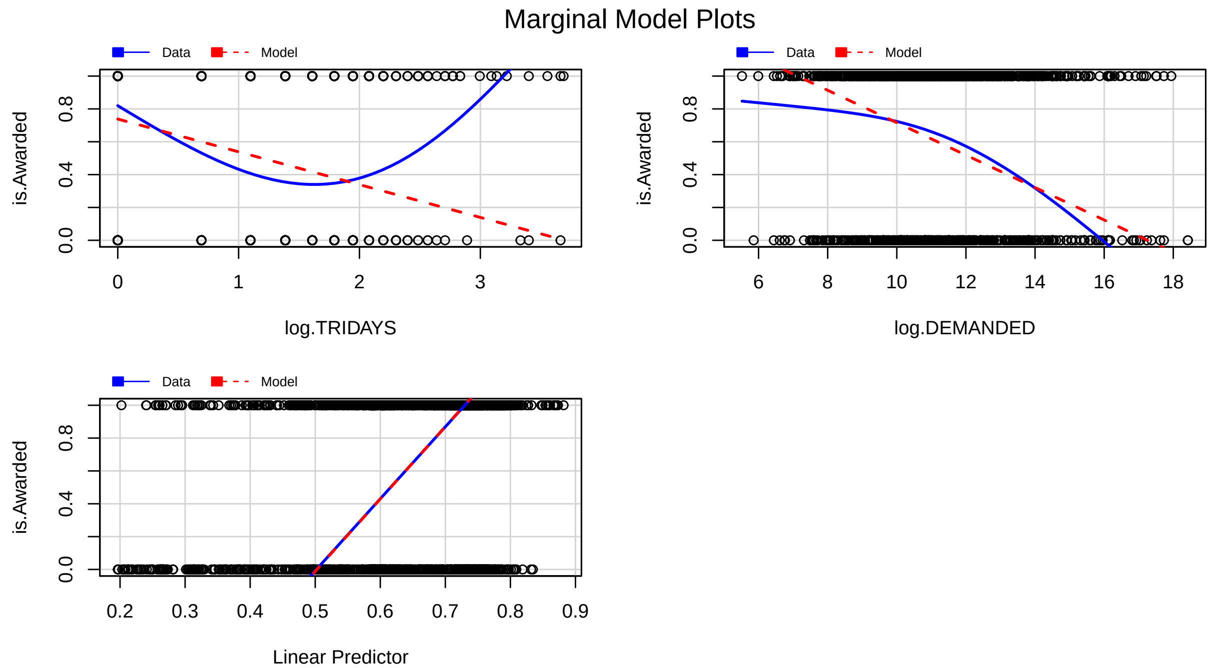




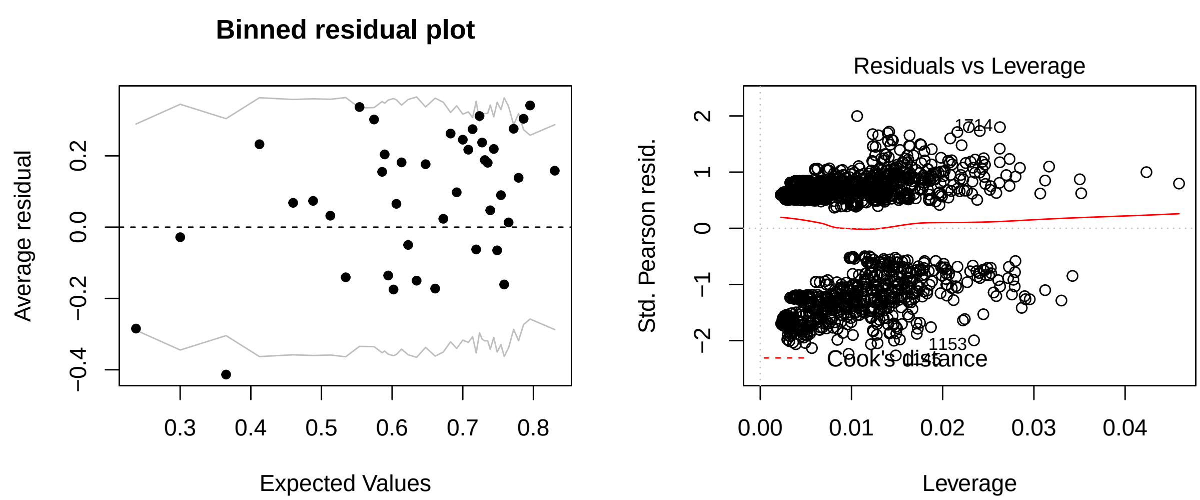
Our log transformation has failed to significantly improve our model fitness. The AIC value increased while the binned residual plot appears to show little improvement. The previous leverage points are no longer an issue, but the transformation created new leverage points with high standardized residual. Therefore, we will not move forward with this transformation.

```
df.isaward$log.TRIDAYS = log(df.isaward$TRIDAYS)
glm.2 <- glm(is.Awarded ~ log.TRIDAYS + factor(DECORP) + factor(DEGOVT) +
    factor(YEAR.recat) + factor(CLAIMTYPE) + factor(TOTALNOPL) +
    factor(TOTALNODE) + factor(BODINJ) * log.DEMANDED, data = df.isaward,
    family = "binomial")
# pander(summary(glm.2))

mmps(glm.2)
```

```
## Warning in mmps(glm.2): Interactions and/or factors skipped
```




```r
par(mfrow = c(1, 2))
binnedplot(fitted(glm.2), resid(glm.2))
plot(glm.2, which = 5)
```

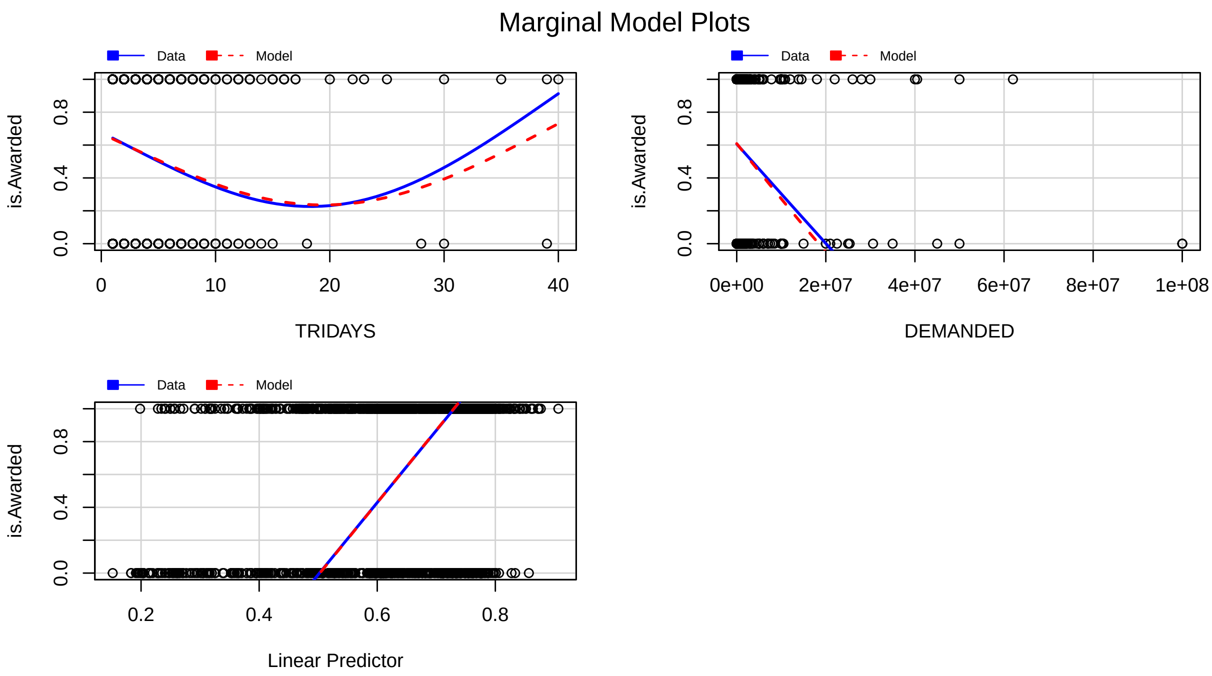


When transforming both TRIDAYS and DEMANDED with log, our binned residual plot showed slight improvement in terms of being contained within the confidence band, and our residual vs leverage point appears to have smaller leverage. However, our marginal models plot suggested that this model is a poor fit on the data, and our AIC also increased. Thus we will not go with this transformation.

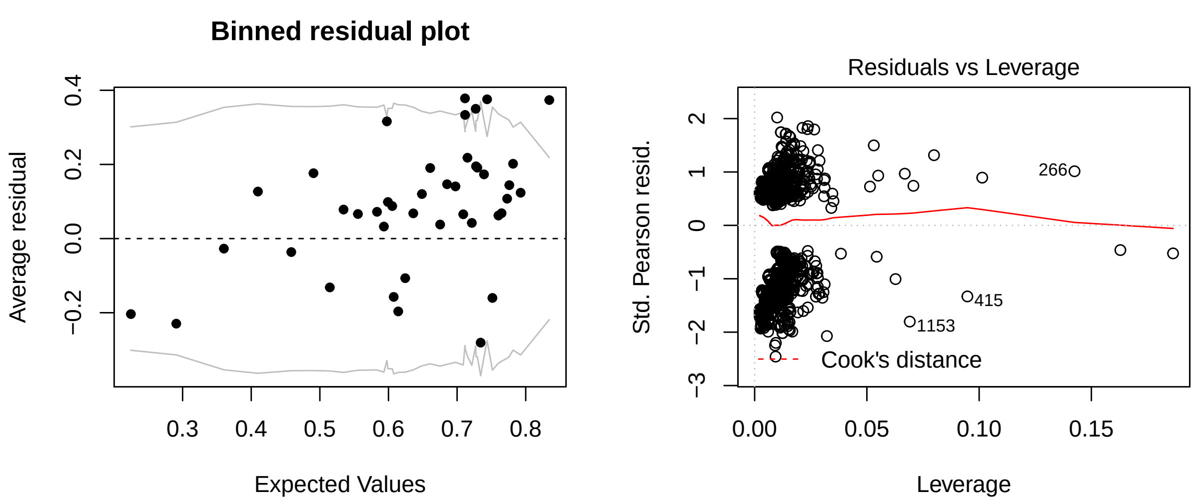
```r
glm.2 <- glm(is.Awarded ~ TRIDAYS + factor(DECORP) + factor(DEGOVT) +
    factor(YEAR.recat) + factor(CLAIMTYPE) + factor(TOTALNOPL) +
    factor(TOTALNODE) + factor(BODINJ) + DEMANDED, data = df.isaward,
    family = "binomial")
```

```r
# summary(glm.2) pander(summary(glm.2))

mmps(glm.2)
```

```
## Warning in mmps(glm.2): Interactions and/or factors skipped
```




```r
par(mfrow = c(1, 2))
binnedplot(fitted(glm.2), resid(glm.2))
plot(glm.2, which = 5)
```



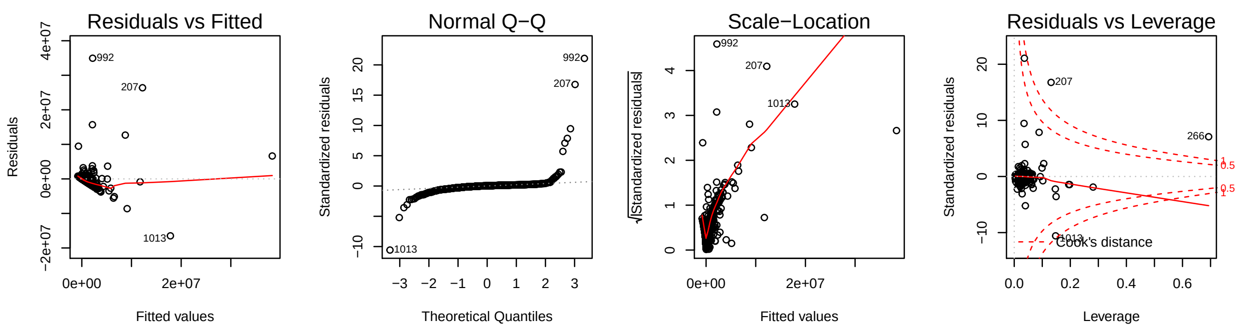
We then want to check if the interaction term between injury and damage demanded can be dropped to improve model fitness. Our AIC shows no improvement (marginal difference in value, not significant enough to claim model improvement). Our residual plots also did not show improvement from our previous model. Thus we will keep the interaction term in our model.

After multiple attempts, we decide to stick with our original model for predicting whether a case will be awarded with damage:

$logit(is.Awarded) = \beta_0 + \beta_1 * TRIDAYS + \beta_2 * DECORP + \beta_3 * DEGOVT + \beta_4 * YEAR.recat + \beta_5 * CLAIMTYPE + \beta_6 * TOTALNOPL + \beta_7 * TOTALNODE + \beta_8 * BODINJ + \beta_9 * DEMANDED + \beta_{10} * DEMANDED{:}BODINJ$

Award Size Linear Model

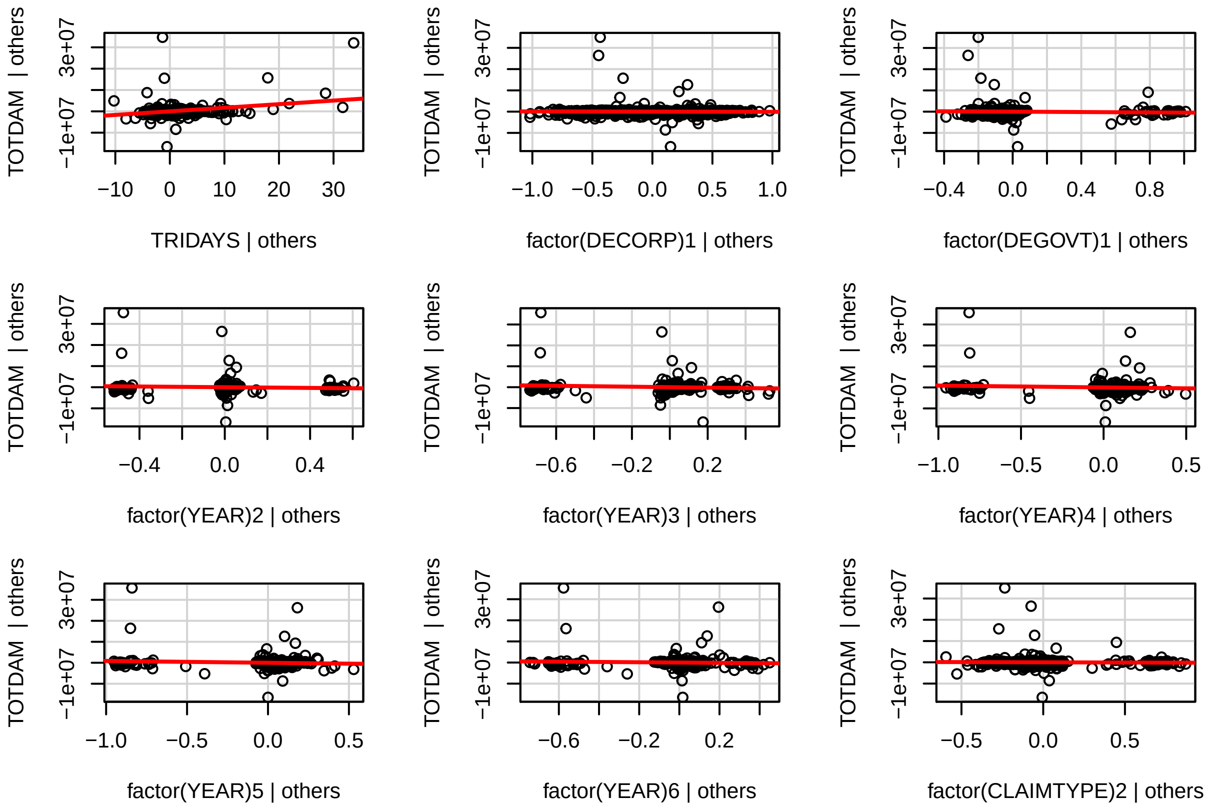
```
par(mfrow = c(1, 4))
plot(lm.1)
```

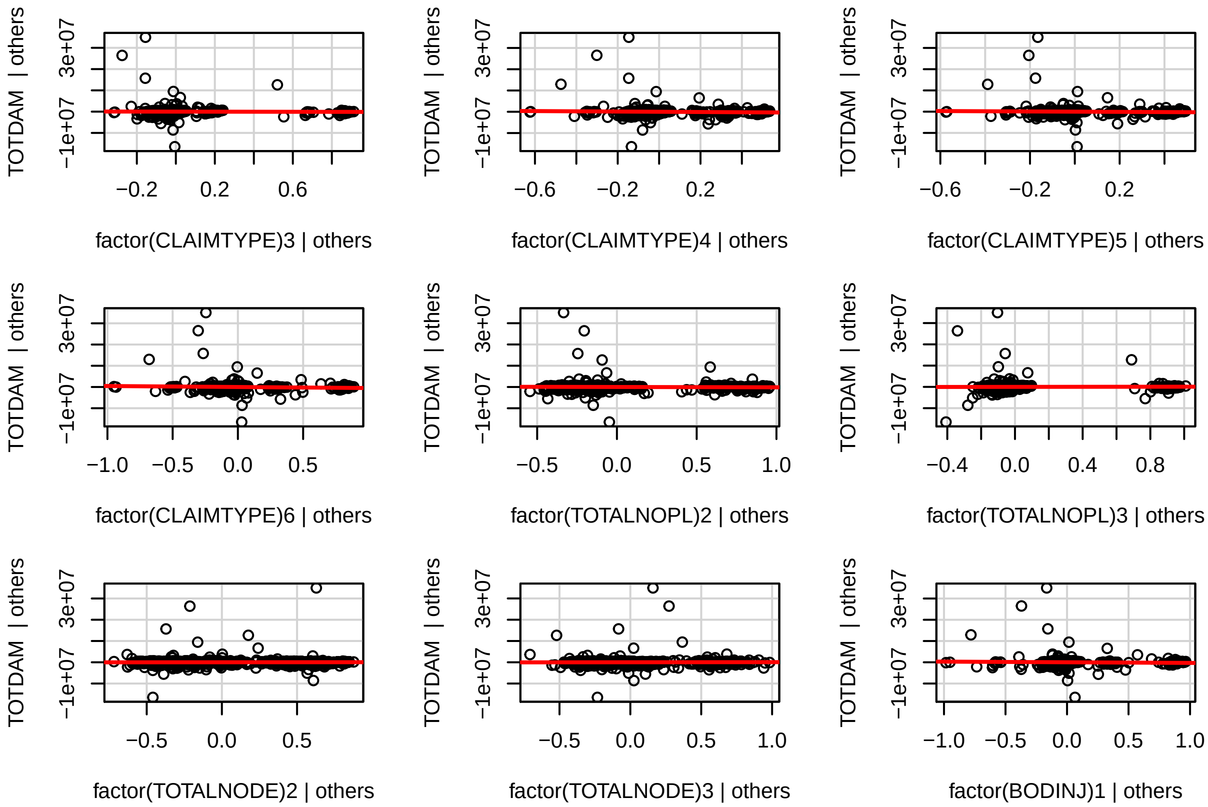


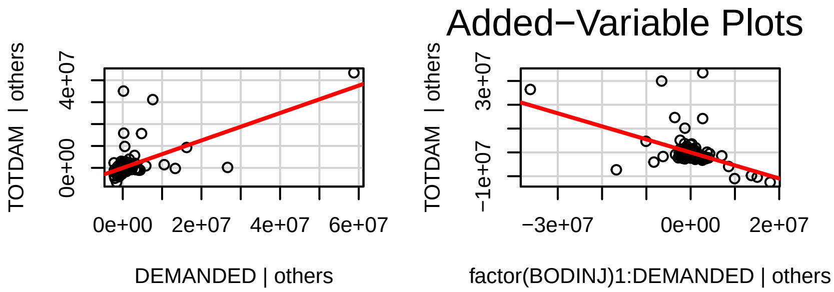
From our diagnostic plot for the linear model, we can see form the scale location plot that the constant variance assumption is violated here. We also do not see zero expectation or residual independence, as the residuals appears to be correlated in the residuals vs fitted plot. Our normality assumption is mostly met, but the plot suggested that there are several potential outliers. We can see that point 266, 207, 1013 are potential outliers with high leverage, but we do not have enough information to rule them as outliers.

```
av.plots(lm.1)
```

```
## Warning: 'av.plots' is deprecated.
## Use 'avPlots' instead.
## See help("Deprecated") and help("car-deprecated").
```

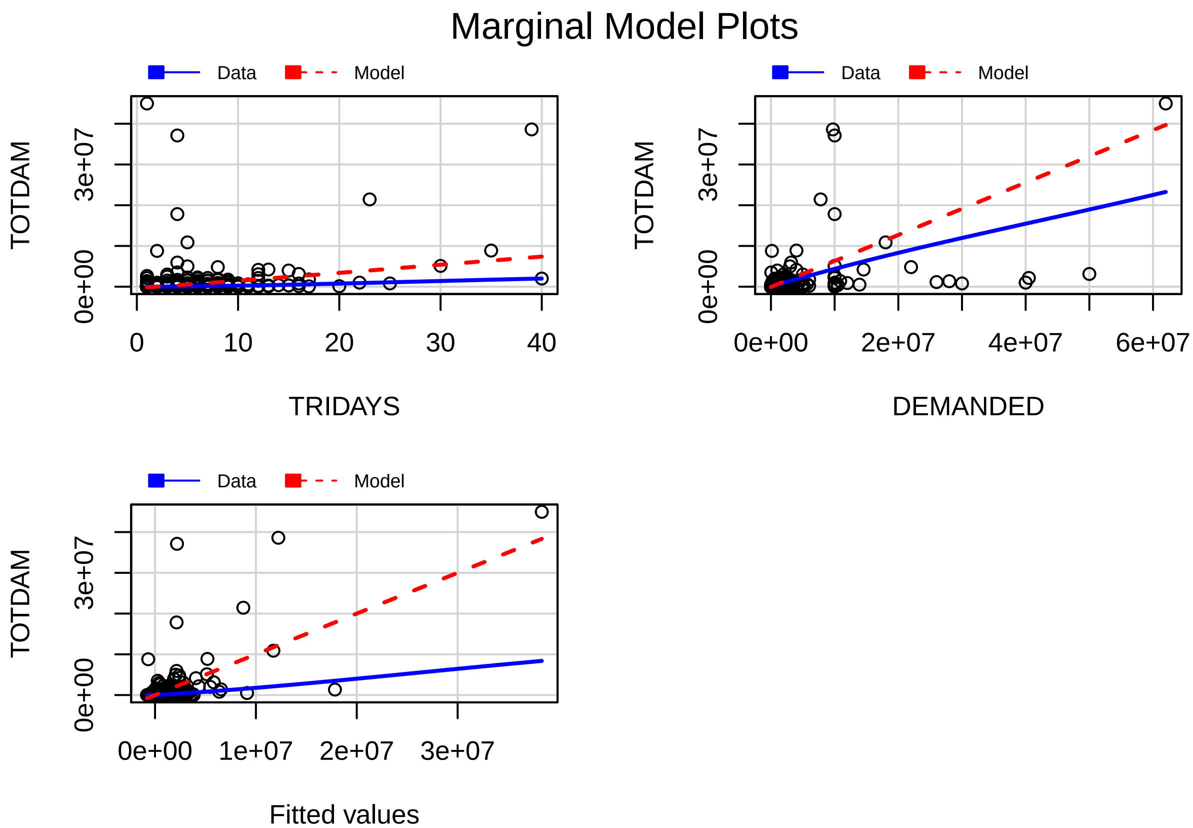

TOTDAM | others
−1e+07
3e+07
TRIDAYS | others
factor(DECORP)1 | others
factor(DEGOVT)1 | others
factor(YEAR)2 | others
factor(YEAR)3 | others
factor(YEAR)4 | others
factor(YEAR)5 | others
factor(YEAR)6 | others
factor(CLAIMTYPE)2 | others


TOTDAM | others
−1e+07
3e+07
−0.2
0.2
0.6
factor(CLAIMTYPE)3 | others
TOTDAM | others
−1e+07
3e+07
−0.6
−0.2
0.2
factor(CLAIMTYPE)4 | others
TOTDAM | others
−1e+07
3e+07
−0.6
−0.2
0.2
factor(CLAIMTYPE)5 | others
TOTDAM | others
−1e+07
3e+07
−1.0
−0.5
0.0
0.5
factor(CLAIMTYPE)6 | others
TOTDAM | others
−1e+07
3e+07
−0.5
0.0
0.5
1.0
factor(TOTALNOPL)2 | others
TOTDAM | others
−1e+07
3e+07
−0.4
0.0
0.4
0.8
factor(TOTALNOPL)3 | others
TOTDAM | others
−1e+07
3e+07
−0.5
0.0
0.5
factor(TOTALNODE)2 | others
TOTDAM | others
−1e+07
3e+07
−0.5
0.0
0.5
1.0
factor(TOTALNODE)3 | others
TOTDAM | others
−1e+07
3e+07
−1.0
−0.5
0.0
0.5
1.0
factor(BODINJ)1 | others



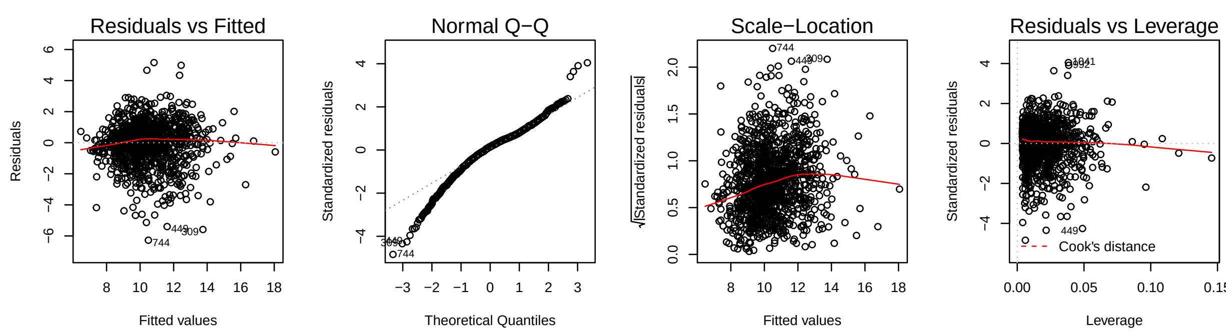

```r
mmps(lm.1)
```

```
## Warning in mmps(lm.1): Interactions and/or factors skipped
```



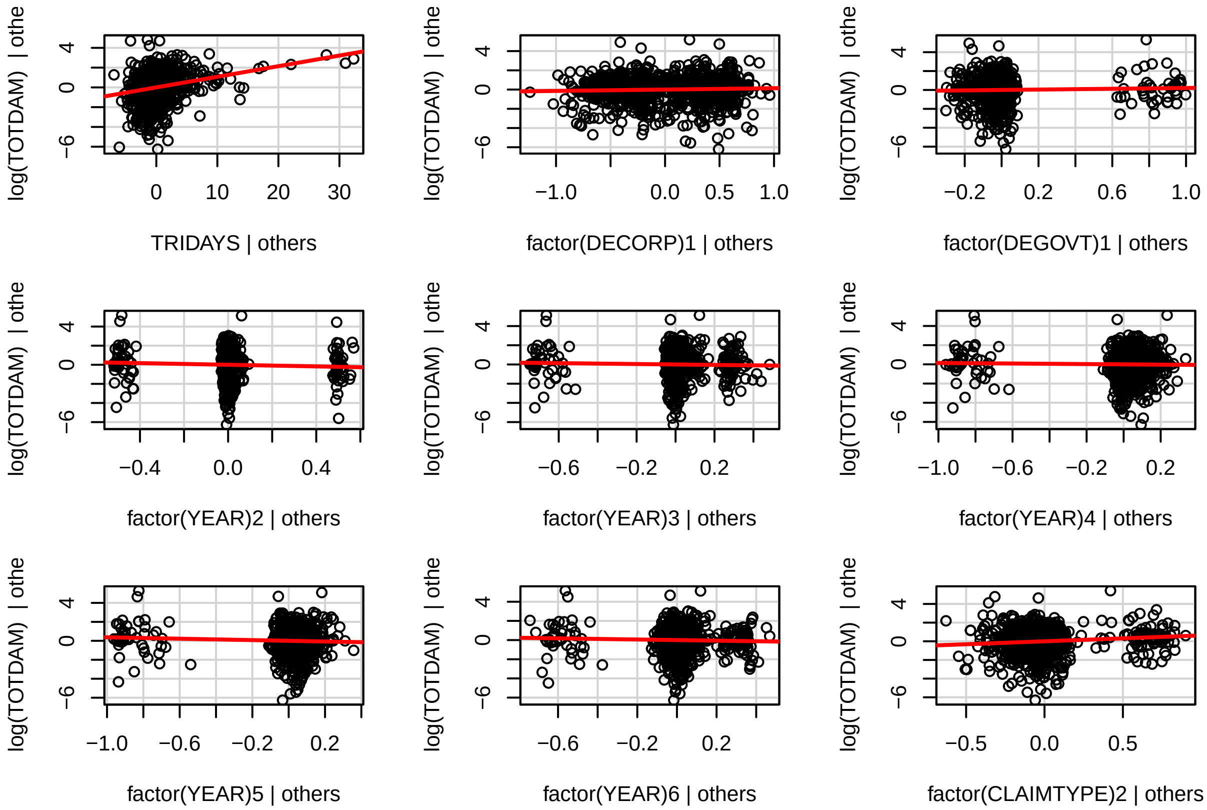
From the added variable plot we can see that the interaction term with DEMANDED appears to have very clustered residuals, and the marginal model plotting also suggests that our model is not an approiate fit. Therefore, we want to see if log transformation on the continuious variables would improve our model.

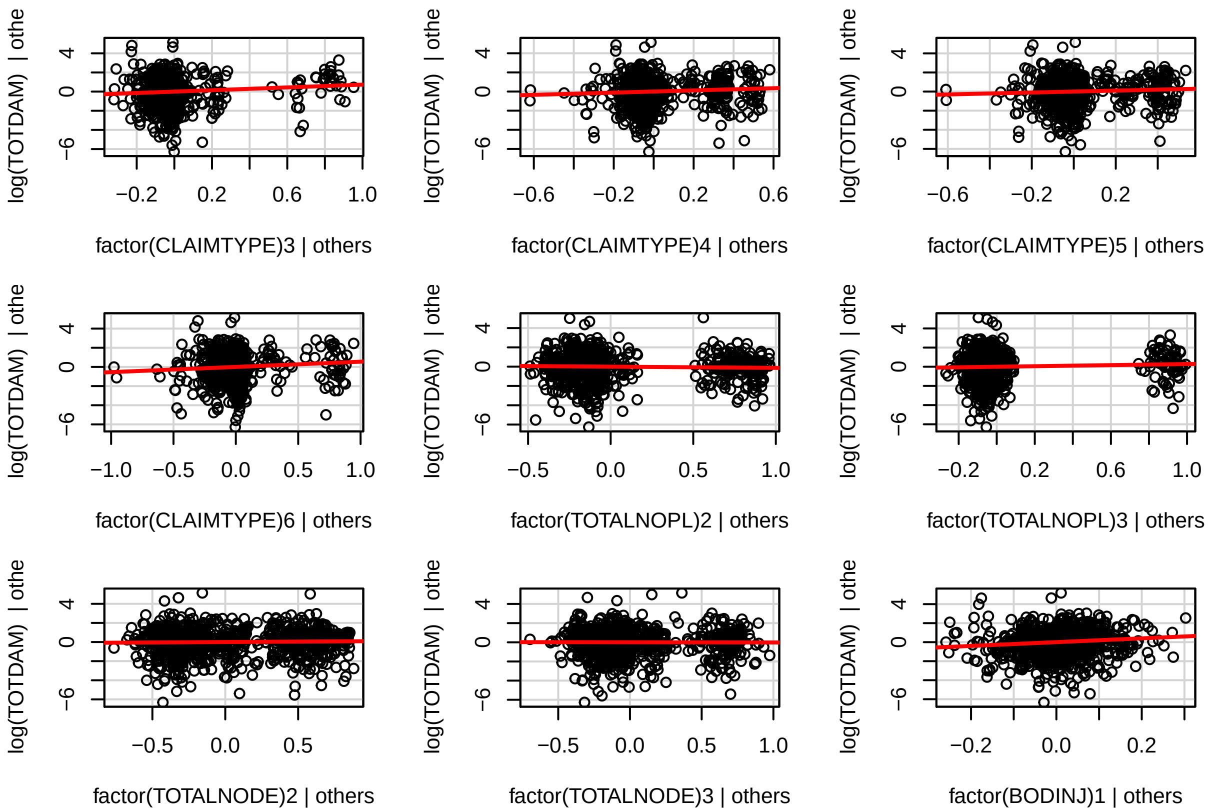
```
lm.2.1 <- lm(log(TOTDAM) ~ TRIDAYS + factor(DECORP) + factor(DEGOVT) +
    factor(YEAR) + factor(CLAIMTYPE) + factor(TOTALNOPL) + factor(TOTALNODE) +
    factor(BODINJ) * log(DEMANDED), data = df.awardsize)
# summary(lm.2)
par(mfrow = c(1, 4))
plot(lm.2.1)
```

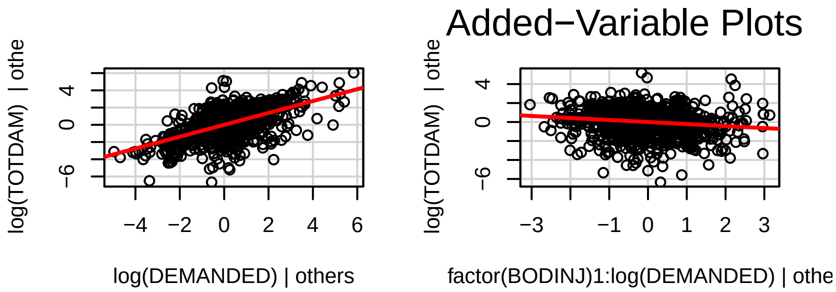



```
av.plots(lm.2.1)
```

```
## Warning: 'av.plots' is deprecated.
## Use 'avPlots' instead.
## See help("Deprecated") and help("car-deprecated").
```

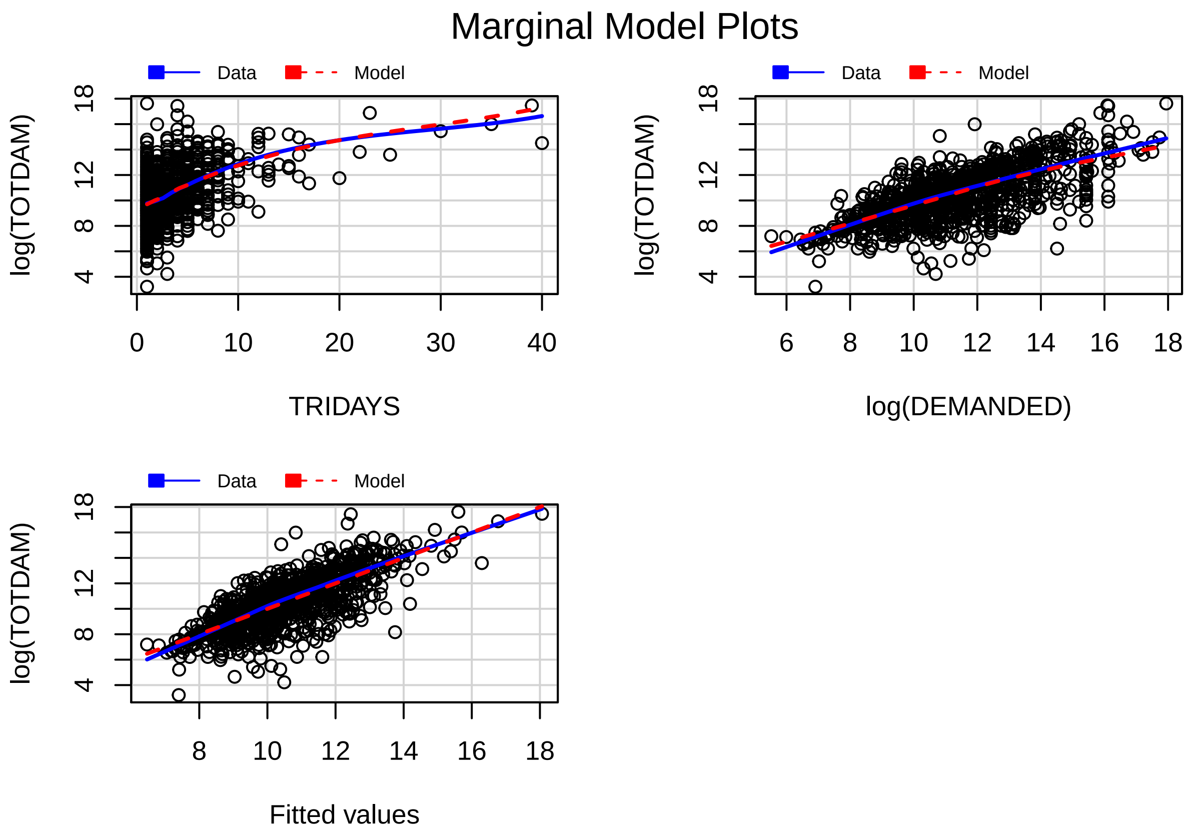



log(TOTDAM) | othe
factor(CLAIMTYPE)3 | others
factor(CLAIMTYPE)4 | others
factor(CLAIMTYPE)5 | others
factor(CLAIMTYPE)6 | others
factor(TOTALNOPL)2 | others
factor(TOTALNOPL)3 | others
factor(TOTALNODE)2 | others
factor(TOTALNODE)3 | others
factor(BODINJ)1 | others




```r
mmps(lm.2.1)
```

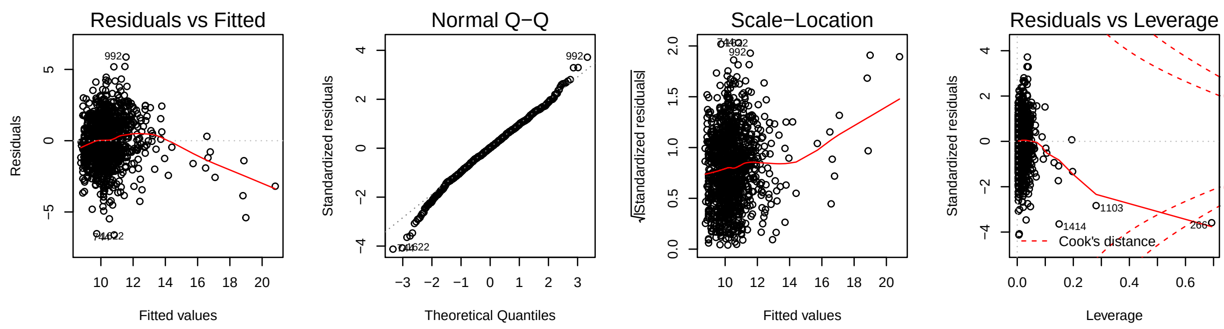
```
## Warning in mmps(lm.2.1): Interactions and/or factors skipped
```



After using both log transformation on TOTDAM and DEMANED, we see significant improvement in our residual plots. Our variance has stabilized and the residuals appears to be more evenly spread along both sides of the zero line. The residual also appears to be independent. However, the normal QQ plot has improved, but still contains a left tail with several potential outliers (744, 309, 499). Our residual vs Leverage plot has resolved the potential outlier with high leverage form the previous model. We now have some new potential outliers: 1041, 922, 449. However, with the limited information we cannot rule them out from the model.

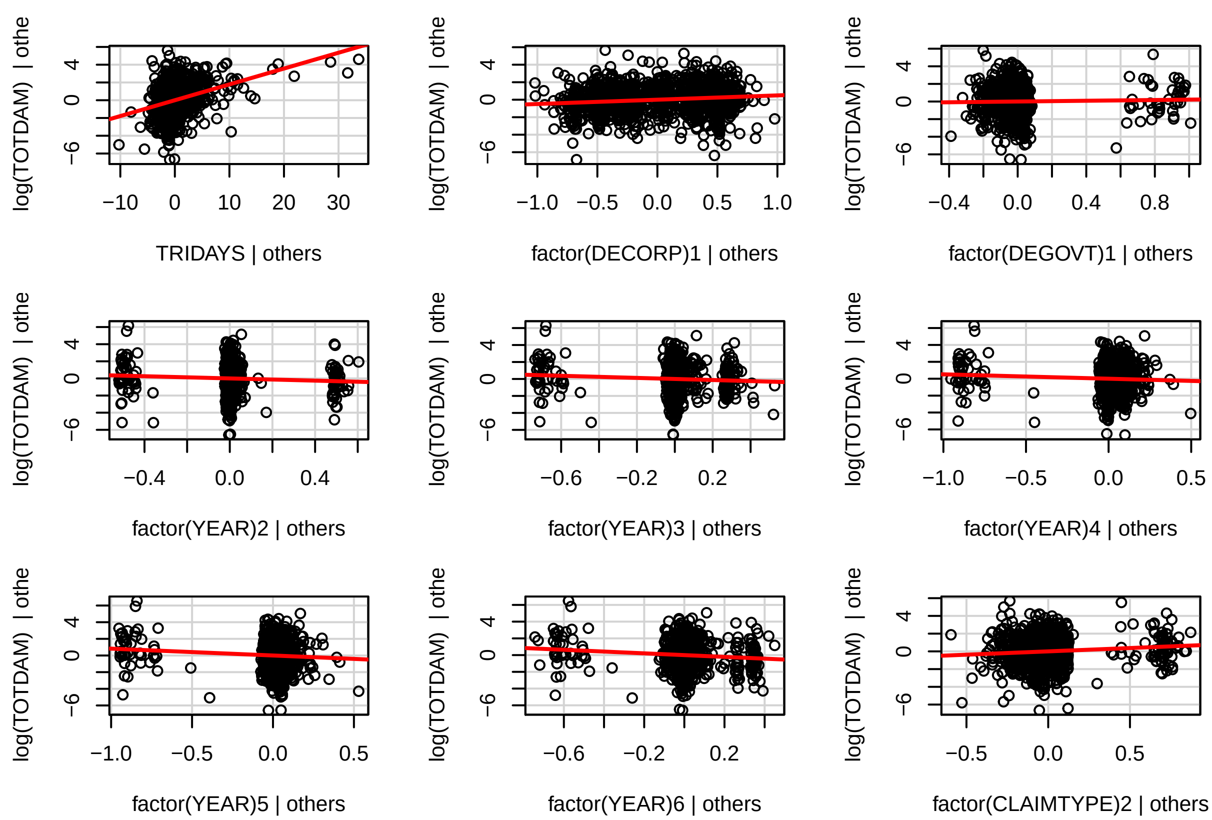
Our Added Variable plots and marginal model plots have also shown significnat improvement, but it suggests that a transformation on TRIDAYS can help improve our model fitness as well. Before doing that, we just want to experiment with only log transforming one variable.

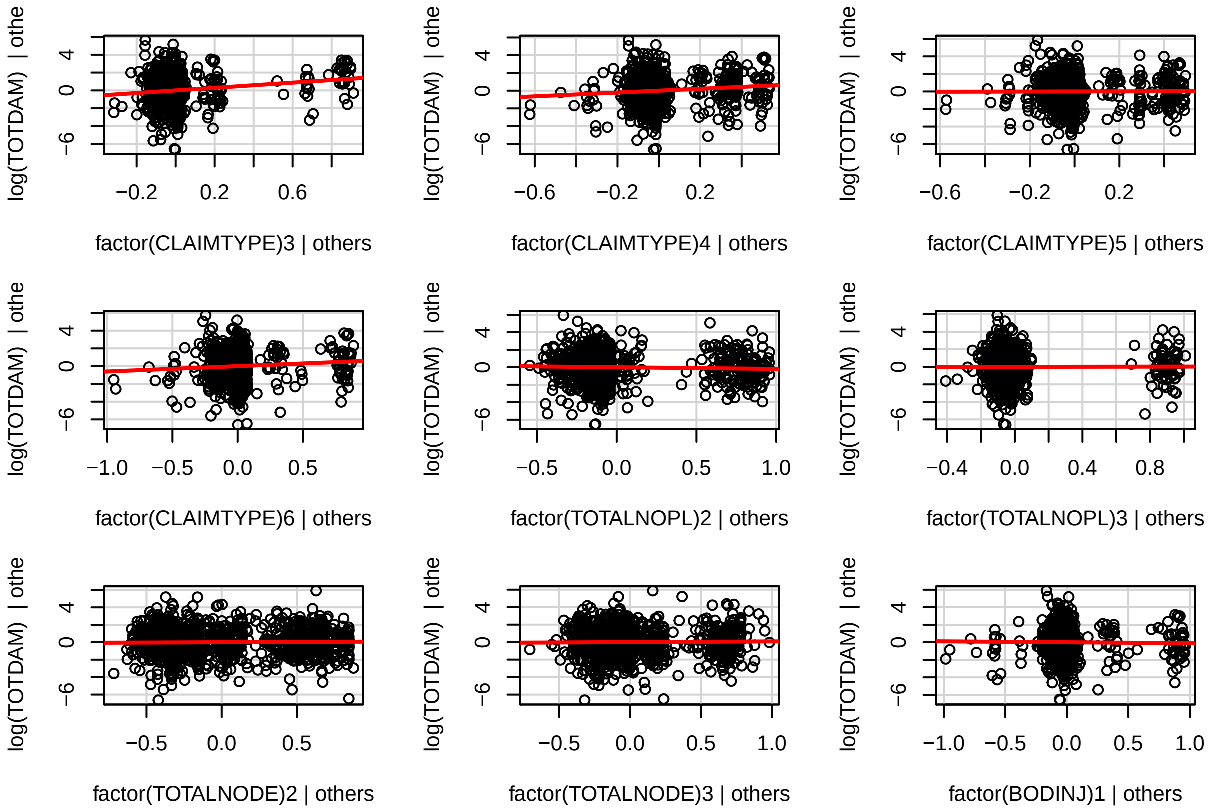
```
# log transform only response
lm.2.2 <- lm(log(TOTDAM) ~ TRIDAYS + factor(DECORP) + factor(DEGOVT) +
    factor(YEAR) + factor(CLAIMTYPE) + factor(TOTALNOPL) + factor(TOTALNODE) +
    factor(BODINJ) * DEMANDED, data = df.awardsize)
# summary(lm.2)
par(mfrow = c(1, 4))
plot(lm.2.2)
```

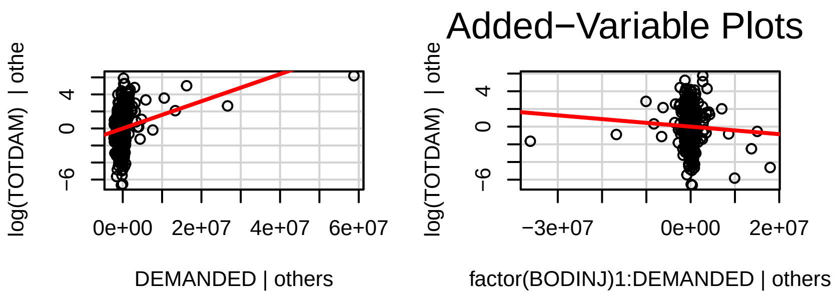



```r
av.plots(lm.2.2)
```

```
## Warning: 'av.plots' is deprecated.
## Use 'avPlots' instead.
## See help("Deprecated") and help("car-deprecated").
```

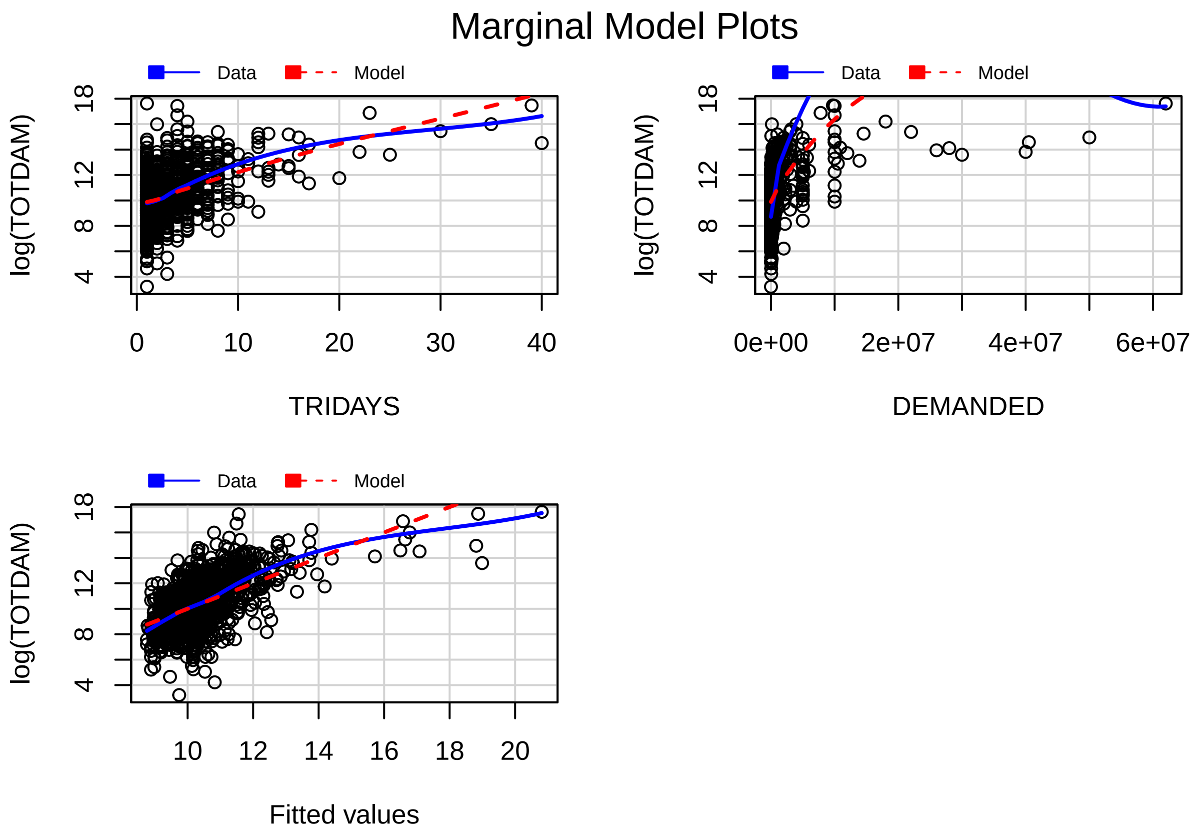



log(TOTDAM) | othe
factor(CLAIMTYPE)3 | others
−0.2 0.2 0.6
factor(CLAIMTYPE)4 | others
−0.6 −0.2 0.2
factor(CLAIMTYPE)5 | others
−0.6 −0.2 0.2
factor(CLAIMTYPE)6 | others
−1.0 −0.5 0.0 0.5
factor(TOTALNOPL)2 | others
−0.5 0.0 0.5 1.0
factor(TOTALNOPL)3 | others
−0.4 0.0 0.4 0.8
factor(TOTALNODE)2 | others
−0.5 0.0 0.5
factor(TOTALNODE)3 | others
−0.5 0.0 0.5 1.0
factor(BODINJ)1 | others
−1.0 −0.5 0.0 0.5 1.0
−6 0 4

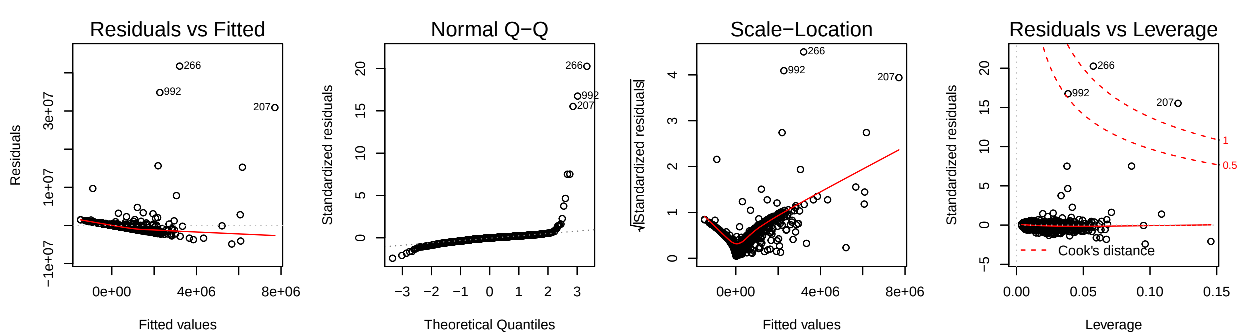



```r
mmps(lm.2.2)
```

```
## Warning in mmps(lm.2.2): Interactions and/or factors skipped
```

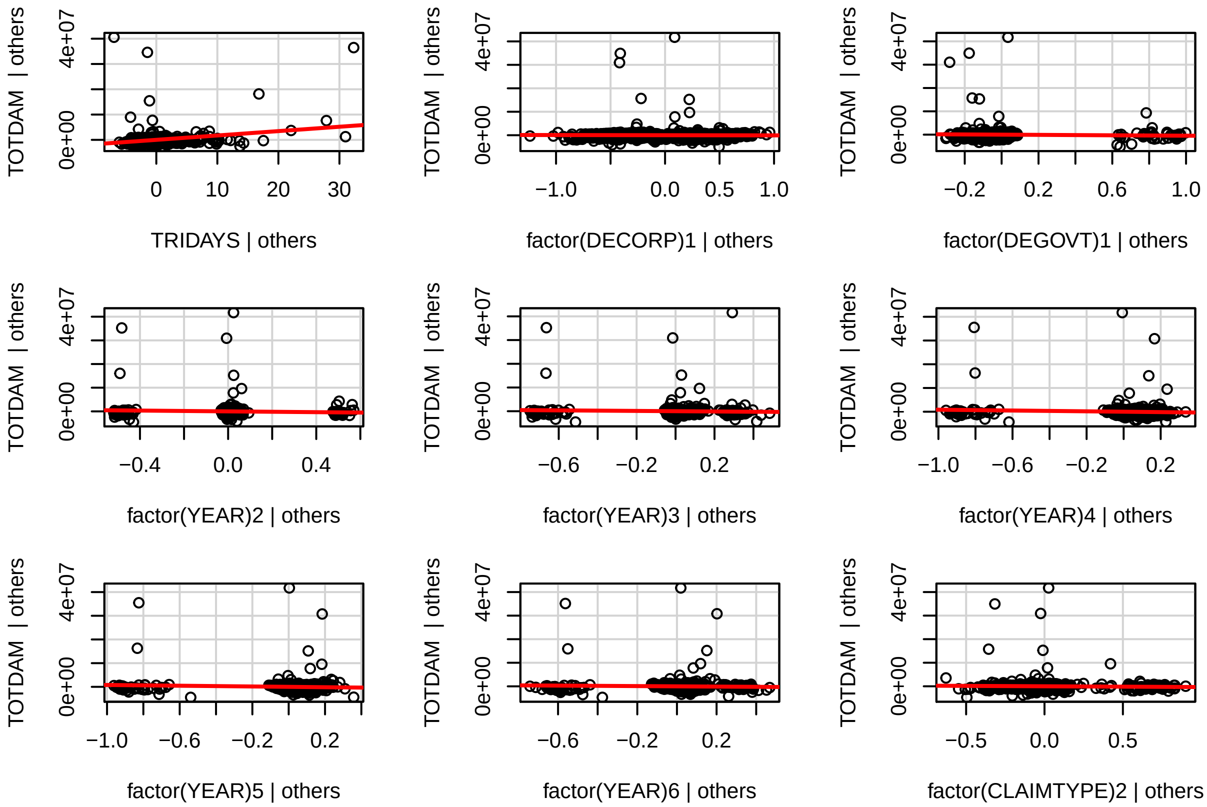


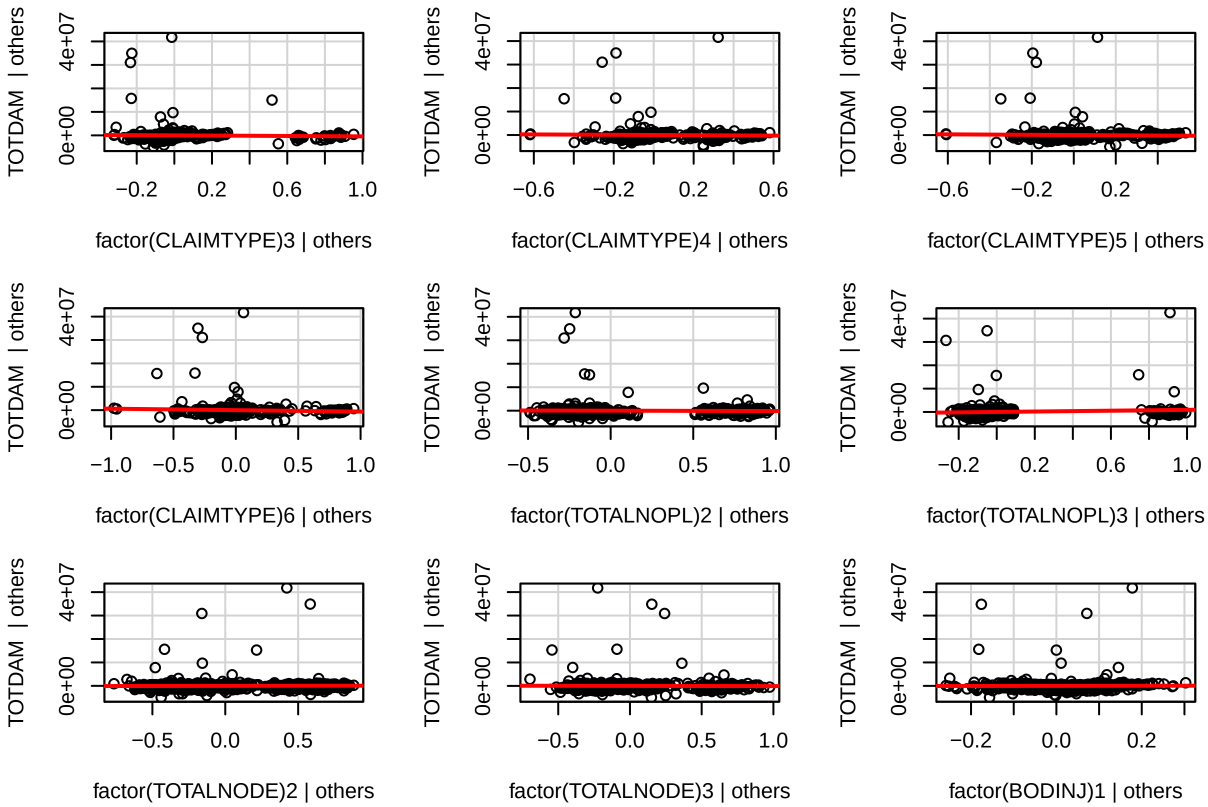

```r
# Log transform only Demanded
lm.2.3 <- lm(TOTDAM ~ TRIDAYS + factor(DECORP) + factor(DEGOVT) +
    factor(YEAR) + factor(CLAIMTYPE) + factor(TOTALNOPL) + factor(TOTALNODE) +
    factor(BODINJ) * log(DEMANDED), data = df.awardsize)
# summary(lm.2)
par(mfrow = c(1, 4))
plot(lm.2.3)
```

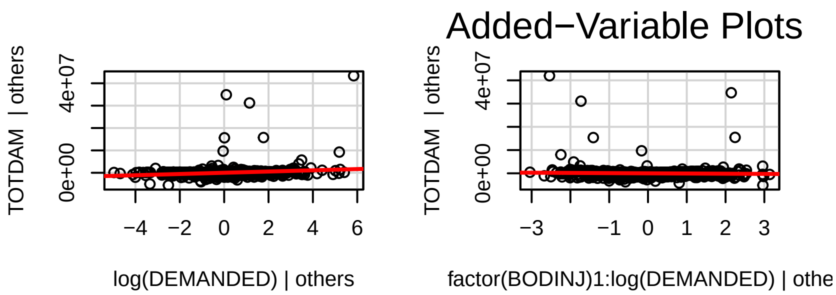



```r
av.plots(lm.2.3)
```

```
## Warning: 'av.plots' is deprecated.
## Use 'avPlots' instead.
## See help("Deprecated") and help("car-deprecated").
```

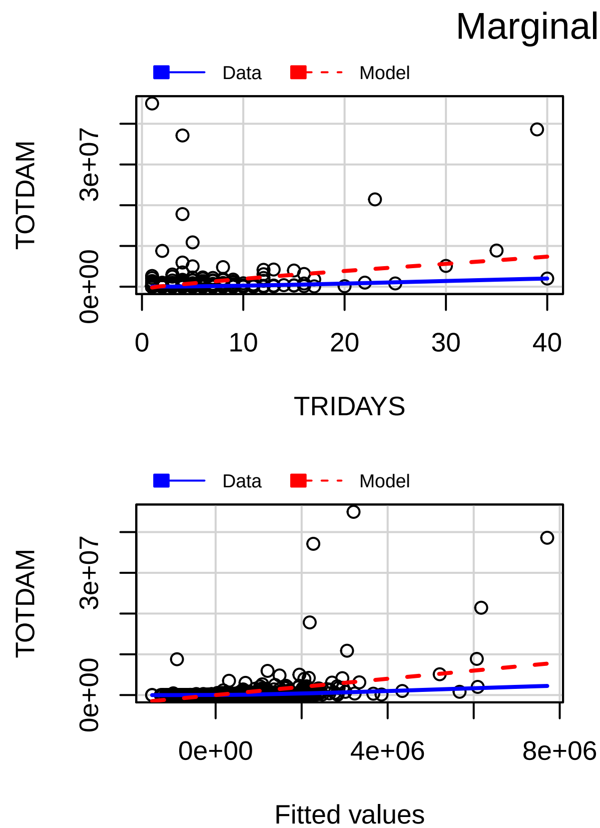

TOTDAM | others
TRIDAYS | others
factor(DECORP)1 | others
factor(DEGOVT)1 | others
factor(YEAR)2 | others
factor(YEAR)3 | others
factor(YEAR)4 | others
factor(YEAR)5 | others
factor(YEAR)6 | others
factor(CLAIMTYPE)2 | others


TOTDAM | others
0e+00
4e+07
−0.2
0.2
0.6
1.0
factor(CLAIMTYPE)3 | others
−0.6
−0.2
0.2
0.6
factor(CLAIMTYPE)4 | others
−0.6
−0.2
0.2
factor(CLAIMTYPE)5 | others
−1.0
−0.5
0.0
0.5
1.0
factor(CLAIMTYPE)6 | others
−0.5
0.0
0.5
1.0
factor(TOTALNOPL)2 | others
−0.2
0.2
0.6
1.0
factor(TOTALNOPL)3 | others
−0.5
0.0
0.5
factor(TOTALNODE)2 | others
−0.5
0.0
0.5
1.0
factor(TOTALNODE)3 | others
−0.2
0.0
0.2
factor(BODINJ)1 | others




```r
mmps(lm.2.3)
```

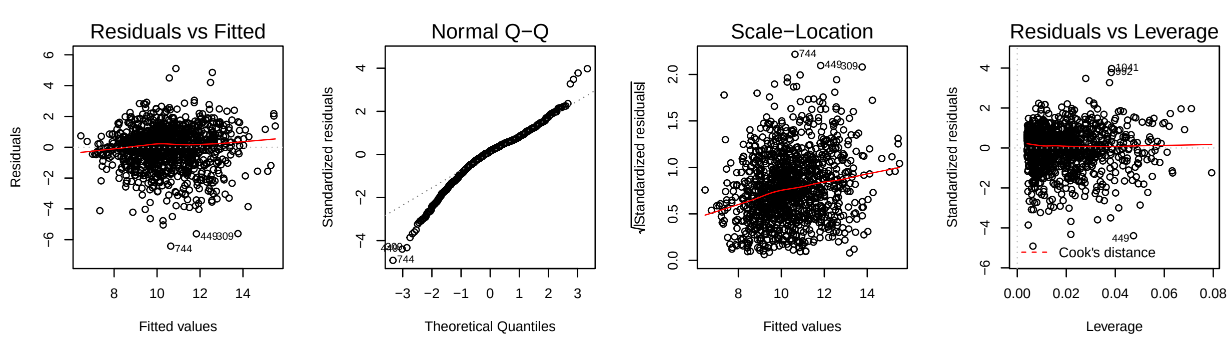
```
## Warning in mmps(lm.2.3): Interactions and/or factors skipped
```



As we can observe from the plots above, our model with transformation on only one of the variables (DEMANDED, TOTDAM) does not improve our model fitness. The residual assumptions are not satisfying and the marginal model plots are suggesting poor fit.

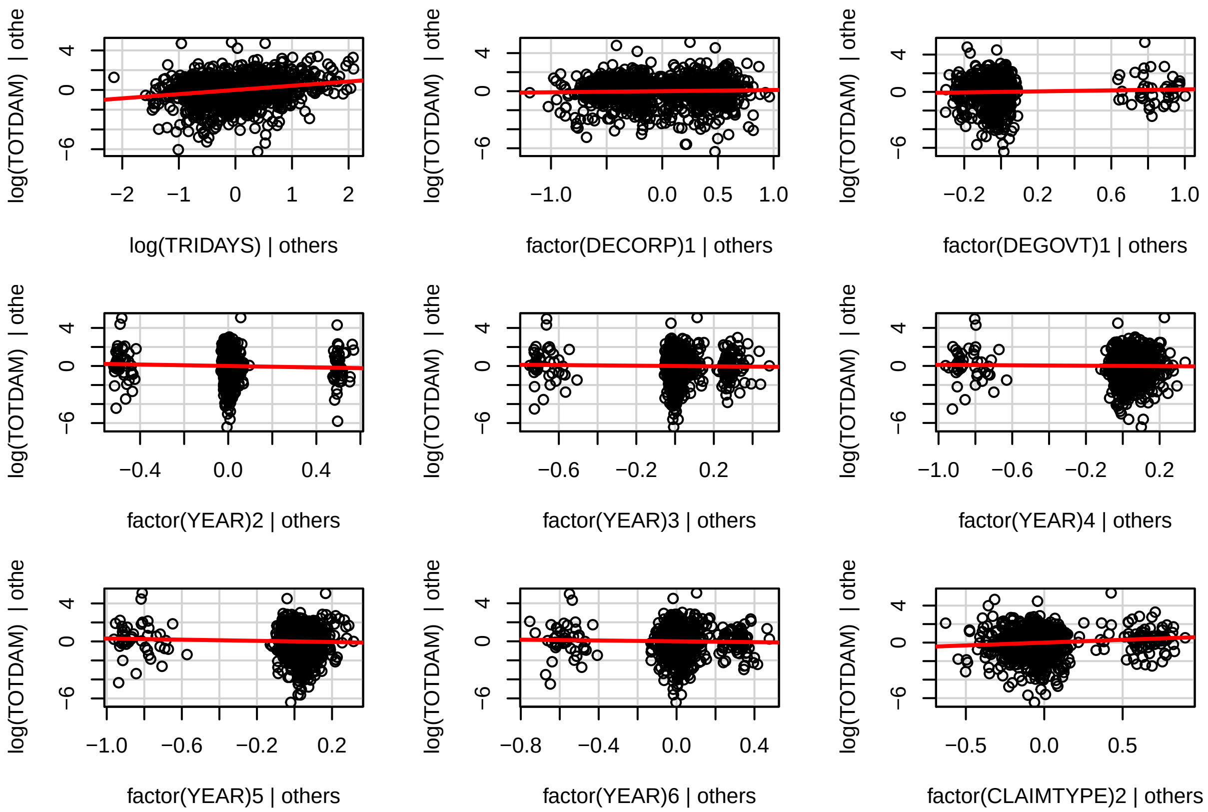
Next, we will use the model that transforms both TOTDAM and DEMANDED, and transform TRIDAYS to see if the model fitness improves.

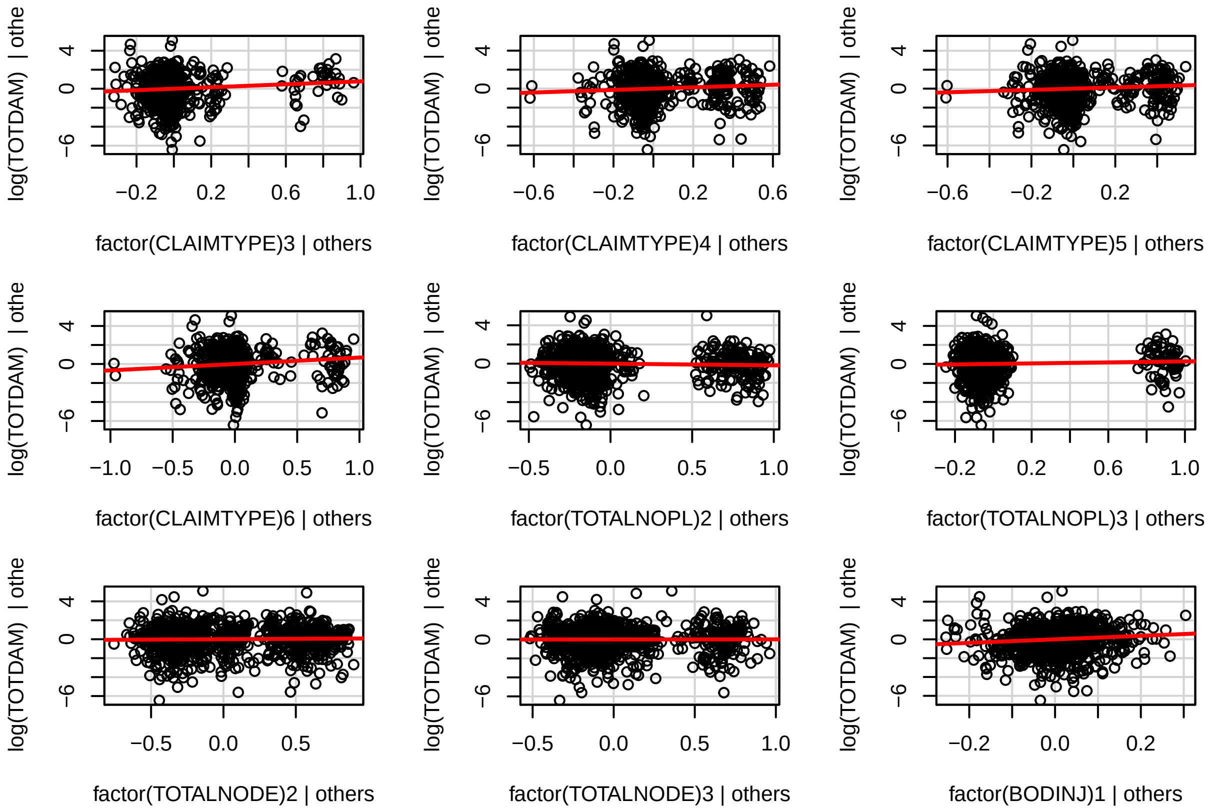
```r
# Log transform only Demanded
lm.2.4 <- lm(log(TOTDAM) ~ log(TRIDAYS) + factor(DECORP) + factor(DEGOVT) +
    factor(YEAR) + factor(CLAIMTYPE) + factor(TOTALNOPL) + factor(TOTALNODE) +
    factor(BODINJ) * log(DEMANDED), data = df.awardsize)
# summary(lm.2.4)
par(mfrow = c(1, 4))
plot(lm.2.4)
```

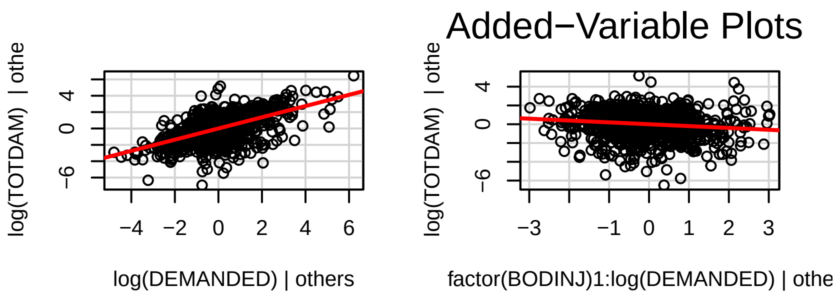


```r
av.plots(lm.2.4)
```

```
## Warning: 'av.plots' is deprecated.
## Use 'avPlots' instead.
## See help("Deprecated") and help("car-deprecated").
```

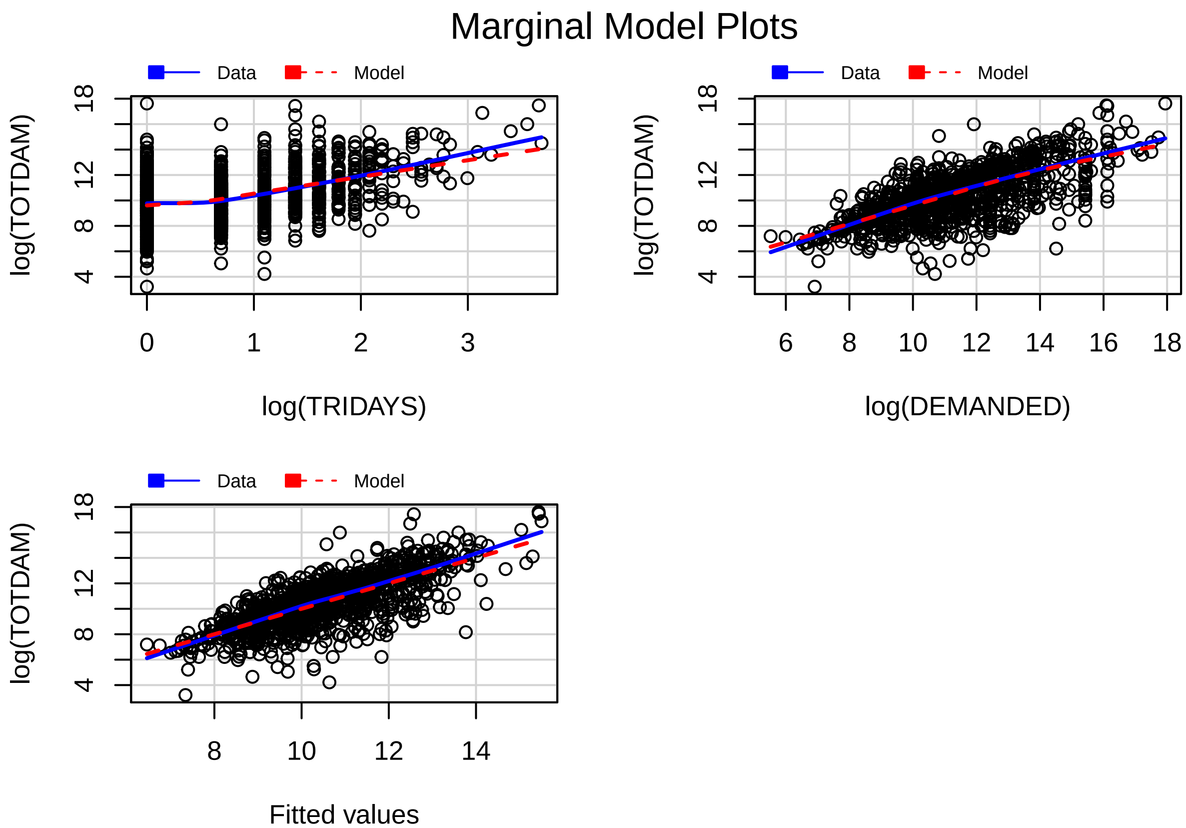



log(TOTDAM) | othe
factor(CLAIMTYPE)3 | others
factor(CLAIMTYPE)4 | others
factor(CLAIMTYPE)5 | others
factor(CLAIMTYPE)6 | others
factor(TOTALNOPL)2 | others
factor(TOTALNOPL)3 | others
factor(TOTALNODE)2 | others
factor(TOTALNODE)3 | others
factor(BODINJ)1 | others



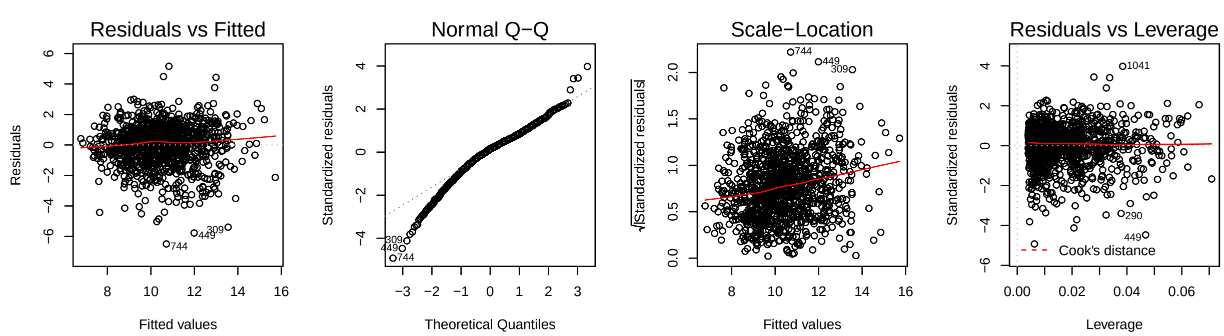

```r
mmps(lm.2.4)
```

```
## Warning in mmps(lm.2.4): Interactions and/or factors skipped
```



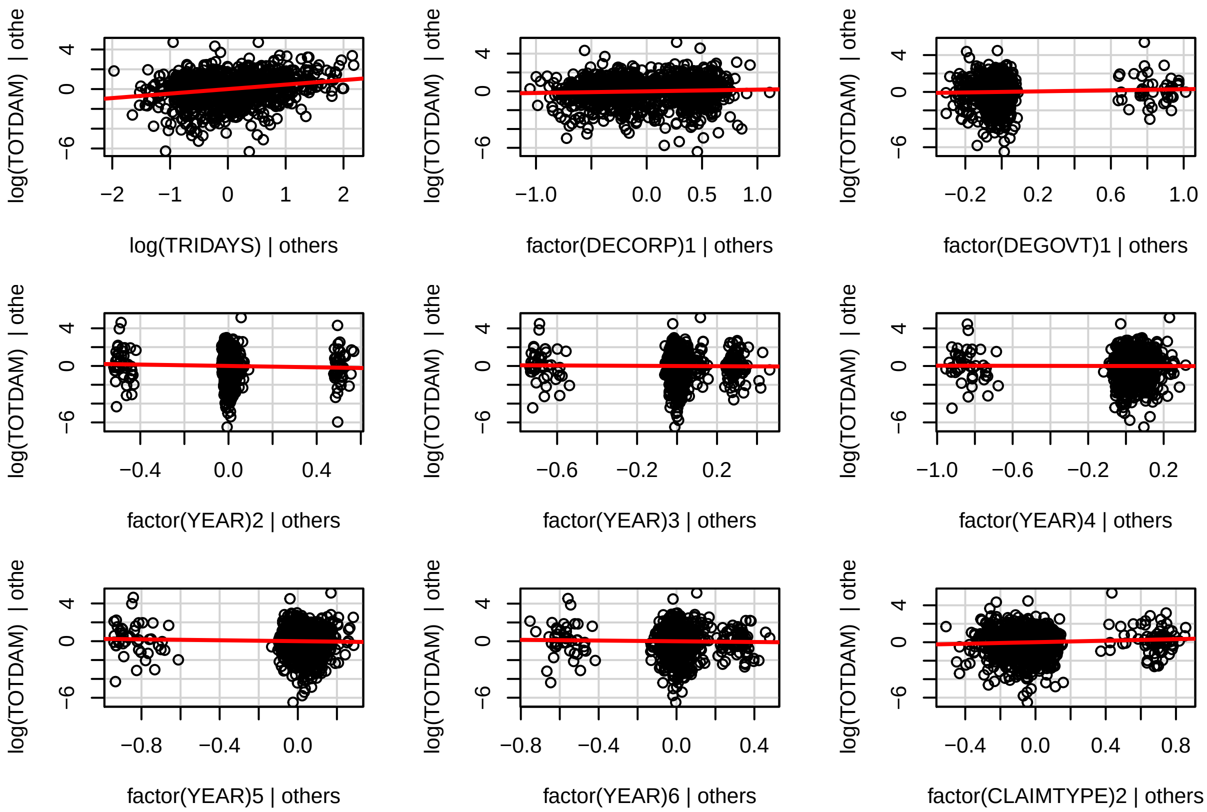
Log transformation on TRIDAYS seems to stabilize the variance further from the previous model, and we do not observe any fitness issues in the marginal model and added variable plots. The fitness on the TRIDAYS improved with the transformation. Therefore, this is the best full model that includes the interaction term. But before we proceed to model inference, we want to see if the interaction term is necessary.

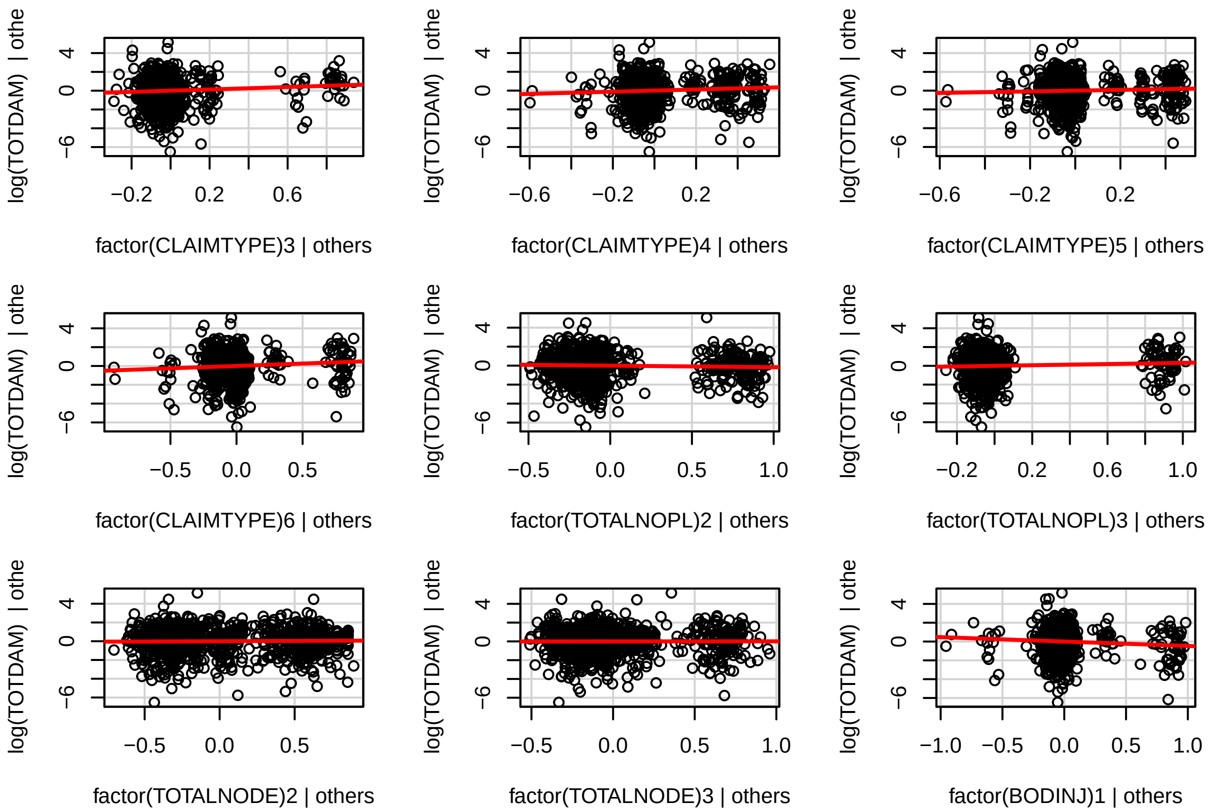
```r
# Log transform only Demanded
lm.2.5 <- lm(log(TOTDAM) ~ log(TRIDAYS) + factor(DECORP) + factor(DEGOVT) +
    factor(YEAR) + factor(CLAIMTYPE) + factor(TOTALNOPL) + factor(TOTALNODE) +
    factor(BODINJ) + log(DEMANDED), data = df.awardsize)
# summary(lm.2.4)
par(mfrow = c(1, 4))
plot(lm.2.5)
```



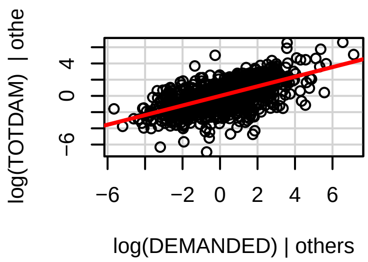
```r
av.plots(lm.2.5)
```

```
## Warning: 'av.plots' is deprecated.
## Use 'avPlots' instead.
## See help("Deprecated") and help("car-deprecated").
```



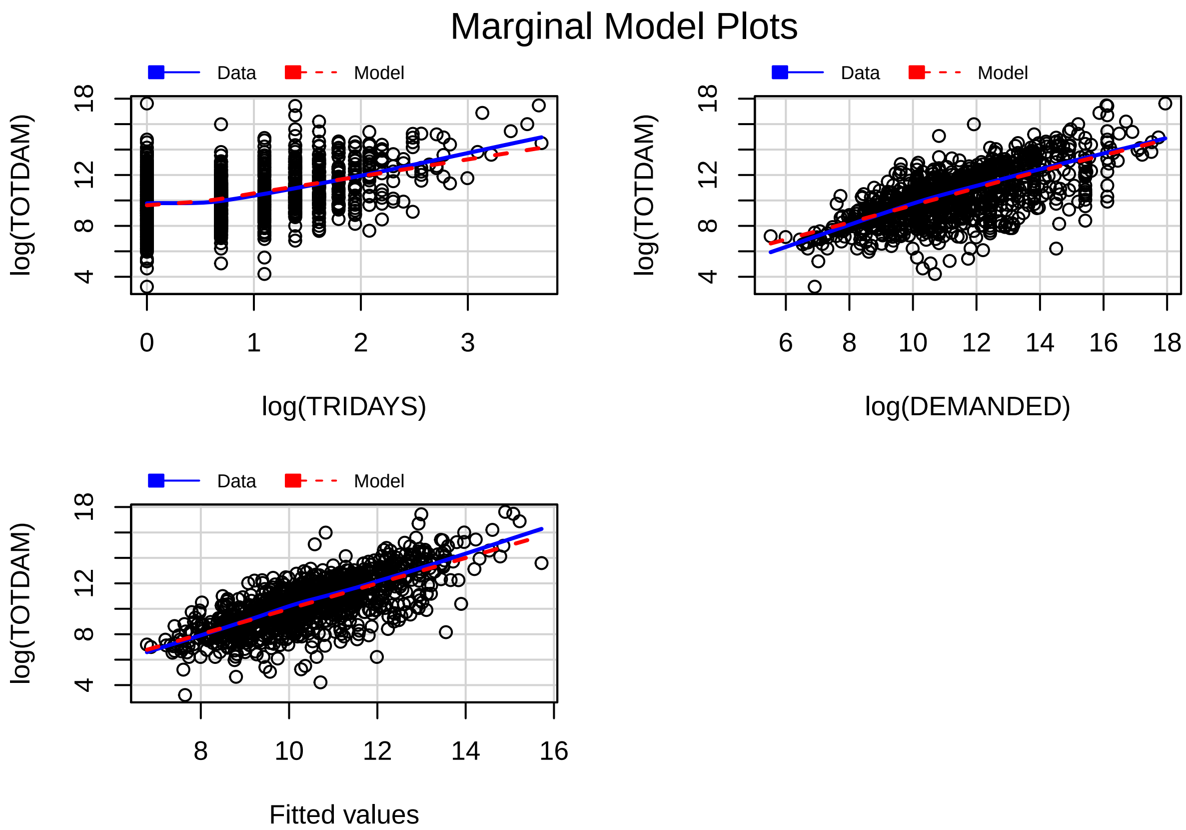

log(TOTDAM) | othe
factor(CLAIMTYPE)3 | others
factor(CLAIMTYPE)4 | others
factor(CLAIMTYPE)5 | others
factor(CLAIMTYPE)6 | others
factor(TOTALNOPL)2 | others
factor(TOTALNOPL)3 | others
factor(TOTALNODE)2 | others
factor(TOTALNODE)3 | others
factor(BODINJ)1 | others

Added−Variable Plots



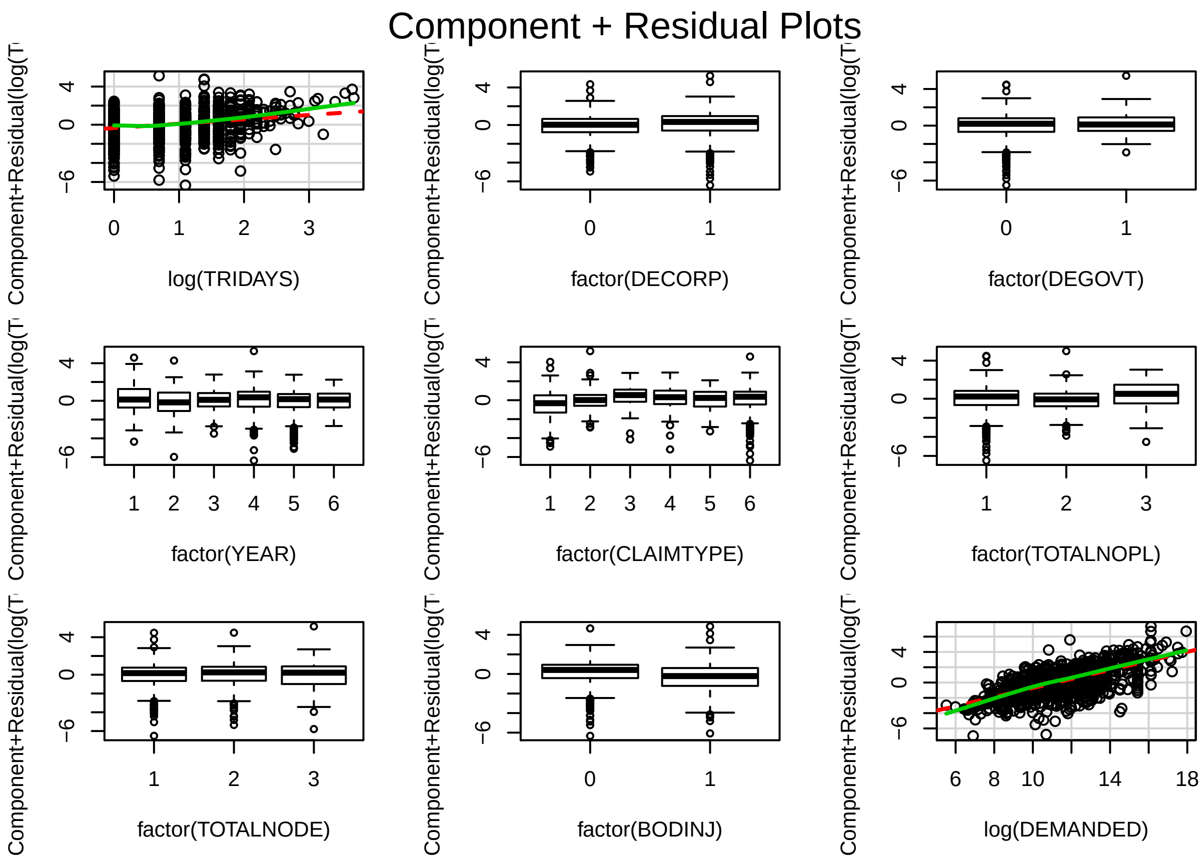

```r
mmps(lm.2.5)
```

```
## Warning in mmps(lm.2.5): Interactions and/or factors skipped
```




```
cr.plots(lm.2.5)
```

```
## Warning: 'cr.plots' is deprecated.
## Use 'crPlots' instead.
## See help("Deprecated") and help("car-deprecated").
```



When we removed the interaction term, our residual plots appear to be similar. The assumptions of constant variance, normality, zero expectation and independence is mostly met. Our added variable plot, marginal model plot and C-R plots does not show anything alarming. We then will use partial F-test to determine whether the interaction term is needed in the model.

```
kable(anova(lm.2.1, lm.2.5, test = "F"), caption = "Partial F-test")
```

Table 7: Partial F-test

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1154	1949.944	NA	NA	NA	NA
1155	2019.652	-1	-69.70888	41.25455	0

From the partial F-test, we can see that with the small P-value, we favor the full model over the reduced. Therefore, we will keep the interaction term in our final model whose predictive value is the size of damage received by the plaintiff.

$log(TOTDAM) = \beta_0 + \beta_1 * log(TRIDAYS) + \beta_2 * DECORP + \beta_3 * DEGOVT + \beta_4 * YEAR + \beta_5 * CLAIMTYPE + \beta_6 * TOTALNOPL + \beta_7 * TOTALNODE + \beta_8 * BODINJ + \beta_9 * log(DEMANDED) + \beta_{10} * log(DEMANDED){:}BODINJ$

Model Inference

Binary Award Logistic Model

Our final model is: $logit(is.Awarded) = \beta_0 + \beta_1 * TRIDAYS + \beta_2 * DECORP + \beta_3 * DEGOVT + \beta_4 * YEAR.recat + \beta_5 * CLAIMTYPE + \beta_6 * TOTALNOPL + \beta_7 * TOTALNODE + \beta_8 * BODINJ + \beta_9 * DEMANDED + \beta_{10} * DEMANDED{:}BODINJ$

```
pander(summary(glm.1.a))
```

	Estimate	Std. Error	z value	Pr(>|z|)
TRIDAYS	0.03467	0.01737	1.997	0.04585
factor(DECORP)1	0.08316	0.1175	0.7075	0.4793
factor(DEGOVT)1	-0.5641	0.2336	-2.415	0.01575
factor(YEAR.recat)1	0.5417	0.1284	4.219	2.452e-05
factor(CLAIMTYPE)2	-0.4113	0.2062	-1.994	0.04613
factor(CLAIMTYPE)3	-1.303	0.2485	-5.243	1.577e-07
factor(CLAIMTYPE)4	-0.8594	0.304	-2.827	0.004697
factor(CLAIMTYPE)5	-0.4512	0.3339	-1.351	0.1767
factor(CLAIMTYPE)6	-0.4243	0.2305	-1.84	0.06574
factor(TOTALNOPL)2	-0.03461	0.1387	-0.2496	0.8029
factor(TOTALNOPL)3	0.6343	0.261	2.43	0.01509
factor(TOTALNODE)2	0.2613	0.1217	2.148	0.03175
factor(TOTALNODE)3	0.07657	0.1629	0.4702	0.6382
factor(BODINJ)1	-0.9493	0.2206	-4.304	1.677e-05
DEMANDED	-3.76e-08	2.305e-08	-1.631	0.1029
factor(BODINJ)1:DEMANDED	3.405e-08	2.572e-08	1.324	0.1855
(Intercept)	0.7616	0.2697	2.824	0.00474

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	2399 on 1835 degrees of freedom
Residual deviance:	2251 on 1819 degrees of freedom

```
glm.result = cbind(round(exp(glm.1.a$coefficients), 3), round(exp(confint(glm.1.a)),
    3), ifelse(round(summary(glm.1.a)$coeff[, 4], 3) == 0, "<0.001",
    round(summary(glm.1.a)$coeff[, 4], 3)))
```

```
## Waiting for profiling to be done...
```

```
colnames(glm.result) = c("AOR", "CI lb", " CI ub %", "Pval")
rownames(glm.result) = c("Intercept", "Trial Days", "Is.Corporate",
    "Is.Gov't", "Is.after.1999", "premises liability claims",
    "malpractice claims", "fraud claims", "rental claims", "other claims",
    "Two Plaintiffs", "Three or more Plaintiffs", "Two Defendant",
    "Three or more Defendants", "Injury Claims", "Damage Demanded",
    "Injury:Demanded")
pander(glm.result)
```

	AOR	CI lb	CI ub %	Pval
Intercept	2.142	1.262	3.636	0.005
Trial Days	1.035	1.001	1.072	0.046
Is.Corporate	1.087	0.863	1.368	0.479
Is.Gov't	0.569	0.359	0.898	0.016
Is.after.1999	1.719	1.336	2.21	<0.001
premises liability claims	0.663	0.442	0.993	0.046
malpractice claims	0.272	0.165	0.439	<0.001
fraud claims	0.423	0.234	0.77	0.005
rental claims	0.637	0.333	1.236	0.177
other claims	0.654	0.416	1.028	0.066
Two Plaintiffs	0.966	0.737	1.27	0.803
Three or more Plaintiffs	1.886	1.148	3.205	0.015
Two Defendant	1.299	1.024	1.651	0.032
Three or more Defendants	1.08	0.786	1.489	0.638
Injury Claims	0.387	0.251	0.596	<0.001
Damage Demanded	1	1	1	0.103
Injury:Demanded	1	1	1	0.185

We can interpret the final result in terms of adjusted odds ratio. To answer our research questions, we first look at relationship between the odds of being awarded damage and whether the defendant is corporation. In this model, we did not have enough evidence to suggest that the odds were any different when defendant is a corporation. We concluded that the relationship between the odds of being awarded damage and the amount of damage requested was not statistically significant. We did not have enough evidence to conclude that the relationship between the odds of being awarded damage and the amount demanded could be altered by whether the case involves bodily injuries. In cases with three or more plaintiffs, however, were associated with higher odds on average of being awarded damage (AOR = 1.982). Contrary to what we see in the EDA, our model suggested that between 1999 - 2001, the odds of a plaintiff being awarded damage was higher when compared to the cases before 1999 (AOR = 1.385). Note that when the defendant is the government, the odds of the plaintiff receiving any damage was decreased when comparing to the cases where the defendant is not the government (AOR = 0.546). Of all the claim types, motor vehicle cases had the highest odds of awarding damage to plaintiff because the AOR of all other claim types are less than 1, and claims of malpractice has the lowest odds for damage award (AOR = 0.254).

```
lm.result = cbind(round((lm.2.4$coefficients), 3), round((confint(lm.2.4)),
    3), ifelse(round(summary(lm.2.4)$coeff[, 4], 3) == 0, "<0.001",
    round(summary(lm.2.4)$coeff[, 4], 3)))
rownames(lm.result) = c("Intercept", "log(Trial Days)", "Is.Corporate",
    "Is.Gov't", "1997", "1998", "1999", "2000", "2001", "premises liability claims",
    "malpractice claims", "fraud claims", "rental claims", "other claims",
    "Two Plaintiffs", "Three or more Plaintiffs", "Two Defendant",
    "Three or more Defendants", "Injury Claims", "log(Damage Demanded)",
    "Injury:log(Demanded)")
colnames(lm.result) = c("AOR", "CI lb", " CI ub %", "Pval")
pander(lm.result)
```

	AOR	CI lb	CI ub %	Pval
Intercept	2.159	1.291	3.027	<0.001
log(Trial Days)	0.426	0.308	0.543	<0.001
Is.Corporate	0.121	-0.054	0.295	0.175
Is.Gov't	0.271	-0.153	0.695	0.21

	AOR	CI lb	CI ub %	Pval
1997	-0.385	-0.944	0.174	0.177
1998	-0.164	-0.639	0.31	0.496
1999	-0.109	-0.533	0.316	0.616
2000	-0.312	-0.737	0.114	0.151
2001	-0.219	-0.732	0.294	0.403
premises liability claims	0.608	0.239	0.977	0.001
malpractice claims	0.775	0.286	1.264	0.002
fraud claims	0.679	0.187	1.17	0.007
rental claims	0.646	0.126	1.165	0.015
other claims	0.674	0.287	1.061	0.001
Two Plaintiffs	-0.171	-0.389	0.047	0.123
Three or more Plaintiffs	0.266	-0.064	0.597	0.114
Two Defendant	0.084	-0.093	0.261	0.351
Three or more Defendants	0.013	-0.232	0.257	0.919
Injury Claims	1.873	0.83	2.915	<0.001
log(Damage Demanded)	0.685	0.627	0.744	<0.001
Injury:log(Demanded)	-0.197	-0.279	-0.115	<0.001

This model was built on cases that were awarded damage. Because our linear model was built with the log-transformed amount of damages awarded and requested, we could interpret the coefficients in terms of ratio. An increase in 1% of demanded damage from the plaintiffs were, on average, associated with an 0.68% increase in awarded damages when no bodily injures were involved in the case. However, the same 1% increase in demanded damage in cases with bodily injury claims were on average only associated with 0.49% in awarded damage. Nevertheless, cases with bodily injury claims on average were awarded 5.33 times amount in damage than the cases without such claims, holding all other variables constant. We did not have enough evidence to suggest that there was a relationship between the amount awarded and whether the case occurred in recent years. We also conclude that the relationship between the awarded damage amount and whether the defendant was a corporation was statistically insignificant. Additionally, our model did not have evidence that there was a relationship between the number of plaintiffs and the amount of damage awarded. Note that an increase in 1% of trial length was on average associated with an 0.425% increase in money awarded, and that all the other claim types other than motor vehicle claims received less in damage award.

Therefore, to address our initial hypothesis, we concluded that a more recent trail and cases with three or more plaintiffs are predictive of a higher probability that damages were awarded. We did not have enough evidence to support that the relationship exists with the other variables that we have proposed.

In terms of the amount awarded, we concluded that there existed a relationship between the damage requested and the damage awarded, and whether the case involved a bodily injury claim alters this relationship. For cases with bodily injury claims, the amount demanded had a weaker effect on the total damages. We did not found any statistically significant relationships between the other hypothesized variables and the amount rewarded.