Investigating the Features of Lawsuit Cases to Predict the Size of Total Damage Awards

To Jared the Statistician

Mengran He

16 December, 2016

Exploratory Data Analysis

The data collection was carried out by the National Center for State Courts for the goal of investigating what features of a particular case are predictive of the probability that damages are awarded and the size of total damage awards. In particular, the litigation advocacy watchdog group intially hypothesized that a more recent trial, a corporation as the defendant, and an increased number of plaintiffs are predictive of a higher probability that damages are awarded and also higher damages. Plus, they assumed that the amount of demanded has a larger effect on the total damages when there is a bodily injury. There are 1836 civil lawsuit cases in the dataset, 64% of them received damage awards (1175 out of 1836 cases). We first converted the following variables into factor: `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"), `claimtype` type of claim the plaintiff made (1:motor vehicle; 2:premises liability; 3:malpractice; 4:fraud; 5:rental/lease; 6:other). Variables total amount of damages awarded to plaintiff `totdam` and total amount of damages requested by plaintiff `demanded` are converted to continuous because they represents the amount of money. We also created a dummy variable `award` indicating whether the damages was awarded, if total amount of damages is \$0, `award` = 0, otherwise `award` = 1.

Variable `year` means the year the civil lawsuit was file (1: pre-1997; 2: 1997; 3:1998; 4:1999; 5:2000; 6:2001). Figure 1a shows that the average amount of damages awarded to plaintiff generally decreased from years before 1997 to year 2001 except a small bump in year 1998. The barplot in figure 1b shows that the average amount of damages requested by plaintiff also decreased from the beginning of year 1997 to year 2001. We assumed a negative relationship between the year the civil lawsuit was filed and the average of damages awarded or requested, so we converted the `year` as numeric variable. Figure 1c and 1d suggest that as the number of plaintiffs `totalnopl` (1:one plaintiff; 2: two plaintiffs; 3: 3 or more plaintiffs) increased, the average amount of damages awarded and requested also increased, the number of plaintiffs `totalnopl` is considered as integer variable. Figure 1e and 1f show that the average amount of damages awarded and requested increased as the number of defendants increased `totalnode` (1:one defendant; 2:two defendants; 3: 3 or more defendants), the number of defendants `totalnode` is considered as integer variable. Figure 1g shows 1076 out of 1836 trials lasted 2 or less days, the distribution of days the trial lasted `tridays` is heavily right skewed, we converted `tridays` as numeric variable.

Figure 2 summarizes the frequency of whether damages were awarded to plaintiff `award` (1:"Yes", 0: "No") based on different categories. Among 1175 cases with damage awards received, 411 cases included bodily injury and 764 cases didn't include body injury (1:"Yes", 0: "No"); 27 cases where the defendants were corporation and 548 cases where defendants were not corporation (1:"Yes", 0:"No"); 267 cases were motor vehicle claims, 76 cases were premises liability claims, 36 were malpractice, 81 were fraud, 67 were rental/lease, and 648 are others; 26 cases had 1 plaintiff, 181 cases had 2 plaintiffs, 68 cases had 3 or more plaintiffs; 642 cases had 1 defendant, 363 cases had 2 defendants, 170 cases had 3 or more defendants. By eyeballing figure 2, we can see that the number of cases which didn't include bodily injury but received damages awards is more than the number of cases which included bodily injury and received damages awards. The number of cases which didn't have corporation as defendants but received damages awards is more than the number of cases which had corporation as defendants and received damages awards. The number of cases which had 2

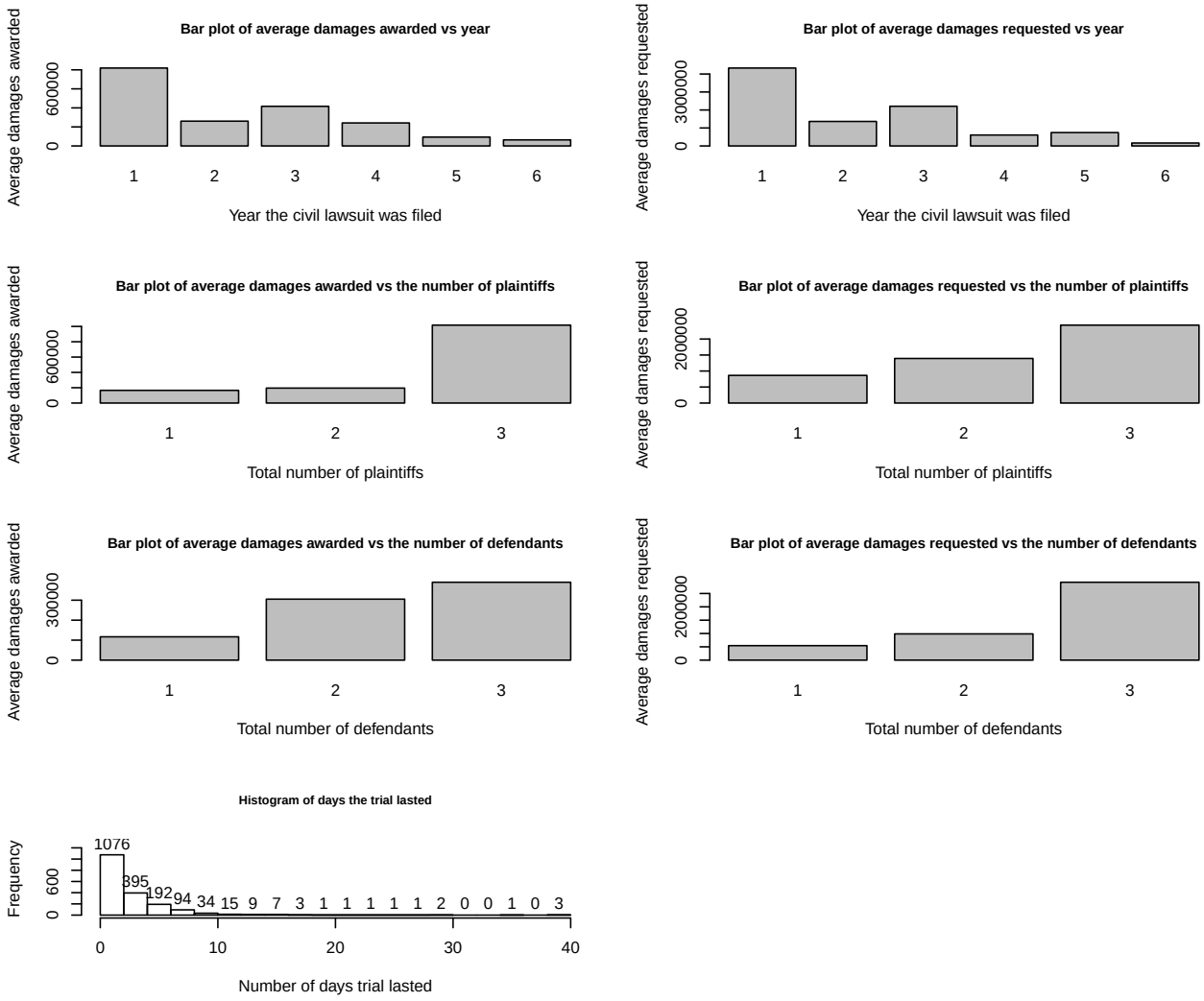


Figure 1: a) Bar plot of average damages awarded vs year the lawsuit was filed, b) Bar plot of average damages requested vs year the lawsuit was filed, c) Bar plot of average damages awarded vs the number of plaintiffs, d) Bar plot of average damages requested vs the number of plaintiffs, e) Bar plot of average damages awarded vs the number of defendants, f) Bar plot of average damages requested vs the number of defendants, g) Histogram of days the trial lasted

Summary of Whether Damages Awards Were Received (1: Yes, 0: No)

	Bodily Injury				Corporation	
Award	0	1		Award	0	1
0	306	355		0	360	301
1	764	411		1	548	627
	Number of Plaintiffs					
Award	1	2	3			
0	515	122	24			
1	926	181	68			
	Number of Defendants					
Award	1	2	3			
0	378	179	104			
1	642	363	170			
	Claim Type					
Award	1	2	3	4	5	6
0	167	82	71	47	25	269
1	267	76	36	81	67	648

Figure 2: Summary of whether damage awards were received

plaintiffs and received the damages awards is greater than the number of cases which had 3 or more plaintiffs and received the damages awards.

Table 1 summarizes the total amount of damages requested by plaintiff `demanded`, where the minimum requested amount is \$250, the maximum is \$100000000, the average is \$1030000 and the median is \$60290. It suggests that the mean is far greater than the median, the distribution of damages requested by plaintiff is right skewed. Table 2 summarizes the total amount of damages awarded to plaintiff `totdam` where the minimum is \$0, the maximum is \$44970000, the average is \$213100 and the median is \$7795, which means the distribution of damages awarded to plaintiff is also right skewed.

Table 1: Summary of total amount of damages requrested (in \$)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
250	23500	60300	1030000	305000	100000000

Table 2: Summary of total amount of damages awarded to plaintiff (in \$)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	7800	213000	50300	45000000

Figure 3 suggests that there are outliers in both variables. The case of highest damages requested was case 965: plaintiff requested \$100025000 but no damages was awarded, trial lasted 2 days, a bodily injury was included in premises liability claim, the lawsuit was filed in 1998, and one plaintiff and more than 3 defendants were included. The case of highest damages awarded was case 266: more than three plaintiffs requested \$62000000 and they were awarded \$44968563, the trial lasted 1 day, no bodily injury was incldued in fraud claim, the lawsuit was filed in 1998, and there were 2 defendants. Figure 4a suggests that the total amount of damages awarded to plaintiff is not linearly related to the total amount of damages requested by plaintiff. By taking logarithms transformation on these two variables shown in figure 4b, we can see a linear trend in this plot surrounded by noises.

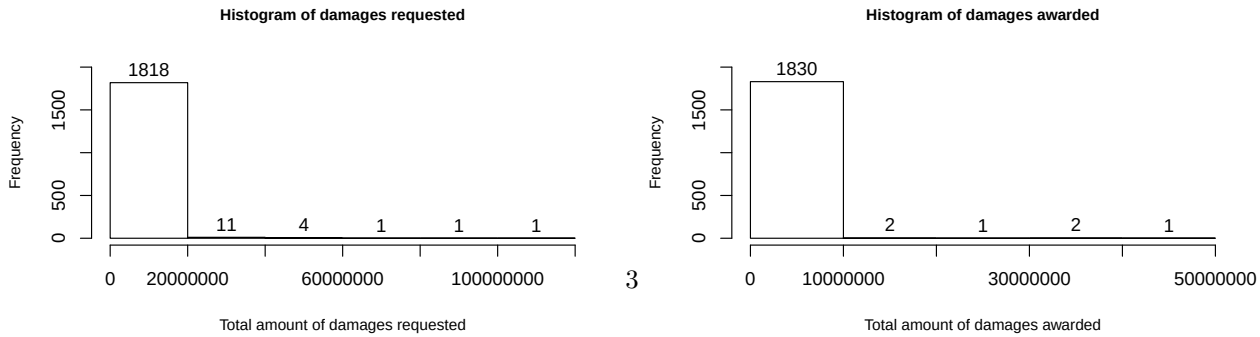


Figure 3: a) Histogram of damages requested by plaintiff, b) Histogram of damages awarded to plaintiff

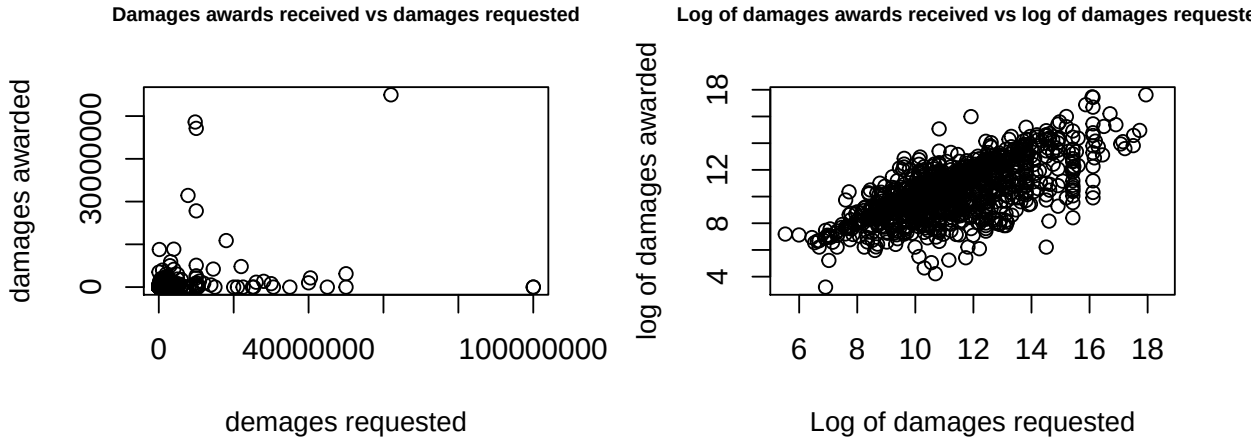


Figure 4: a) Scatter plot of damages awards received vs damages requested, b) Scatter plot of Log of damages awards received vs log of damages requested

Model Selection

Part 1: Regression model of predicting whether any damages are awarded

Initial Modeling To predict whether any damages are awarded, we fitted a logistic regression model where response variable is log odds of probability that the damages are awarded, p(award = 1). In our first model, model 1: $log\frac{p}{1-p} = \beta_0 + \beta_1 demanded + \beta_2 tridays + \beta_3 bodinj + \beta_4 decorp + \beta_5 degovt + \beta_6 year + \beta_7 claimtype + \beta_8 totalnopl + \beta_9 totalnode + \epsilon$. Table 3 suggests that variables `bodinj` whether bodily injury was included, `degovt` where the defendant was the government, `year` the civil lawsuit was filed, claimtype are statistically significant in predicting the probability of whether the damages are awarded.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.1287	0.3509	0.37	0.7137
demanded	-0.0000	0.0000	-1.18	0.2360
tridays	0.0365	0.0177	2.06	0.0391
bodinj1	-0.9034	0.2177	-4.15	0.0000
decorp1	0.0805	0.1167	0.69	0.4903
degovt1	-0.5244	0.2342	-2.24	0.0251
year	0.1834	0.0466	3.94	0.0001
claimtype2	-0.4057	0.2050	-1.98	0.0478
claimtype3	-1.3235	0.2474	-5.35	0.0000
claimtype4	-0.8975	0.3020	-2.97	0.0030
claimtype5	-0.4744	0.3325	-1.43	0.1537
claimtype6	-0.4181	0.2289	-1.83	0.0677
totalnopl	0.1746	0.0974	1.79	0.0732
totalnode	0.0986	0.0769	1.28	0.1999

Table 3: Summary of model 1, predicting whether any damages are awarded

Diagnostics of Initial Model Nevertheless, in figure 5, the binned residuals plots of model 1 suggest that the binned residuals are not independent from the amount of damages requested and pattern shows in binned residuals vs fitted values plot. In addition, in figure 6, the component+reidual plots of model 1 suggests that the damages requested `demanded` doesn't have linear relationship with response variable. So, model 1 doesn't fit the data very well.

Tranformations and Diagnositcs According to figure 4a and 4b, the amount of damages requested is

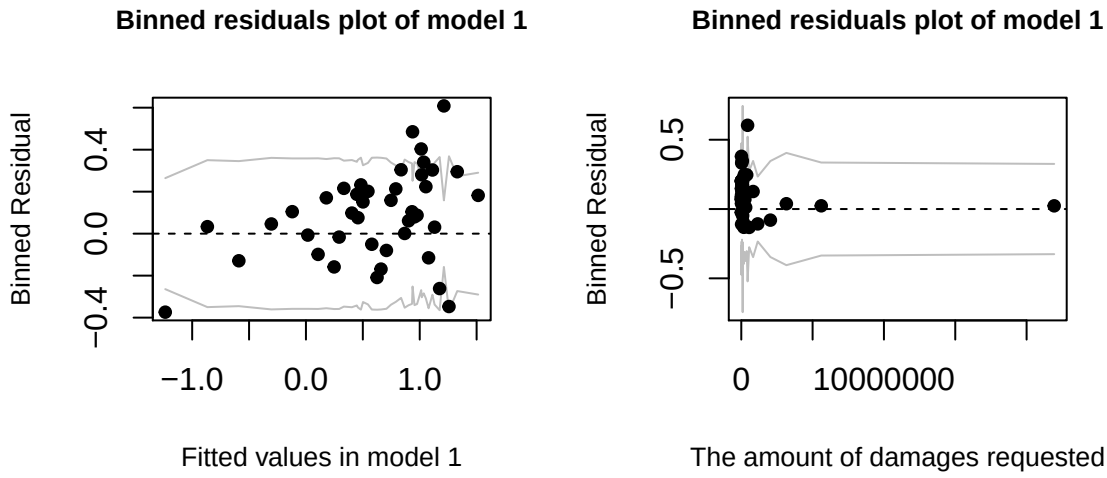


Figure 5: a) Binned residuals vs fitted values in model 1, b) Binned residuals vs damages requested by plaintiffs

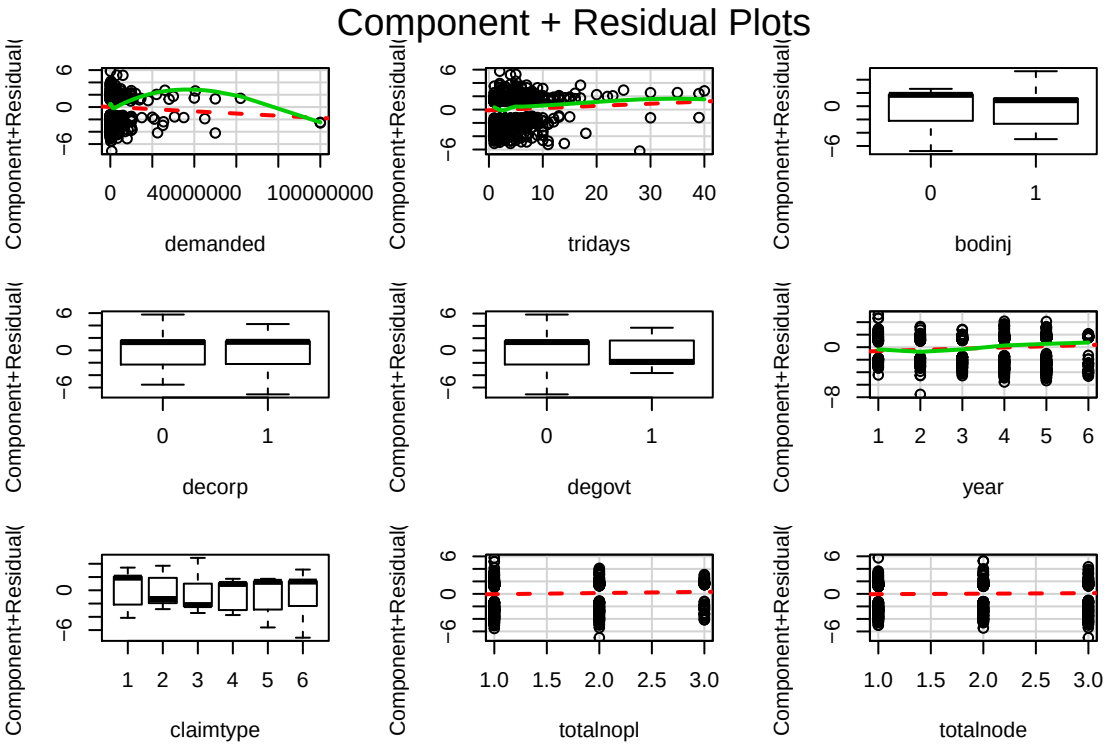


Figure 6: Component+Residual plots of model 1

not linearly related to the amount of damages awarded, but the log of the amount of damages requested is positively related to the log of the amount of damages awarded. We took logarithms transformation on variable `demanded` the amount of damages requested and fitted next logistic regression model 1.1: $log\frac{p}{1-p} = \beta_0 + \beta_1 log(demanded) + \beta_2 tridays + \beta_3 bodinj + \beta_4 decorp + \beta_5 degovt + \beta_6 year + \beta_7 claimtype + \beta_8 totalnopl + \beta_9 totalnode + \epsilon$. Figure 7 shows that there is no pattern in binned residual plots and the binned residuals are independent from the log of damages requested by plaintiffs.

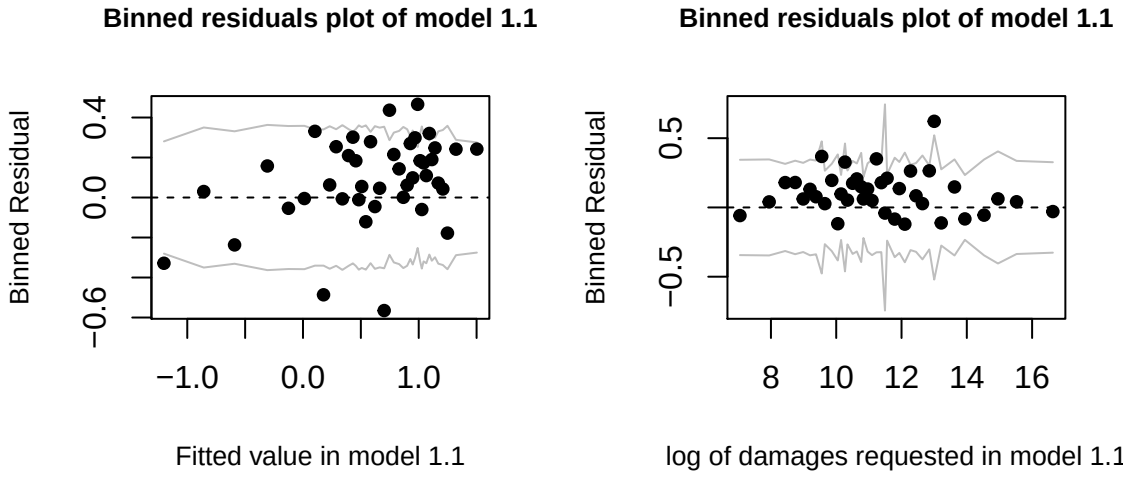


Figure 7: Binned residuals vs fitted values in model 1.1 , b) Binned residuals vs damages requested in model 1.1

The component+residual plots and marginal model plots in figure 8 and figure 9 indicate there is no serious problem on this model fitting and no outliers exist in the model, so after taking log transformation on the amount of damages requested `demanded`, model 1.1 is better than model 1.

Model Inference and Results Table 4 summarizes the estimated coefficients of model 1.1: $log\frac{p}{1-p} = 0.4353 - 0.028 log(demanded) + 0.0383 tridays - 0.8937 bodinj + 0.0926 decorp1 - 0.5226 degovt1 + 0.1817 year - 0.4064 claimtype2 - 0.1.293 claimtype3 - 0.8831 claimtype4 - 0.4924 claimtype5 - 0.4151 claimtype6 + 0.1692 totalnopl + 0.0941 totalnode$. The estimated coefficients of days the trial lasted, whether a bodily injury was included, whether the defendant was the government, the year the civil lawsuit was filed and claimtype are statistically significant as their P-values are less than $\alpha = 0.05$. When the days the trial lasted one day longer, the odds of probability that the damages will be awarded is expected to increase by 3.9% ($= e^{0.0383} - 1$). The odds of the probability that the damages will be awarded is expected to decrease by 59% ($= e^{-0.8937} - 1$) once the bodily injury is included in the claim. From the dataset, 764 cases which didn't include bodily injury had damages awarded, while 411 cases which included bodily injury had received damages awards. The odds of the probability that the damages will be awarded is expected to decrease by 41% ($= e^{-0.5226} - 1$) if the defendants was the government. The dataset shows that 1131 cases whose defendants were not government had damages awarded, while only 44 cases whose defendants were government had damages awarded. The odds of probability of damages awarded is expected to increase by 19.9% ($= e^{0.1817} - 1$) if the trial is more recent. Besides, the coefficients of the number of plaintiff and whether the defendant is corporation are not statistically significant in this model. As shown in the dataset, the cases where defendants were not corporation but received damage awards were more than the cases where defendants were corporation and received damage awards. Also, the cases which had 2 plaintiffs and received damage awards were more than the cases which had 3 or more plaintiffs and received damage awards. Fiture 10 suggests that the log of damages requested may have smaller effect on the probability of damages awarded if there is a bodily injury, as shown in green dots, even though log of damages requested alone doesn't significantly affect the proability of damages awarded.

To sum up, a more recent trial or longer lasting trial are predictive of a higher probability that damages are awarded, which are consistent with initial assumption, while the factors of whether a corporation is the defendant or an increased number of plaintiffs don't have strong indication on the higher probability that

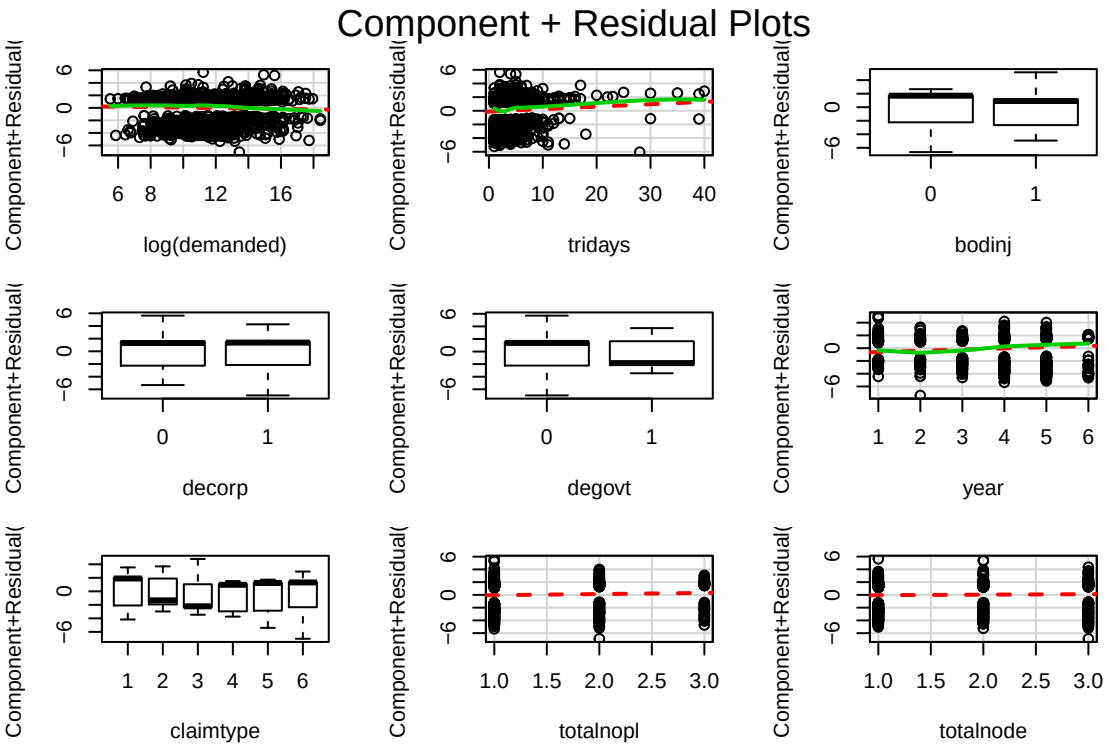


Figure 8: Component+Residual plots of model 1.1

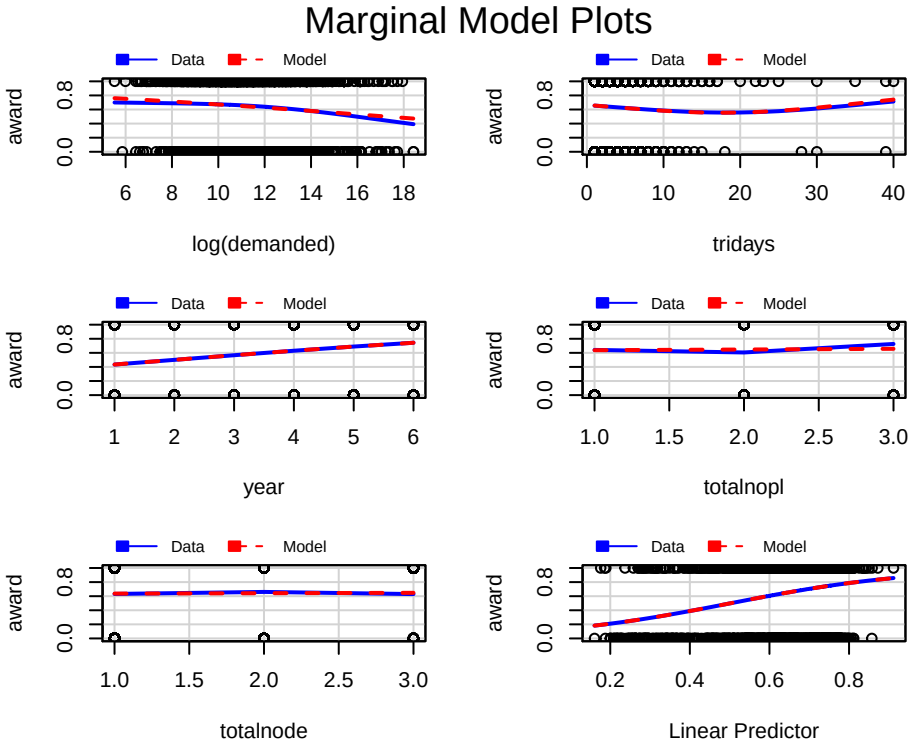


Figure 9: Marginal Model Plots of model 1.1

damages are awarded as assumed by the advocacy watchdog group. The inclusion of bodily injury claim reduces the effect of log of damages requested on the probability of damages awarded, which also disagrees with intial hypothesis.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.4353	0.4499	0.97	0.3333
log(demanded)	-0.0280	0.0276	-1.01	0.3102
tridays	0.0383	0.0180	2.12	0.0338
bodinj1	-0.8937	0.2190	-4.08	0.0000
decorp1	0.0926	0.1172	0.79	0.4294
degovt1	-0.5226	0.2339	-2.23	0.0254
year	0.1817	0.0468	3.89	0.0001
claimtype2	-0.4064	0.2051	-1.98	0.0475
claimtype3	-1.2930	0.2508	-5.16	0.0000
claimtype4	-0.8831	0.3036	-2.91	0.0036
claimtype5	-0.4924	0.3323	-1.48	0.1384
claimtype6	-0.4151	0.2292	-1.81	0.0701
totalnopl	0.1692	0.0973	1.74	0.0821
totalnode	0.0941	0.0766	1.23	0.2193

Table 4: Summary of model 1.1 predicting whether any damages are awarded

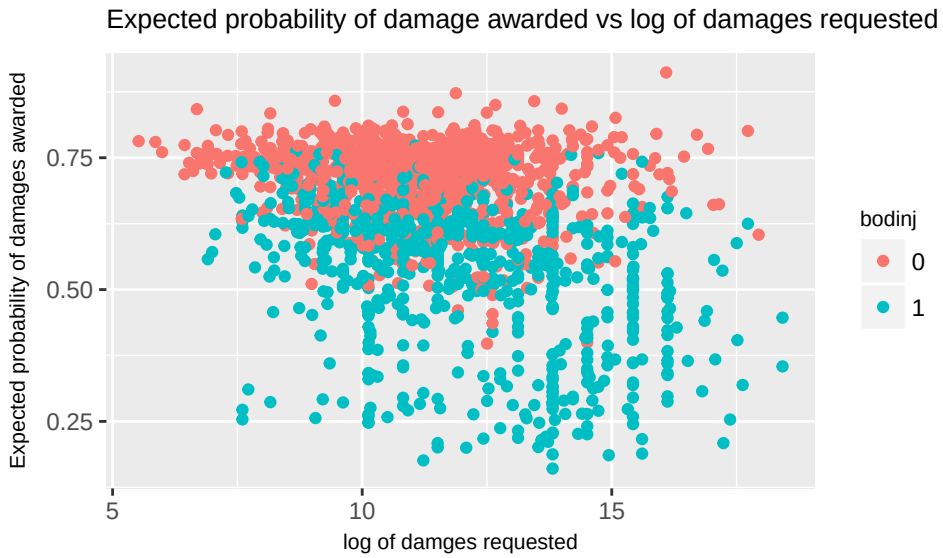


Figure 10: Expected probability of damages awarded vs log(demanded) with effect of bodily injury claim

Part 2: Regression model of predicting how much the damages is awarded

Initial Modeling Given that there was an award, we fitted a multiple linear regression model of predicting how much is awarded. In the intial model, the amount of damages awarded `totdam` is response variable, such that model 2: $totdam = \beta_0 + \beta_1 demanded + \beta_2 tridays + \beta_3 bodinj + \beta_4 decorp + \beta_5 degovt + \beta_6 year + \beta_7 claimtype + \beta_8 totalnopl + \beta_9 totalnode + \epsilon$. Table 5 shows the summary of model 2. The estimated coefficients of damages requested by plaintiffs, the days trial lasted, the inclusion of bodily injury claim, and claimtypes are statistically significant.

Diagnostics of Initial Model Figure 11 shows the diagnositc plots of model 2, as we can see that, there is a pattern in residuals plot and points 207, 226 992 are potential influential points. The points on the two

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	498416.4061	421635.5901	1.18	0.2374
demanded	0.2905	0.0166	17.54	0.0000
tridays	154630.2556	18523.9242	8.35	0.0000
bodinj1	-875449.7502	271529.8788	-3.22	0.0013
decorp1	9881.2903	128025.5703	0.08	0.9385
degovt1	-541090.5457	316782.4331	-1.71	0.0879
year	-15653.7852	55354.6652	-0.28	0.7774
claimtype2	-564138.4624	267487.2407	-2.11	0.0352
claimtype3	-166749.7656	361511.3829	-0.46	0.6447
claimtype4	-614893.0721	364861.6983	-1.69	0.0922
claimtype5	-782978.6443	374389.9143	-2.09	0.0367
claimtype6	-778490.0884	286440.0516	-2.72	0.0067
totalnopl	72252.3087	104533.4440	0.69	0.4896
totalnode	12.2094	85454.7198	0.00	0.9999

Table 5: Summary of model 2, predicting how much the damages is awarded

tails deviate from the normality assumption line in residuals QQplot, so model 2 violates residuals normality assumption. Also, the scale-location plot suggests that the variance are not equal in model 2. Marginal model plots and component residual plot in figure 12 and 13 further suggest that model 2 doesn't fit the data very well, it fails to have strong predictability on how much the damages are awarded.

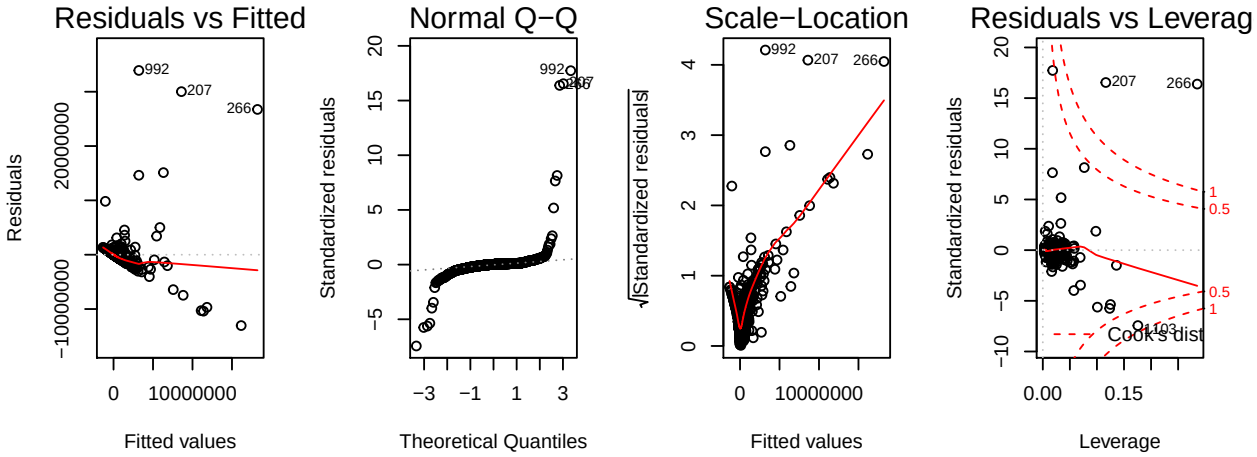


Figure 11: a) Residuals vs Fitted value, b) Residuals QQplot, c) Scale-Location plot, d) Residuals vs Leverage plot

Transformations and Diagnostics In order to fix the problems from model 2, we fitted the next model by log transforming the predictor the damages requested, such that model 2.1: $totdam = \beta_0 + \beta_1 log(demanded) + \beta_2 tridays + \beta_3 bodinj + \beta_4 decorp + \beta_5 degovt + \beta_6 year + \beta_7 claimtype + \beta_8 totalnopl + \beta_9 totalnode + \epsilon$. The diagnostic plots in figure 14 shows that the problems are not solved, the outliers still exist, residuals are not independent from fitted values and variances are still unequal. So, model 2.1 doesn't fit data as well.

According to figure 4b, the log of damages requested and the log of damages awarded are positively correlated, so we fitted the next model by transforming both reponse variable and the amount of damages requested, such that model 2.2: $log(totdam) = \beta_0 + \beta_1 log(demanded) + \beta_2 tridays + \beta_3 bodinj + \beta_4 decorp + \beta_5 degovt + \beta_6 year + \beta_7 claimtype + \beta_8 totalnopl + \beta_9 totalnode + \epsilon$. The diagnositc plots in figure 15 shows that model 2.2 is better than previous two models as residuals are normally distributed and variances are equal, and no outliers exist in this model. Nevertheless, figure 15e and 15f indicates that there are some lined up pattern in residuals vs log(demanded) plot and expected value vs log(demanded) plot.

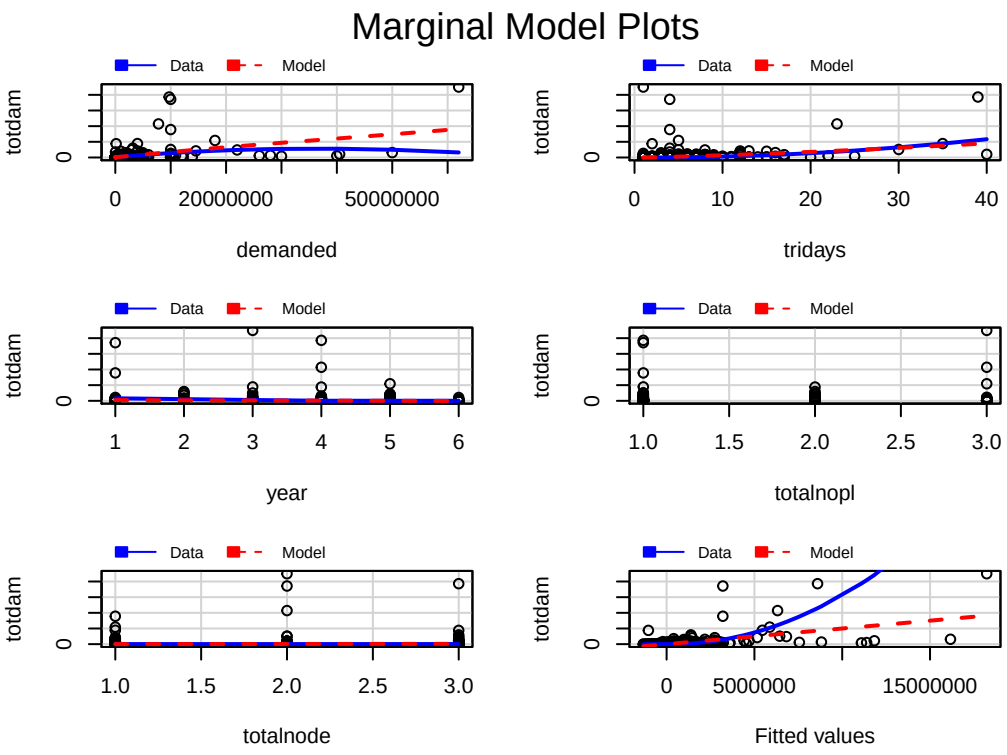


Figure 12: Marginal Model plots of model 2

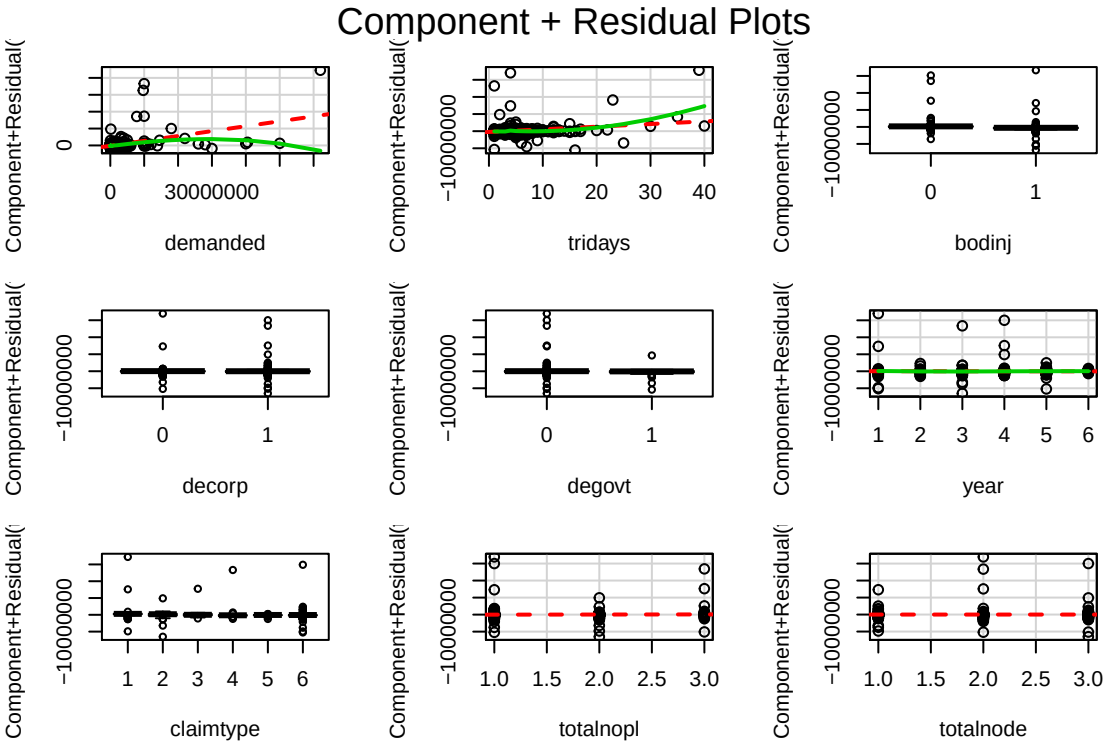


Figure 13: Component+Residual Plots of model 2

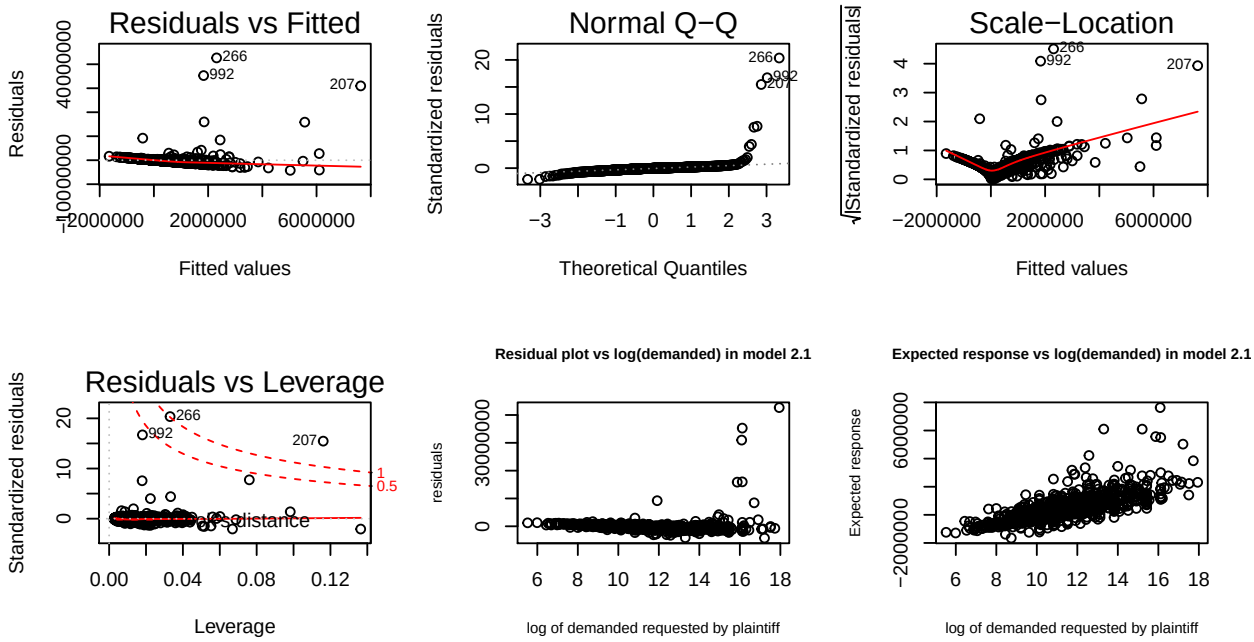


Figure 14: Model 2.1: a) Residuals vs Fitted value, b) Residuals QQplot, c) Scale-Location plot, d) Residuals vs Leverage plot, e) Residuals vs log(demanded), f) Expected response vs log(demanded)

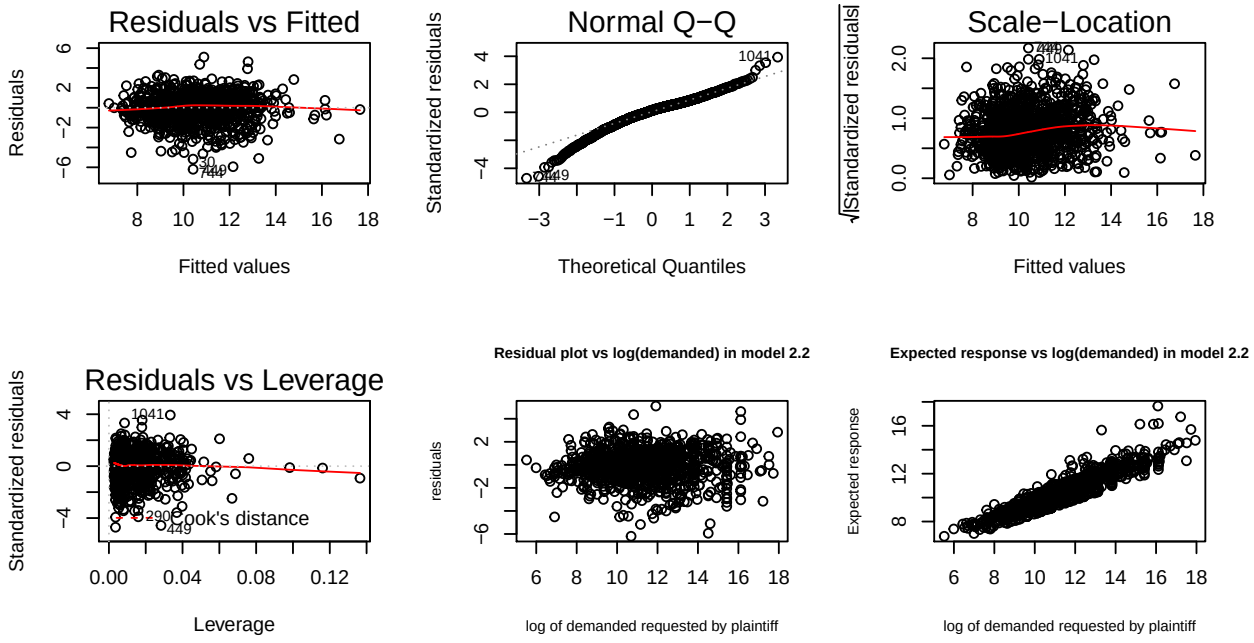


Figure 15: Model 2.2: a) Residuals vs Fitted value, b) Residuals QQplot, c) Scale-Location plot, d) Residuals vs Leverage plot, e) Residuals vs log(demanded), f) Expected response vs log(demanded)

Fiture 16 indicates that the inclusion of bodily injury claim may affect the relationship between the log of damages awarded and log of damages requested, and all the lined up points represent the cases when the bodily injury was included in the claim.

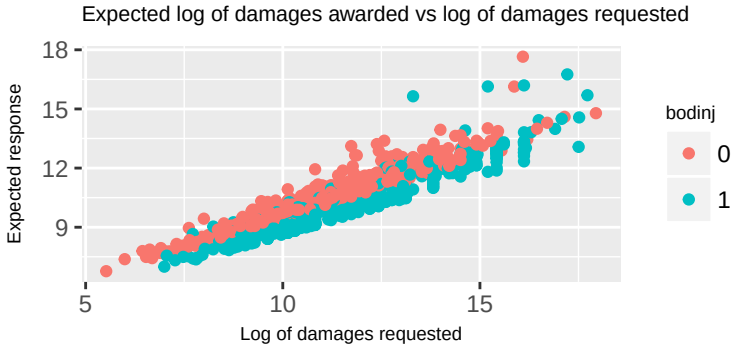


Figure 16: Expected response vs log(demanded) with effect of whether bodily injury claim was included (1:"Yes", 0:"No")

Then, we fitted the next model by including the interaction of the log of demanded requested and whether the bodily injury was included in the claim, model 2.3: $log(totdam) = \beta_0 + \beta_1 log(demanded) + \beta_2 tridays + \beta_3 bodinj + \beta_4 decorp + \beta_5 degovt + \beta_6 year + \beta_7 claimtype + \beta_8 totalnopl + \beta_9 totalnode + \beta_1 0 log(demanded) : bodinj + \epsilon$. The diagnostic plots of mdoel 2.3 shows that model 2.3 fulfills the linear model assumptions, residuals are independent from covariates and no outliers or influential points exist, except that the residuals QQplot has some points deviating from lower left tail. Marginal model plots and add-variable plots futher support that mdoel 2.3 fit the data very well.

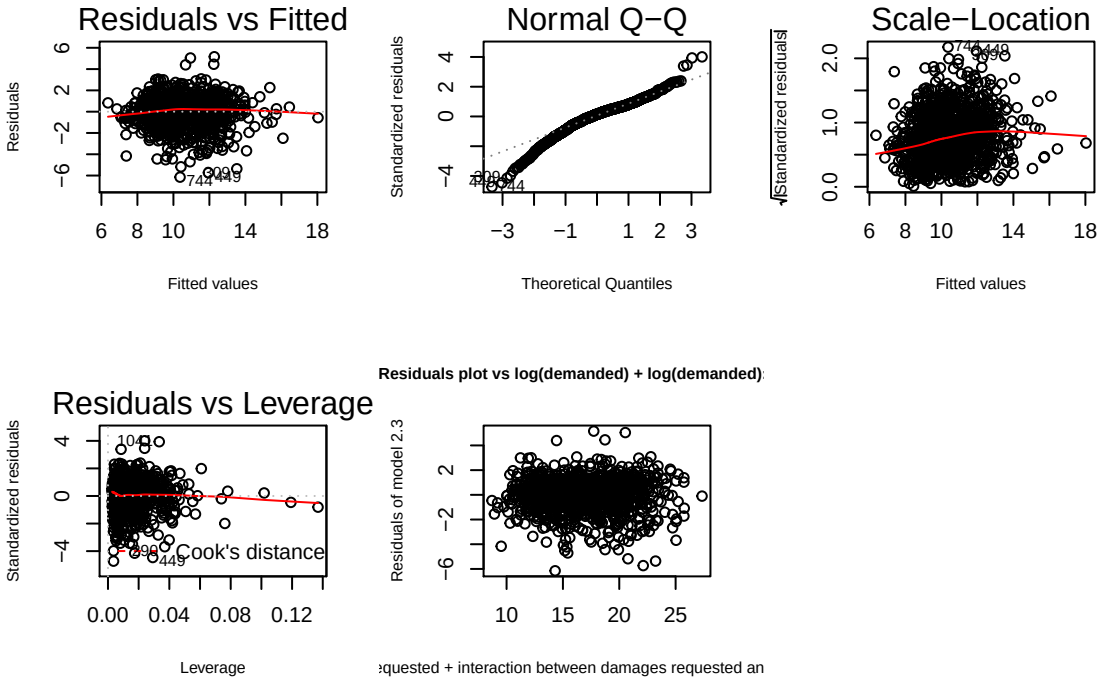


Figure 17: Model 2.3: a) Residuals vs Fitted value, b) Residuals QQplot, c) Scale-Location plot, d) Residuals vs Leverage plot, e) Residuals plot vs log(damanded) with effect of bodily injury effect

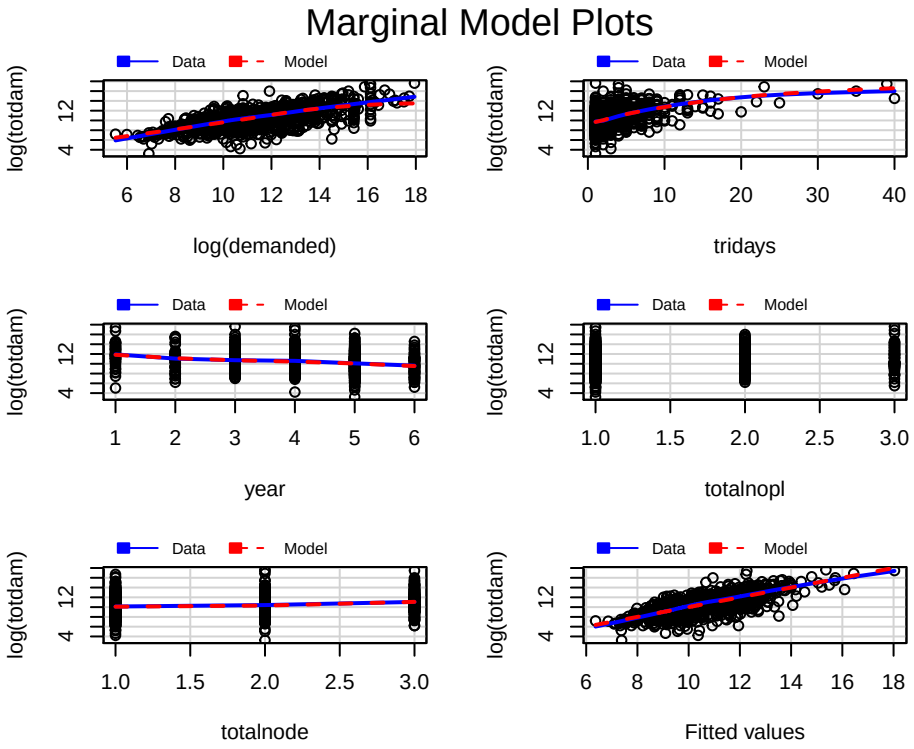
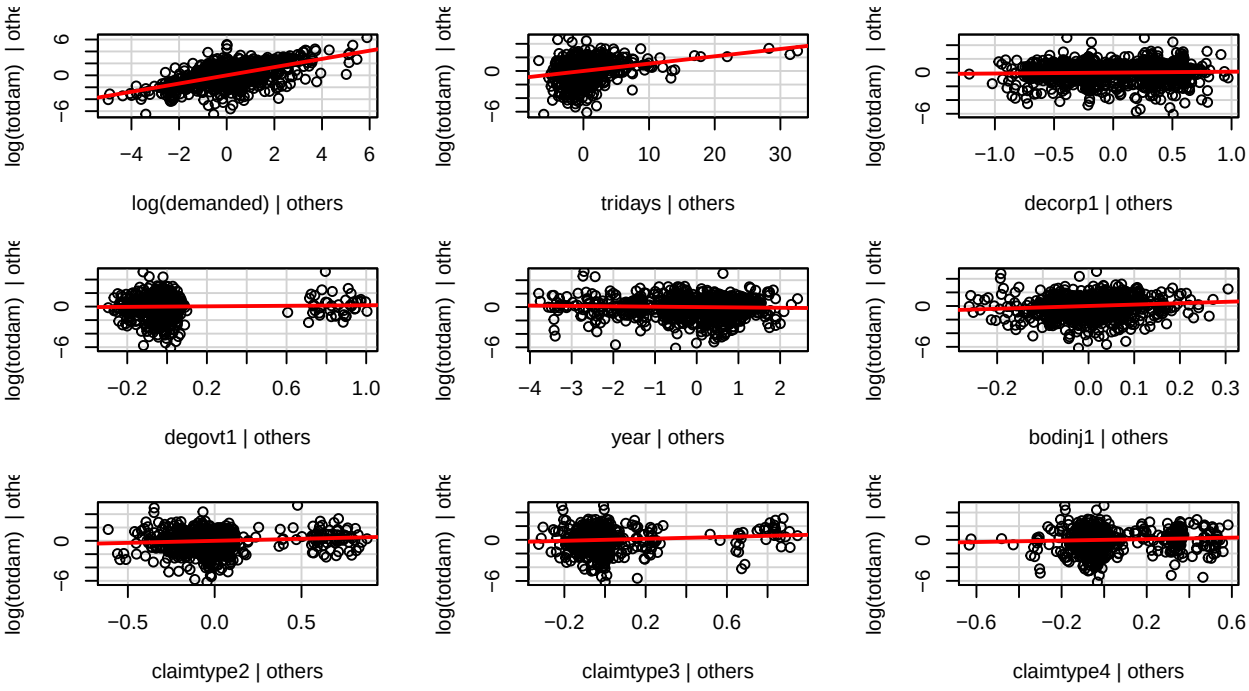
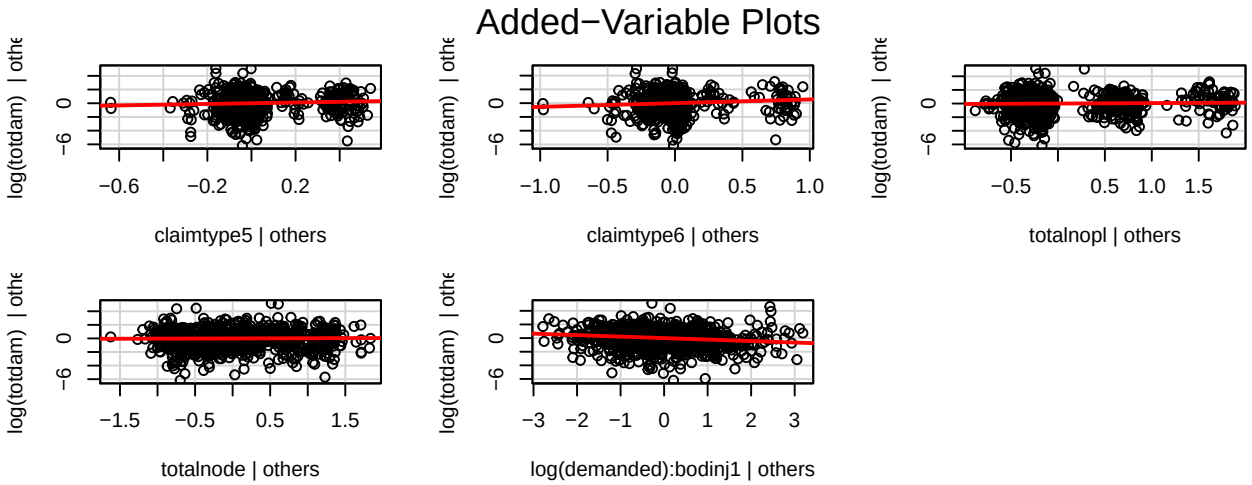


Figure 18: Marginal Model Plots of Model 2.3





	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.1290	0.4361	4.88	0.0000
log(demanded)	0.6890	0.0291	23.68	0.0000
tridays	0.1081	0.0128	8.44	0.0000
decorp1	0.1374	0.0881	1.56	0.1192
degovt1	0.1971	0.2138	0.92	0.3566
year	-0.0546	0.0376	-1.45	0.1463
bodinj1	2.0391	0.5236	3.89	0.0001
claimtype2	0.6094	0.1853	3.29	0.0010
claimtype3	0.7273	0.2476	2.94	0.0034
claimtype4	0.5355	0.2483	2.16	0.0313
claimtype5	0.5292	0.2583	2.05	0.0407
claimtype6	0.5380	0.1969	2.73	0.0064
totalnopl	0.0644	0.0704	0.91	0.3607
totalnode	0.0283	0.0577	0.49	0.6233
log(demanded):bodinj1	-0.2172	0.0412	-5.28	0.0000

Table 6: Summary of model 2.3, predicting how much the damages is awarded

Model Inference and Results Table 6 summarizes the estimated coefficients of model 2.3: $log(totdam) = 2.129 + 0.689log(demanded) + 0.1081tridays + 2.0391bodinj1 + 0.1374decorp1 + 0.1971degovt1 - 0.0546year + 0.6094claimtype2 + 0.7273claimtype3 + 0.5355claimtype4 + 0.5292claimtype5 + 0.5380claimtype6 + 0.0644totalnopl + 0.0283totalnode - 0.2172log(demanded) \times bodinj$. The intercept, the log of damages requested by plaintiffs, the days that trial lasted, the inclusion of bodily injury claim, the types of claim, and the interaction between the log of damages requested and whether the bodily injury was included are statistically significant. If a bodily injury was not part of the claim, the total amount of damages awarded to plaintiff is expected to increase by 6.79% given 10% increase in total amount of damages requested by plaintiff. If a bodily injury was included as part of the claim, the toal amount of damages awarded is expected to increase by 4.6% given 10 % increase in total amount of damages requested by plaintiff. Figure 16 supports this indication, as we can see that given the same log of damages requested by plaintiff, the expected value of log of damages awarded to plaintiff when bodily injury was included (shown in blue dots) is lower than those when bodily injury wasn't included (shown in red dots). For the days the trial lasted, for one day increase in the days the trial lasted, the total amount of damages awarded is expected to increase by 11.42%. For the type of claim, motor vehicle claim as reference, the amount of damages awarded is expected to increase by 84% when shifting from motor vehicle claim type to premises liability type, increase by 107% when shifting to malpractice type, increase by 70.8% when shifting to fraud type, increase by 69.76% when shifting to rental/lease type, increased by 71.3% when shifting to other claims. The number of plaintiffs, whether the defendant was corporation, and the year the lawsuit was filed are not

statistically significant in this model.

Therefore, we concluded that the increase of damages requested by plaintiff and longer lasting trial days and malpractice as claim type are predictive of higher damages awarded, which agree with our intial hypothesis. However, the amount of damages requested may have smaller effect on the amount of damages awarded when bodily injury is part of the claim, which controdicts to our initial hypothesis. Whether the defendant is corporation and increased number of plaintiffs are not as predictive of higher damages awarded as hypothesized initially.

Appendix

```
library(knitr)  # We need the knitr package to set chunk options
library(arm)
library(car)
library(pander)
library(grid)
library(xtable)
library(ggplot2)
options(scipen = 1, digits = 2)
options(scipen = 999)
########### Data Cleaning and EDA ##########
damage <- read.csv("damages.csv", header = TRUE)
names(damage) <- tolower(names(damage))
# change the following integer to factor
damage$bodinj <- as.factor(damage$bodinj)
damage$decorp <- as.factor(damage$decorp)
damage$degovt <- as.factor(damage$degovt)
damage$claimtype <- as.factor(damage$claimtype)
# change the following integer to continuous
# variable
damage$totdam <- as.numeric(damage$totdam)
damage$demanded <- as.numeric(damage$demanded)
# Create a dummy variable `award` indicating
# whether the damages is awarded (1 = Yes, 0 = No),
# if total amount of damages is 0, `award` = 0,
# otherwise `award` = 1.
damage$award <- as.factor(ifelse(damage$totdam == 0,
    0, 1))

yearcheck <- sapply(split(damage$totdam, damage$year),
    mean)
yearcheck1 <- sapply(split(damage$demanded, damage$year),
    mean)
par(mfrow = c(4, 2))
barplot(yearcheck, xlab = "Year the civil lawsuit was filed",
    ylab = "Average damages awarded", main = "Bar plot of average damages awarded vs year",
    cex.main = 0.9, cex.lab = 1)
barplot(yearcheck1, xlab = "Year the civil lawsuit was filed",
    ylab = "Average damages requested", main = "Bar plot of average damages requested vs year",
    cex.main = 0.9, cex.lab = 1)
damage$year <- as.numeric(damage$year)
# As the number of plaintiffs increases, the
# average amount of damages awarded and requested
# are also increased, and the number of plaintiffs
# is descrete, we should keep `totalnopl` as
# integer variable.
noplcheck <- sapply(split(damage$totdam, damage$totalnopl),
    mean)
barplot(noplcheck, xlab = "Total number of plaintiffs",
    ylab = "Average damages awarded", main = "Bar plot of average damages awarded vs the number of plai
    cex.main = 0.9, cex.lab = 1)
noplcheck1 <- sapply(split(damage$demanded, damage$totalnopl),
```

```
    mean)
barplot(noplcheck1, xlab = "Total number of plaintiffs",
    ylab = "Average damages requested", main = "Bar plot of average damages requested vs the number of p
    cex.main = 0.9, cex.lab = 1)
# The average anmount of damages awarded and
# requested increases as the number of defendatns
# increases, so we keep variable `totalnode` as
# integer variable
nodecheck <- sapply(split(damage$totdam, damage$totalnode),
    mean)
barplot(nodecheck, xlab = "Total number of defendants",
    ylab = "Average damages awarded", main = "Bar plot of average damages awarded vs the number of defe
    cex.main = 0.9, cex.lab = 1)
nodecheck1 <- sapply(split(damage$demanded, damage$totalnode),
    mean)
barplot(nodecheck1, xlab = "Total number of defendants",
    ylab = "Average damages requested", main = "Bar plot of average damages requested vs the number of d
    cex.main = 0.9, cex.lab = 1)
damage$totalnode <- as.integer(damage$totalnode)
damage$totalnopl <- as.integer(damage$totalnopl)
# There were 1076 out of 1836 trials lasted 2 or
# less days, the distribution of days the trial
# lasted `tridays` is heavily right skewed, we
# should convert `tridays` as continuous variable.
hist(damage$tridays, breaks = 20, ylim = c(0, 1300),
    labels = TRUE, main = "Histogram of days the trial lasted",
    xlab = "Number of days trial lasted", cex.main = 0.8,
    cex.lab = 1)
damage$tridays <- as.numeric(damage$tridays)

# 36% of cases didn't receive damage award
# length(damage$totdam[which(damage$totdam ==
# 0)])/length(damage$totdam)
damaward <- damage[which(damage$totdam != 0), ]  #dataset for cases that damages were awarded

attach(damage)
## Among cases that received damage awards: 411
## cases which included bodily injury vs 764 cases
## which didn't include body injury.
table(award, bodinj)

# 627 cases where the defendants were corporation
# had received damage awards Vs 548 where
# defendants were not corporation.
table(award, decorp)

# 267 cases are motor vehicle claims, 76 cases are
# premises liability claims, 36 are malpractice, 81
# are fraud, 67 are rental/lease, and 648 are
# others
table(award, claimtype)

# 926 cases which had 1 plaintiff received damage
```

```r
# awards, 181 cases which had 2 plaintiffs received
# damage awards, 68 cases which had 3 or more
# plaintiffs received damage awards.
table(award, totalnopl)

# 642 cases had one defendant, 363 cases had two
# defendants, 170 cases had 3 or more defendants
table(award, totalnode)

demandsummary <- summary(damage$demanded)
pandoc.table(demandsummary, style = "rmarkdown", "Summary of total amount of damages requrested (in $)"
    split.table = Inf)

totdamsummary <- summary(damage$totdam)
pandoc.table(totdamsummary, style = "rmarkdown", "Summary of total amount of damages awarded to plainti
    split.table = Inf)

par(mfrow = c(1, 2))
hist(damage$demanded, label = TRUE, ylim = c(0, 2000),
    breaks = 6, xlab = "Total amount of damages requested",
    main = "Histogram of damages requested", cex.main = 0.8,
    cex.lab = 0.8)
hist(damage$totdam, label = TRUE, ylim = c(0, 2000),
    breaks = 6, xlab = "Total amount of damages awarded",
    main = "Histogram of damages awarded", cex.main = 0.8,
    cex.lab = "0.8")

par(mfrow = c(1, 2))
plot(totdam ~ demanded, main = "Damages awards received vs damages requested",
    xlab = "demages requested", ylab = "damages awarded",
    cex.main = 0.7, cex.lab = 0.9)
plot(log(totdam) ~ log(demanded), main = "Log of damages awards received vs log of damages requested",
    xlab = "Log of damages requested", ylab = "log of damages awarded",
    cex.main = 0.7, cex.lab = 0.9)
detach(damage)

####### Model Selection: Part 1 #########
model1 <- glm(award ~ demanded + tridays + bodinj +
    decorp + degovt + year + claimtype + totalnopl +
    totalnode, data = damage, family = "binomial")
options(xtable.comment = FALSE)
xtable(summary(model1), caption = "Summary of model 1, predicting whether any damages are awarded")

par(mfrow = c(1, 2))
binnedplot(predict(model1), resid(model1), xlab = "Fitted values in model 1",
    ylab = "Binned Residual", main = "Binned residuals plot of model 1",
    cex.main = 0.8, cex.lab = 0.8)
binnedplot(damage$demanded, resid(model1), xlab = "The amount of damages requested",
    ylab = "Binned Residual", main = "Binned residuals plot of model 1",
    cex.main = 0.8, cex.lab = 0.8)
crPlots(model1)  # variable demanded doesn't have linear relationship with response

model1.1 <- glm(award ~ log(demanded) + tridays + bodinj +
```

```
    decorp + degovt + year + claimtype + totalnopl +
    totalnode, data = damage, family = "binomial")
par(mfrow = c(1, 2))
binnedplot(predict(model1.1), resid(model1.1), xlab = "Fitted value in model 1.1",
    ylab = "Binned Residual", main = "Binned residuals plot of model 1.1",
    cex.main = 0.8, cex.lab = 0.8)
binnedplot(log(damage$demanded), resid(model1.1), xlab = "log of damages requested in model 1.1",
    ylab = "Binned Residual", main = "Binned residuals plot of model 1.1",
    cex.main = 0.8, cex.lab = 0.8)
crPlots(model1.1)
mmps(model1.1)

options(xtable.comment = FALSE)
xtable(summary(model1.1), caption = "Summary of model 1.1 predicting whether any damages are awarded")

ggplot(damage, aes(x = log(demanded), y = predict(model1.1,
    type = "response"), fill = bodinj, colour = bodinj)) +
    geom_point() + labs(title = "Expected probability of damage awarded vs log of damages requested",
    x = "log of damges requested", y = "Expected probability of damages awarded") +
    theme(title = element_text(size = 8))

############## Mode selection: Part 2 ###############
options(xtable.comment = FALSE)
model2 <- lm(totdam ~ demanded + tridays + bodinj +
    decorp + degovt + year + claimtype + totalnopl +
    totalnode, data = damaward, comment = NA, message = NA)
xtable(summary(model2), caption = "Summary of model 2, predicting how much the damages is awarded")

par(mfrow = c(1, 4))
plot(model2, cex.main = 0.8)
mmps(model2)
crPlots(model2)

# after transformation on demanded: still not good,
# the residuals are not independent from
# log(demanded), and outliers exist, and unequal
# variance
model2.1 <- lm(totdam ~ log(demanded) + tridays + decorp +
    degovt + year + bodinj + claimtype + totalnopl +
    totalnode, data = damaward)
par(mfrow = c(2, 3))
plot(model2.1)  #influential points/outliers: pt 992, 266, 207
plot(log(damaward$demanded), resid(model2.1), xlab = "log of demanded requested by plaintiff",
    ylab = "residuals", main = "Residual plot vs log(demanded) in model 2.1",
    cex.main = 0.8, cex.lab = 0.8)
plot(log(damaward$demanded), predict(model2.1), xlab = "log of demanded requested by plaintiff",
    ylab = "Expected response", main = "Expected response vs log(demanded) in model 2.1",
    cex.main = 0.8, cex.lab = 0.8)  #some lined up points exist

# transformation on both totdam and demanded, looks
# better now. but there are some lined up pattern
# in residual vs log(demanded) plot and fitted
# value vs log(demanded) plot.
```

```r
model2.2 <- lm(log(totdam) ~ log(demanded) + tridays +
    decorp + degovt + year + bodinj + claimtype + totalnopl +
    totalnode, data = damaward)
par(mfrow = c(2, 3))
plot(model2.2)  #less influential points/outliers: 744, 449
plot(log(damaward$demanded), resid(model2.2), xlab = "log of demanded requested by plaintiff",
    ylab = "residuals", main = "Residual plot vs log(demanded) in model 2.2",
    cex.main = 0.8, cex.lab = 0.8)  #some lined up points exist
plot(log(damaward$demanded), predict(model2.2), xlab = "log of demanded requested by plaintiff",
    ylab = "Expected response", main = "Expected response vs log(demanded) in model 2.2",
    cex.main = 0.8, cex.lab = 0.8)  #some lined up points exist

ggplot(damaward, aes(x = log(demanded), y = predict(model2.2),
    fill = bodinj, colour = bodinj)) + geom_point() +
    labs(x = "Log of damages requested", y = "Expected response") +
    ggtitle("Expected log of damages awarded vs log of damages requested") +
    coord_fixed(ratio = 0.5) + theme(title = element_text(size = 7))

# add interaction term log(demanded):bodinj, see
# the effect of log(demanded) on response when
# there is a bodily injury
model2.3 <- lm(log(totdam) ~ log(demanded) + tridays +
    decorp + degovt + year + bodinj + claimtype + totalnopl +
    totalnode + log(demanded):bodinj, data = damaward)
par(mfrow = c(2, 3))
plot(model2.3, cex.main = 0.8, cex.lab = 0.8)
plot(log(damaward$demanded) + log(damaward$demanded):damaward$bodinj,
    resid(model2.3), xlab = "log of damages requested + interaction between damages requested and if bo
    ylab = "Residuals of model 2.3", main = "Residuals plot vs log(demanded) + log(demanded):bodinj",
    cex.main = 0.8, cex.lab = 0.8)
mmps(model2.3)
avPlots(model2.3)

options(xtable.comment = FALSE)
xtable(summary(model2.3), caption = "Summary of model 2.3, predicting how much the damages is awarded")
```