Predicting Plaintiff Payout in Civil Litigation

Andrew Resnikoff

10/14/2017

Introduction

Background

A litigation advocacy watchdog group is interested in determining which features of a civil court case are predictive of the size of total damage awards. Since the damages awarded vary greatly between cases, the public has been concerned with why this occurs. In this analysis, we will explore various characteristics of civil court cases and see what their effect on the damages awarded to the plaintiff is. Based on the group's hypothesis, we expect that a more recent trial, a corporation as the defendant, and an increased number of plaintiffs will increase the amount of damages awarded. In addition, it is thought that if there is a bodily injury claim then the relationship between total damages and amount demanded is altered, as the amount demanded should have more of an effect on how much money is ultimately paid out.

Data

We are given a dataset with 1836 observations of the following variables representing a civil lawsuit: the total amount of damages in dollars awarded to a plaintiff, the total amount of damages in dollars requested by the plaintiff, the number of days the trial lasted, an indicator variable that represents whether bodily injury was part of the claim, an indicator variable that represents whether the defendant was a corporation, an indicator variable that represents whether the defendent was the government, the year the lawsuit was filed, the type of claim (motor vehicle, premise liability, malpractice, fraud, rental/lease or other), the number of plaintiffs and the number of defendants[1]. Note that there were no missing values.

Since we are interested in finding the relationship between the amount of damages awarded and the other variables, we will split the dataset based on whether or not the plaintiff was awarded any damages. In the dataset of cases where the plaintiff won, we have 1175 observations. This analysis will focus on predicting how much a plaintiff was paid if they were paid at all.

Methods

Exploratory Data Analysis

In Table 1 below, summary statistics are shown for each continuous variable. From the histograms in Figure 1, we can get a better look at distributions of each variable. Based on the descriptive statistics and histograms, it is clear that both the total damages awarded and the total awards will need to be transformed as the data is highly skewed right[2]. From now on, award amounts will be transformed onto the log scale. Trial days is also skewed right.

In Table 2 and Table 3 below, summary statistics are shown for each categorical variable. Notice that about a third of the trials involve bodily injury and about half involve a corporation. Only about 5% of cases involved the government. Of the different types of claims, motor vehicle related claims were the most of any specific claim, making up about a fifth of the total. 'Other' claims made up about 50% of the cases. This suggests that it may be necessary to recode this variable so that it provides more information. This will be decided after performing bivariate exploratory data analysis. Most of the cases are from the years 2000 (47%)

[1] Variable names and descriptions can be found in Section 1 of the Appendix

[2] extreme values have been omitted in the displayed plots to show the shape

and 1999 (30%), there are about an equal number of cases (~7-8%) each from 1997 and 2001 and also about equal number of cases (~4%) each from 1997 and before 1997. Since we are interested in whether or not more recent trials are predictive of higher damages awards, we can consider recoding the year variable as well. Again, this will need to be confirmed by bivariate exploratory data analysis.

Table 1: Descriptive Statistics for Continuous Variables

	Min	1Q	Median	Mean	3Q	Max	SD
Damages Awarded ($)	25	9243	31000	332982	100226	44968563	2288631
Damages Demanded ($)	250	20000	55516	775727	250000	62000000	3589728
Trial Days	1	1	2	3.004	4	40	3.37

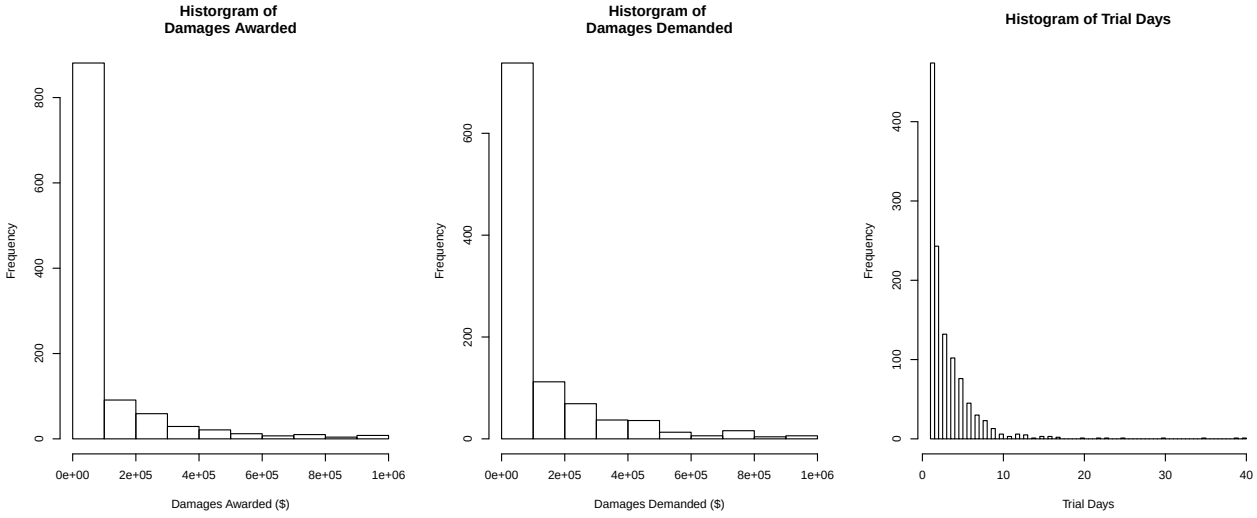
Table 2: Summary Statistics for Categorical Variables

Year	Percent	Claim Type	Percent	No. Plaintiffs	Percent	No. Defendants	Perecent
< 1997	3.74%	Motor Vehicle	22.72%	1	78.8%	1	54.64%
1997	3.57%	Premises Liability	6.47%	2	15.40%	2	30.89%
1998	8.68%	Malpractice	3.06%	3+	5.79%	3+	14.47%
1999	29.79%	Fraud	6.89%				
2000	47.49%	Rental/Lease	5.70%				
2001	6.72%	Others	55.15%				

Table 3: Descriptive Statistics for Binary Variables

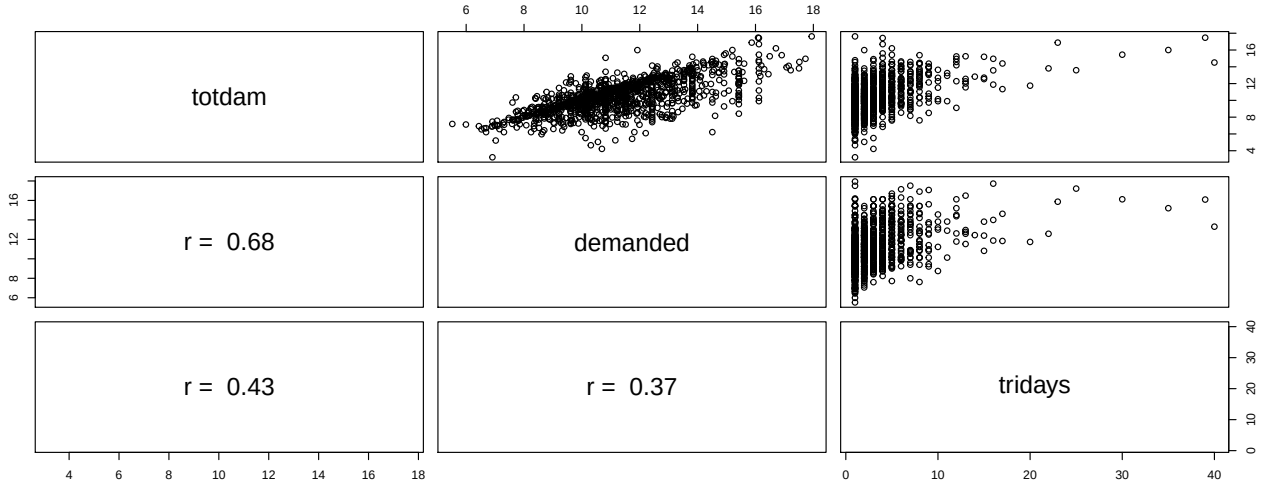
Bodily Injury	Percentage	Corporation Defendant	Percentage	Government Defendant	Percentage
Yes	35%	Yes	53%	Yes	4%
No	65%	No	47%	No	96%

Figure 1: Histograms for Continuous Variables



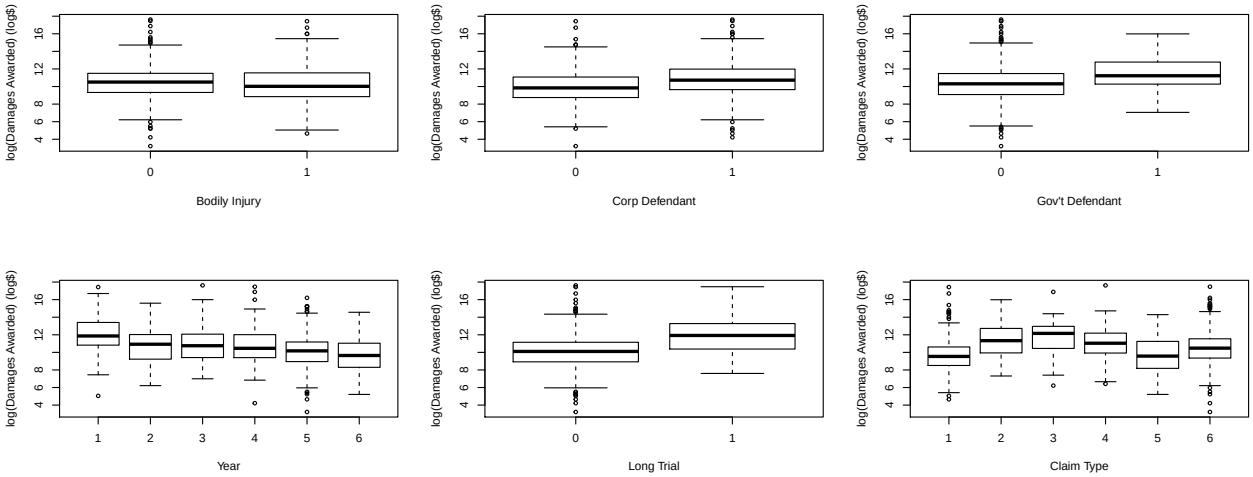
From the pairs plot in Figure 2, we can see that the damages demanded and damages awarded have a fairly strong positive linear relationship. The skew of the number of trial days makes it difficult to see any sort of relationship between either the damages demanded or the damages awarded. The correlation between them is positive and it does appear that there is an increase in payout when the number of days increases, but the relationship is definitely non linear.

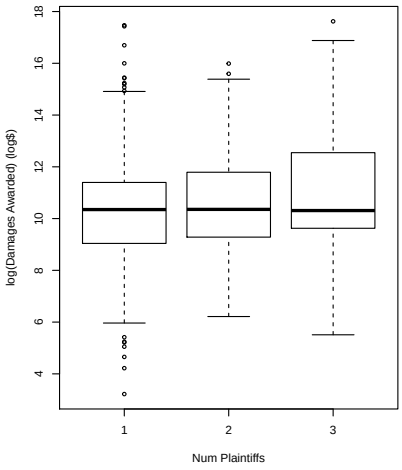
Figure 2: Pairs Plot of Continous Variables

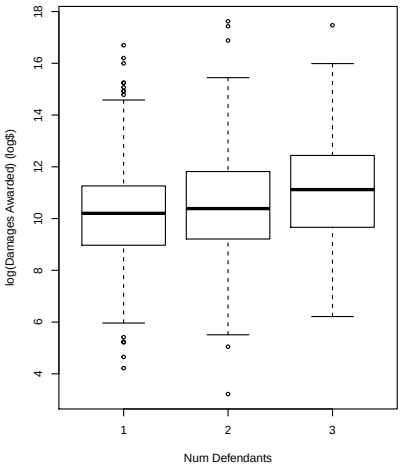


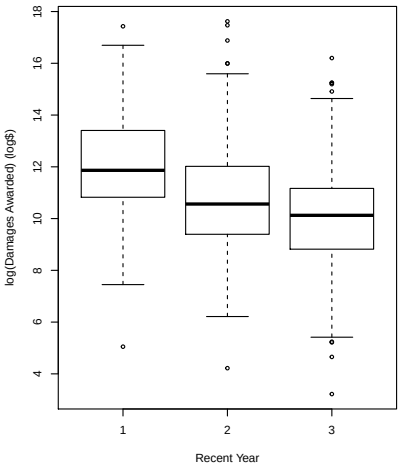
Based on the boxplots in Figure 3, it seems as though the distribution of log(damages awarded) doesn't change based on whether or not bodily injury is involved as the middle 50% of values overlap almost completely. However, for cases where the defendant is either a corporation or the government, we see a slight increase that suggests those may have an effect on the total damages awarded. When looking at the boxplot for the year the lawsuit was filed, notice that there is a lot of overlap of the distributions over the years 1997 - 1999. It seems reasonable to condense these variables into three categories: recent, not recent and old. The distribution of log(damages awarded) seems to increase as the number of plaintiffs and defendants increase. It also varies based on the claim type, so it is not appropriate to recode any of these categorical variables.

Figure 3: Box Plots of Damages Awarded by Categorical Variables.







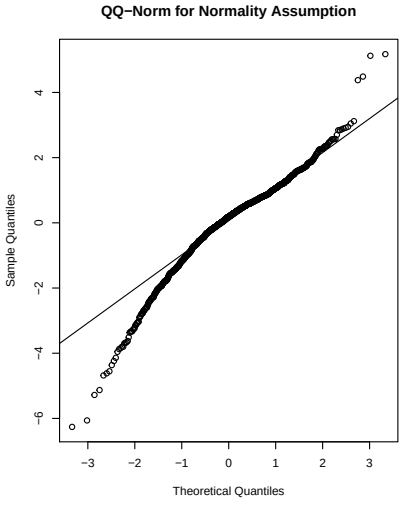


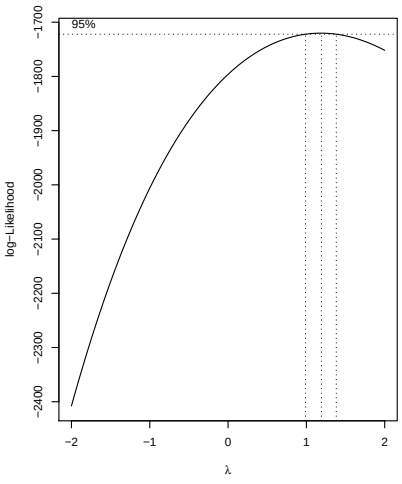
Models

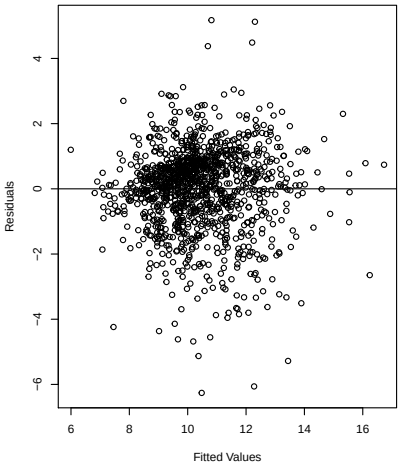
Our initial model will predict log(totdam), the log of the total damages awarded from the following predictor variables: the log of the damages demanded, whether or not bodily injury was involved, whether the defendant is the government, whether the defendant is a corporation, whether the case was filed recently, whether the trial was long and the numbers of plaintiffs and defendants.

Notice that the amount demanded not only affects how much is awarded (intuitively and as shown in the pairs plot), but it will also affect other variables such as how long a trial lasts. Thus, it makes sense to include an interaction term to capture this effect. Another example of this is the claim type - it makes sense that different types of claims will be worth different amounts of money which means plaintiffs will demand different amounts. An interaction is included for this effect as well. The last interaction term included is one that was hypothesized by the litigation watchdog group - that involving a bodily injury changes the amount that is demanded.

A QQ-Norm plot is used to check the assumption that the residuals are approximately normal. The normality assumption appears to potentially be violated, but since we have many observations we can assume it does for the sake of this analyis. A boxcox plot is used to test for transformations. While the plot does seem to suggest that a transformation may be necessary, since the response was already transformed and the spread of the suggested transformation contains 1, a transformation will not be performed for the sake of interpretability. A plot of the model residuals against the fitted values does not show any pattern, but there do appear to be some values with large residuals that may be outliers.

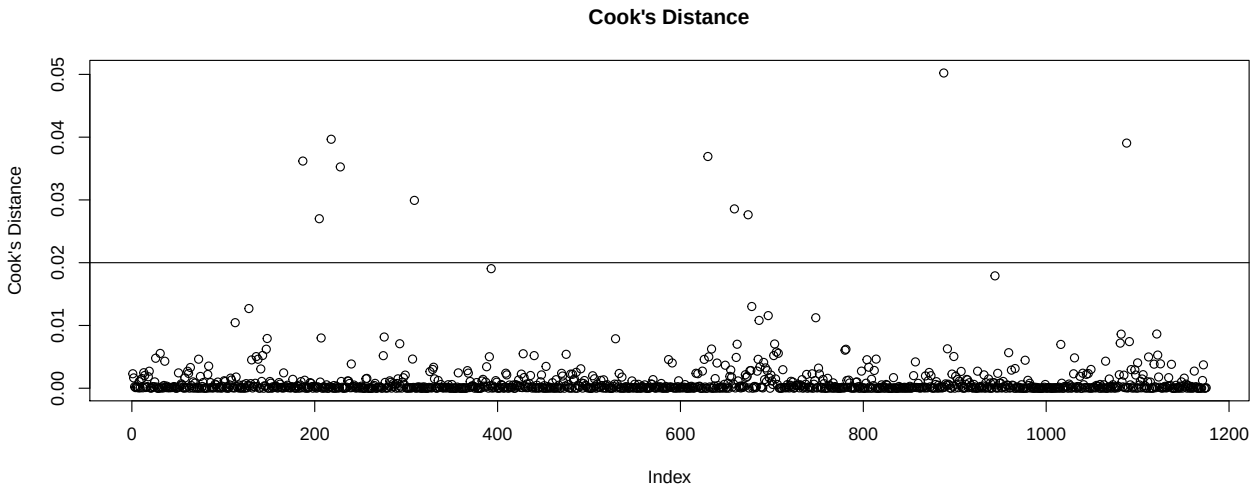






We can use cook's distance to look for leverage points and see if any of the supposed outliers have a large effect on our model. From the plot of cook's distance in Figure 4, we see that there are a few points well above the rest. We can treat these as potential leverage points. By finding the intersection of the potential outliers and the potential leverage points, we have found 7 high leverage potential outliers that are influencing our model. After inspecting these points (and other points), it appears that 3 of them are cases where the plaintiff recieved much more money in damages than they asked for in a very short trial. For the other 4 cases, the amount of money awarded was very little compared to the amount demanded. Since we wouldn't usually expect a lot more money than demanded be awarded very quickly, it seems reasonable to exclude these 3 points from the model. For the low award cases, however, it does seem reasonable that some cases would barely award anything even if a large amount of money is demanded, so these cases will be left in the model. Before moving on, it is necessary to recheck that the assumptions still hold.

Figure 4: Plot of Cook's Distance



Since the model assumptions have been met, we can now test for the appropriateness of our final model. Looking at the p-values for the coefficients in the model, it seems as though the number of defendants, whether or not the trial was recent and whether or not the government was a defendant may not be significant predictors of log(damages awarded). We test this idea using a partial F test.

The test results in an F statistic of 1.68 and a p-value of .13. We therefore fail to reject the null hypothesis, that any of the removed coefficients are non zero, at the $\alpha = .05$ level of significance. Thus, we can remove these from the model.A table of the summary of our final model is shown below in Table 4.

Results

Table 4: Final Model Summary

Variable	Estimate (Std. Error)	95% CI
(Intercept)	1.94 (1.08)	(-.18, 4.05)
log(demanded)	0.69 (0.09)	(.52, .86)
factor(bodinj)1	1.86 (1.04)	(-.19, 3.91)
factor(decorp)1	0.13 (0.08)	(-.03, .3)
tridays	0.22 (0.07)	(.08, .36)
factor(claimtype)2	0.05 (0.93)	(-1.77, 1.87)
factor(claimtype)3	-1.49 (1.75)	(-4.93, 1.95)
factor(claimtype)4	0.99 (1.47)	(-1.89, 3.87)
factor(claimtype)5	-0.64 (1.38)	(-3.36, 2.07)
factor(claimtype)6	0.42 (1.07)	(-1.69, 2.52)
factor(totalnopl)2	-0.08 (0.11)	(-.3, .13)
factor(totalnopl)3	0.33 (0.17)	(.01, .66)
log(demanded):tridays	-0.01 (0.01)	(-.02, .003)
log(demanded):factor(claimtype)2	0.05 (0.07)	(-.09, .2)
log(demanded):factor(claimtype)3	0.18 (0.14)	(-.09, .45)
log(demanded):factor(claimtype)4	-0.03 (0.12)	(-.27, .21)
log(demanded):factor(claimtype)5	0.12 (0.12)	(-.12, .36)
log(demanded):factor(claimtype)6	0.01 (0.09)	(-.16, .18)
log(demanded):factor(bodinj)1	-0.20 (0.08)	(-.36, -.03)

From this model, we can conclude that the amount demanded by the plaintiff, whether or not the case involved a bodily injury, whether or not the defendant is a corporation, the number of trial days, the type of claim and the total number of plaintiffs are significant predictors of the amount of damages awarded. We found that the group's hypothesis was correct in that having a corporation as a defendant corresponds to an increase in damages awarded. Their hypothesis was only partially correct in terms of the number of plaintiffs, having three or more plaintiffs does correspond to an increase in damages awarded. Since the trial being more recent was not a significant predictor (and thus not included in the final model), their hypothesis was incorrect that more recent trials correspond to larger payouts.

Appendix

Variables Descriptions

TOTDAM: total amount of damages awarded to plaintiff (in $)

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 - categorized as follows:
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

R Code

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(cache = TRUE)
knitr::opts_chunk$set(fig.width=12, fig.height=5)
if (!require(MASS)){
    install.packages("MASS")
    library(MASS)
}

data <- read.csv("justice.csv")
# make column names lower case
colnames(data) <- sapply(colnames(data), tolower)

data$won <- rep(0, nrow(data))
data$won[data$totdam > 0] = 1

# create dataframe from subset of data who won their cases
winners <- subset(data, won == 1)
```

```
summary(winners$totdam); sd(winners$totdam)
summary(winners$demanded); sd(winners$demanded)
summary(winners$tridays); sd(winners$tridays)

prop.table(table(winners$bodinj))
prop.table(table(winners$decorp))
prop.table(table(winners$degovt))
prop.table(table(winners$year))
prop.table(table(winners$claimtype))
prop.table(table(winners$totalnopl))
prop.table(table(winners$totalnode))


par(mfrow = c(1,3))
hist(winners$totdam[winners$totdam < 1000000], breaks = 10,
     main = "Historgram of \nDamages Awarded", xlab = "Damages Awarded ($)")
hist(winners$demanded[winners$demanded < 1000000], breaks = 10,
     main = "Historgram of \nDamages Demanded", xlab = "Damages Demanded ($)")
hist(winners$tridays, breaks = 100, main = "Histogram of Trial Days", xlab= "Trial Days")
par(mfrow = c(1,1))

panel.cor <- function(x, y, digits=2, cex.cor)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- cor(x, y)
  txt <- format(c(r, 0.123456789), digits=digits)[1]
  text(0.5, 0.5, paste("r = ",txt), cex=2)
}

pairs.vars <- with(winners, cbind(totdam = log(totdam), demanded = log(demanded), tridays = tridays))
names(pairs.vars) <- c("Damages Awarded", "Damages Demanded", "Trial Days")
pairs(pairs.vars, lower.panel = panel.cor)

winners$longtrial <- rep(0, nrow(winners))
winners$longtrial[winners$tridays > 4] = 1

par(mfrow = c(1,3))
boxplot(log(totdam) ~ bodinj, data = winners, xlab = "Bodily Injury", ylab = "log(Damages Awarded) (log$
boxplot(log(totdam) ~ decorp, data = winners, xlab = "Corp Defendant", ylab = "log(Damages Awarded) (log
boxplot(log(totdam) ~ degovt, data = winners, xlab = "Gov't Defendant", ylab = "log(Damages Awarded) (l
boxplot(log(totdam) ~ year, data = winners, xlab = "Year", ylab = "log(Damages Awarded) (log$)")
boxplot(log(totdam) ~ longtrial, data = winners, xlab = "Long Trial", ylab = "log(Damages Awarded) (log$
boxplot(log(totdam) ~ claimtype, data = winners, xlab = "Claim Type", ylab = "log(Damages Awarded) (log$
boxplot(log(totdam) ~ totalnopl, data = winners, xlab = "Num Plaintiffs", ylab = "log(Damages Awarded)
boxplot(log(totdam) ~ totalnode, data = winners, xlab = "Num Defendants", ylab = "log(Damages Awarded)

recent <- rep(0, nrow(winners))
recent[winners$year == 1] = 1
recent[winners$year > 1 & winners$year < 5] = 2
recent[winners$year >= 5] = 3
recent <- factor(recent)
winners$recent = recent
boxplot(log(totdam) ~ recent, data = winners, xlab = "Recent Year", ylab = "log(Damages Awarded) (log$)
```

```
par(mfrow = c(1,1))



model <- lm(log(totdam) ~ log(demanded) + factor(bodinj) + factor(degovt) + factor(recent) +
                factor(decorp) + tridays +  factor(claimtype) + factor(totalnopl) + factor(totalnode) +
                log(demanded)*tridays + log(demanded)*factor(claimtype) + log(demanded)*factor(bodinj),


par(mfrow = c(1,3))
qqnorm(model$res, main = "QQ-Norm for Normality Assumption"); qqline(model$res)
boxcox(model)

plot(model$res ~ model$fitted.values, xlab = "Fitted Values", ylab = "Residuals"); abline(0,0)

potential.outliers <- which(abs(model$res) > 7)

cooks.d <- cooks.distance(model)
plot(cooks.d, main = "Cook's Distance", ylab = "Cook's Distance"); abline(.02, 0)

potential.leverage <- which(cooks.d > .02)

influential.points <- intersect(potential.outliers, potential.leverage)

influential.rows <- winners[influential.points, ]

similar.points <- which(winners$totdam > 1.5*winners$demanded)

similar.rows <- winners[similar.points, ]

similar.points <- which(winners$totdam < .05*winners$demanded)

similar.rows <- winners[similar.points, ]


leverage.points <- c(992, 1041, 1062)
clean.winners <- winners[-leverage.points, ]
model.clean <- lm(log(totdam) ~ log(demanded) + factor(bodinj) + factor(degovt) + factor(recent) +
                factor(decorp) + tridays +  factor(claimtype) + factor(totalnopl) + factor(totalnode) +
                log(demanded)*tridays + log(demanded)*factor(claimtype) + log(demanded)*factor(bodinj),

par(mfrow = c(1,3))
qqnorm(model$res, main = "QQ-Norm for Normality Assumption"); qqline(model$res)
boxcox(model)

plot(model$res ~ model$fitted.values, xlab = "Fitted Values", ylab = "Residuals"); abline(0,0)

potential.outliers <- which(abs(model$res) > 7)

summary(model.clean)

model.reduced <- lm(log(totdam) ~ log(demanded) + factor(bodinj) +
                factor(decorp) + tridays + factor(claimtype) + factor(totalnopl) +
                log(demanded)*tridays + log(demanded)*factor(claimtype) + log(demanded)*factor(bodinj),
```

```
f.test <- anova(model.reduced, model.clean)
```