Predicting Civil Case Payouts to Plaintiff

Liam Gersten
lgersten

Saturday, March 24 2021

Contents

Introduction

Losing defendants of civil court cases are often ordered to compensate the plaintiff for damages, often in the form of monetary payment. However, the amount of payment tends to vary greatly, and can be based on a number of factors. One might expect that more severe damages may be corelated with higher payments. In some cases, payment could vary based on settlements made after drawn-out, lengthy trials. Claim types, amounts demanded by the plaintiff, or the number of days that a trial lasts may all influence the amount paid by varying degrees. We use the 2001 Civil Justice Survey of State Courts to predict amount paid to plaintiffs from the three factors mentioned

Exploratory Data Analysis

Data

The data obtained via the survey contains 126 entries and 4 variables, 3 of which can be used as quantitative predictors/estimators during analysis, with one categorical variable. Since we are interested in predicting paid amount, we explore and model the relationship between total amount of damages paid (in dollars) and four variables:

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

TRIDAYS: how many days the trial lasted

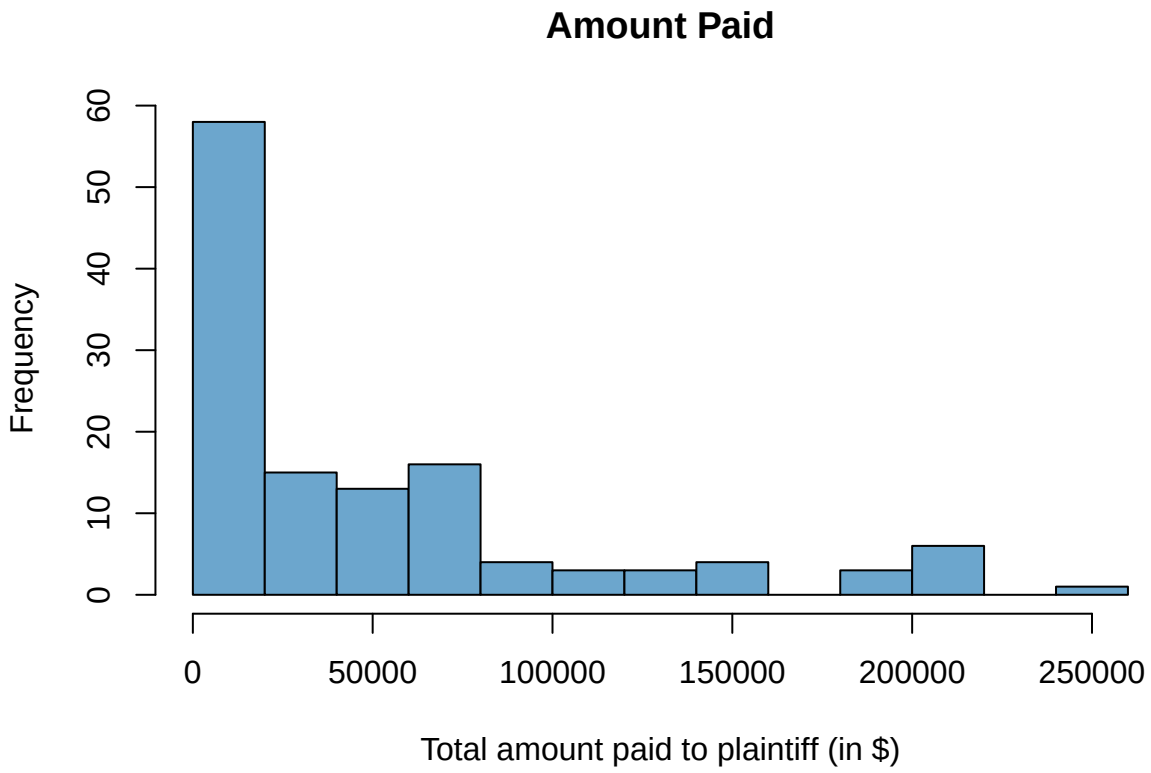
CLAIMTYPE: type of claim made by plaintiff categorized as: 1. motor vehicle 2. premises liability 3. malpractice 4. fraud 5. rental/lease 6. other

The header of the first few lines of the dataset is as follows:

```
## # A tibble: 6 x 4
##   TOTDAM DEMANDED TRIDAYS CLAIMTYPE
##    <dbl>    <dbl>   <dbl> <chr>
## 1  11760    17640       1 Rental
## 2 150000   200000       2 Other
## 3   2831     2870       1 Other
## 4  29863     9900       5 Motor
## 5   2200     2200       2 Other
## 6  70945    58816       2 Other
```

Univariate Exploratory Data Analysis

We begin by displaying histograms, boxplots and numerical summaries to individually explore the patterns observed for each variable.

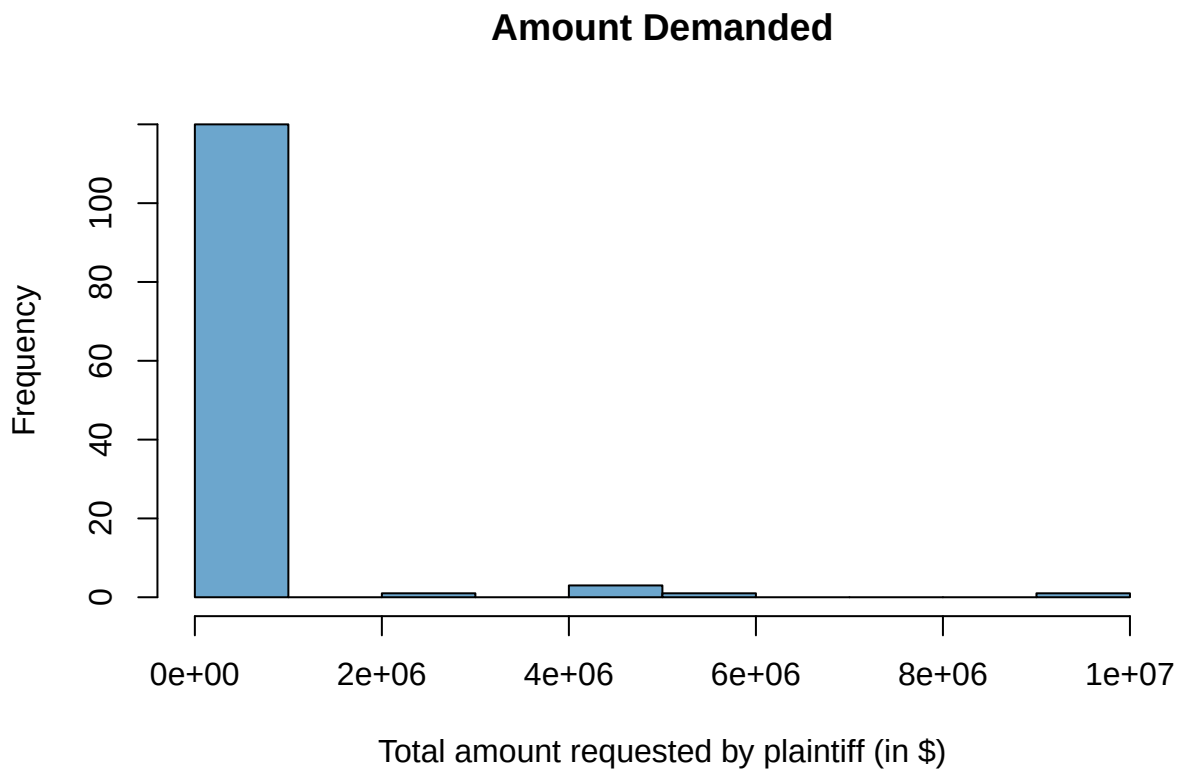


Summary

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     225    7544   26322   51279   70750  248280
```

Observations

The distribution of total amount paid is skewed to the right and unimodal with a single peak in the first quartile. There are known outliers towards the right extrema, which warrant some concern.

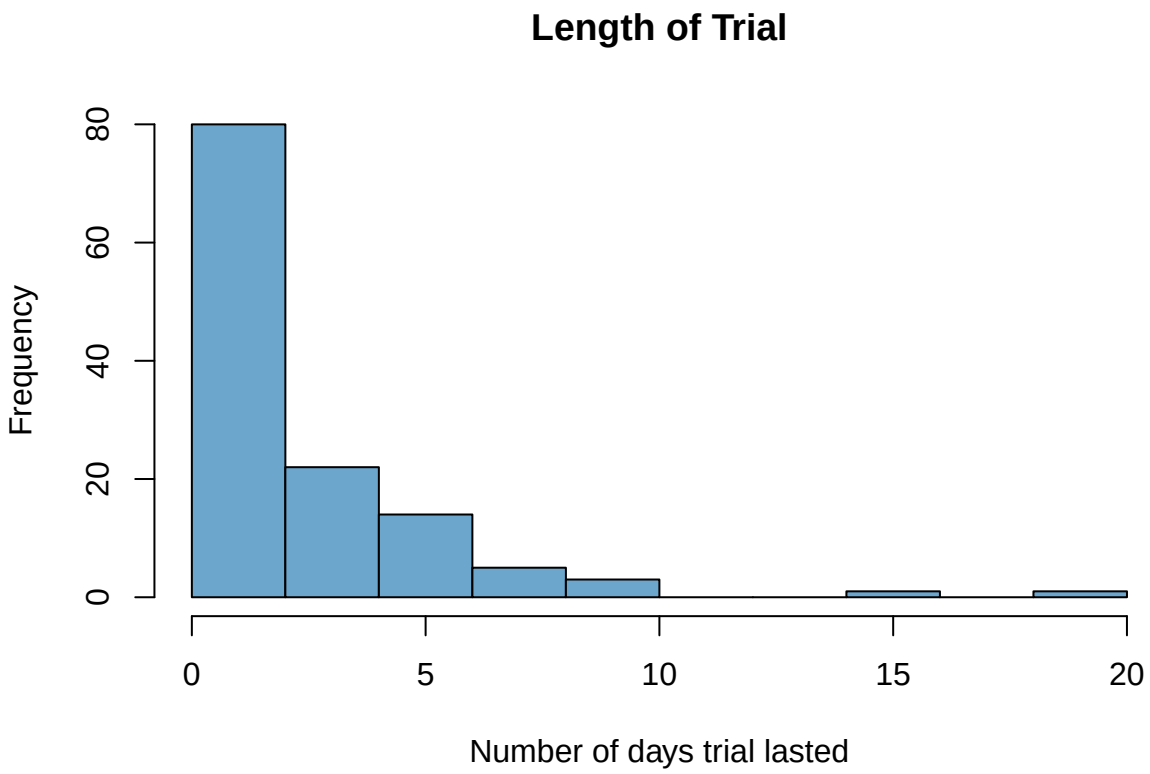


Summary

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     225    7544   26322   51279   70750  248280
```

Observations

The distribution of amount demanded is single-peaked and strongly skewed to the right. The vast majority of requests by the plaintiff do not exceed 100,000. In fact, without outliers, the mean is 36,537, so they alone pull the mean by almost 14,000. Transformations will be critical for modeling.

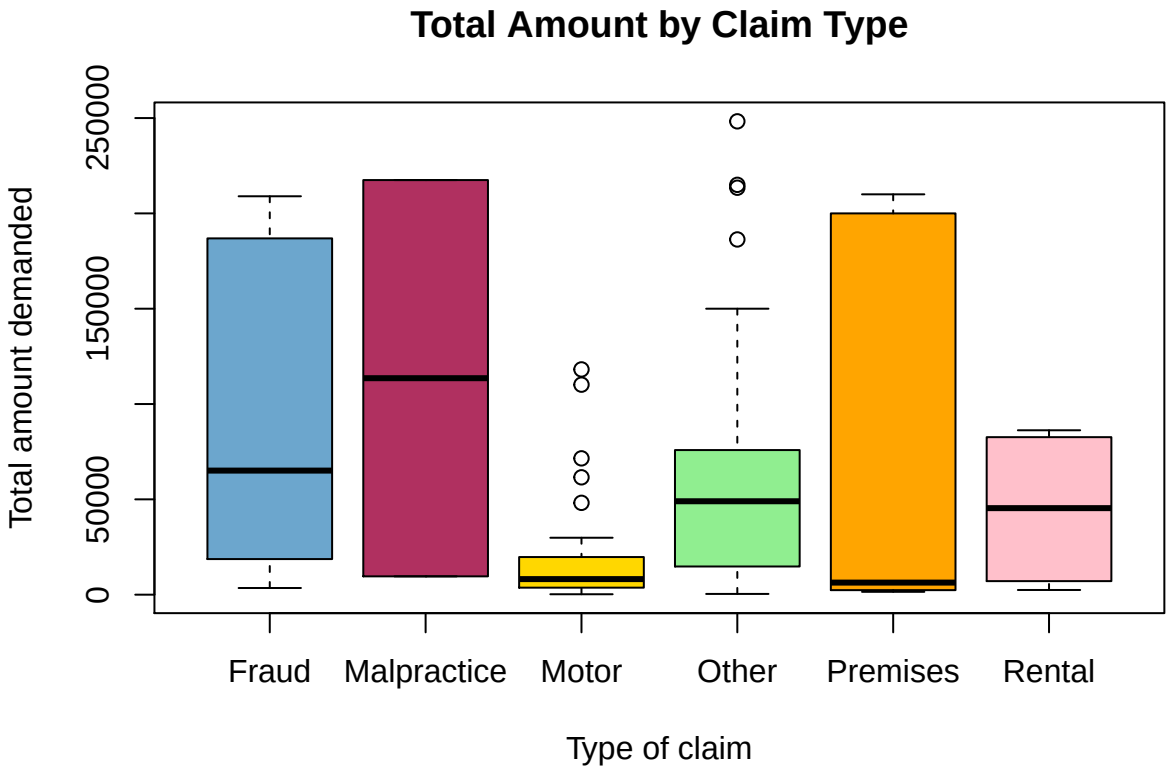


Summary

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   2.833   4.000  20.000
```

Observations

The distribution of trial lengths is single-peaked and skewed to the right with several outliers. While its skewness is more prevalent than that of the amount demanded (the predictor), it does not demonstrate the same severity as the skew seen for amount demanded. Transformations may be necessarry depending on the type of relationship it shows with the predictor.

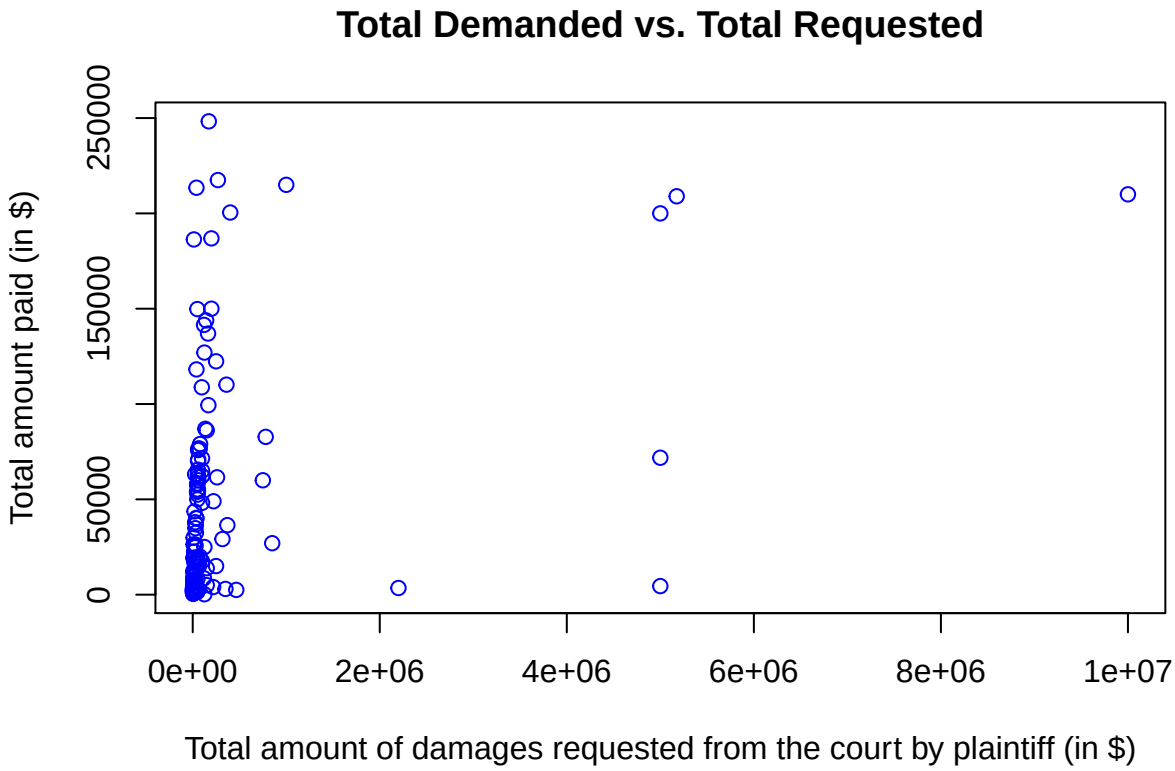


Observations

The distributions of total amount demanded vary greatly by category. Trials involving fraud, malpractice, or premises tend to yield greater payouts for damages. On the other hand, trials involving vehicles or other instances seem to yield smaller payments with some exceptions in the form of outliers.

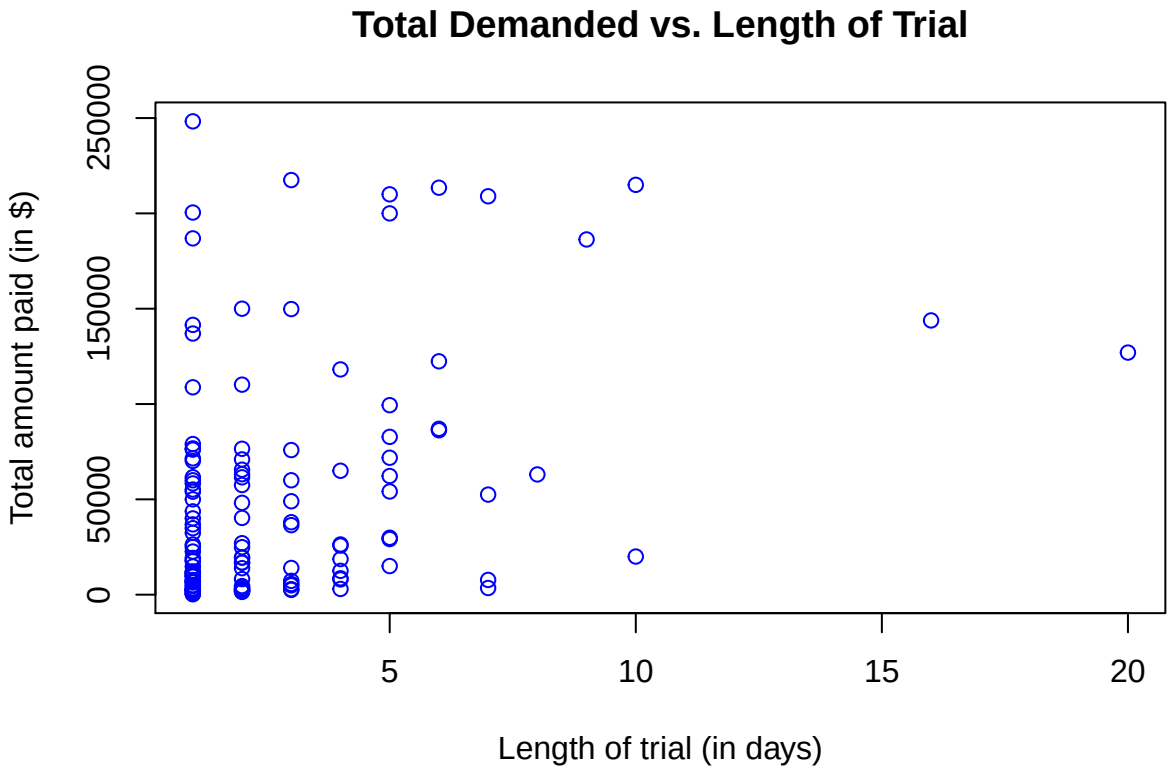
Bivariate Exploratory Data Analysis

Now, we will see and comment on two scatterplots representing the relationships between total amount demanded and its predictors.



Observations

The relationship between total demanded and total requested appears nonlinear. There does seem to be positive association between the two variables, although they will need to be explored further and likely transformed before satisfying the linearity requirements of a simple linear or multiple linear regression model.



Observations

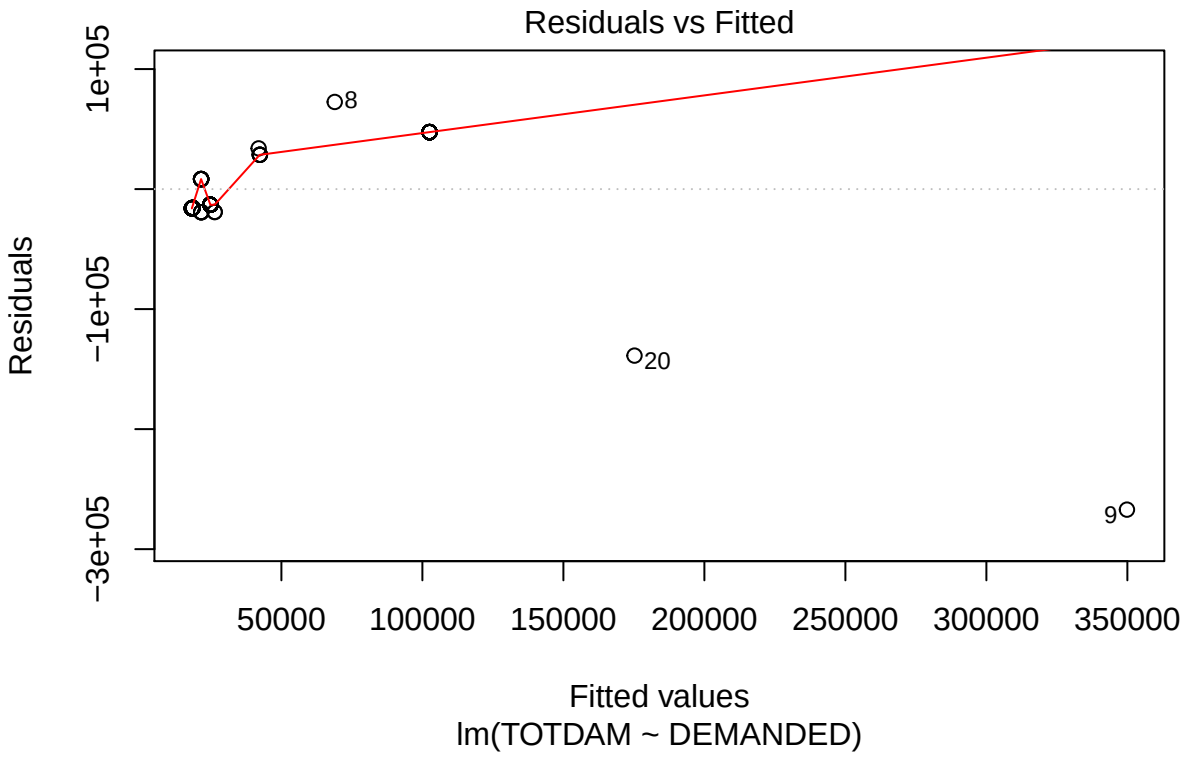
There appears to be a weak, positive linear association between amount paid and length of trial. In general, as trials become longer, the amount paid increases.

Modeling

All three quantitative predictors show signs of skew. The skew of amount paid may be acceptable, but the skew seen for both quantitative predictors is sufficiently concerning. Furthermore, the relationships between amount paid and its predictors are either nonlinear, or weakly linear. This may be due to the skew of said predictors, as the predictor with the stronger skew (total requested) shows a less linear form for its scatterplot. First, we will attempt transformations on our predictor variables. We will then need to revaluate our exploratory data analysis with updated values. We can then validate the necessary assumptions of a multiple linear regression model and check for multicollinearity.

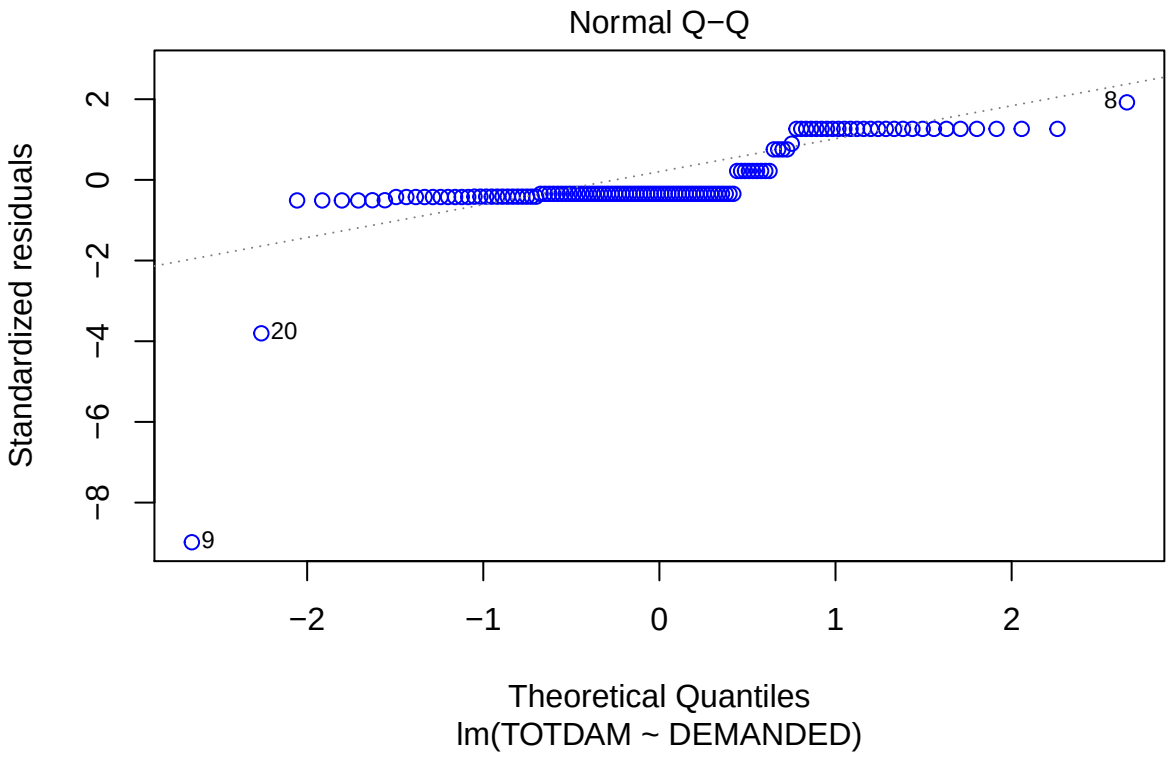
Diagnostics

We will begin by creating a temporary model involving all quantitative predictors. This model will be used to produce residual diagnostics and a Normal Q-Q plot for analysis.



Observations

The residuals show clear signs of pattern and are not constantly spread above and below the 0-line. As such, we cannot validate that the errors of the model have mean zero or have constant standard deviation.

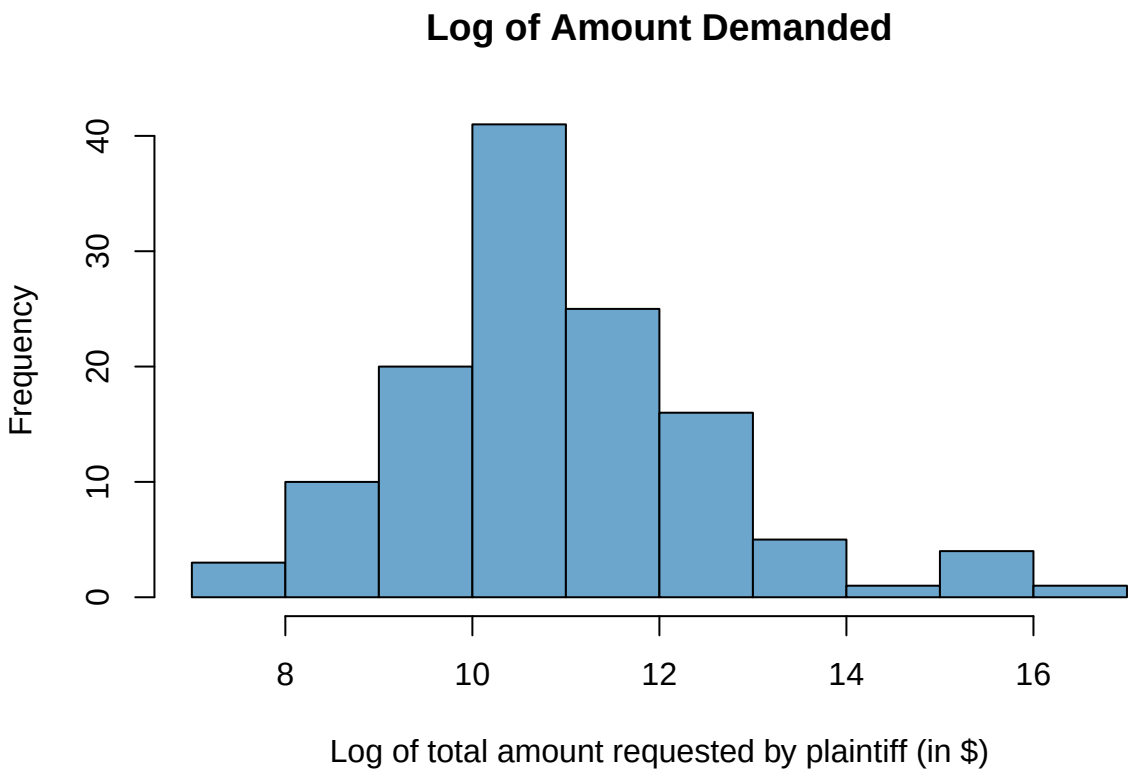


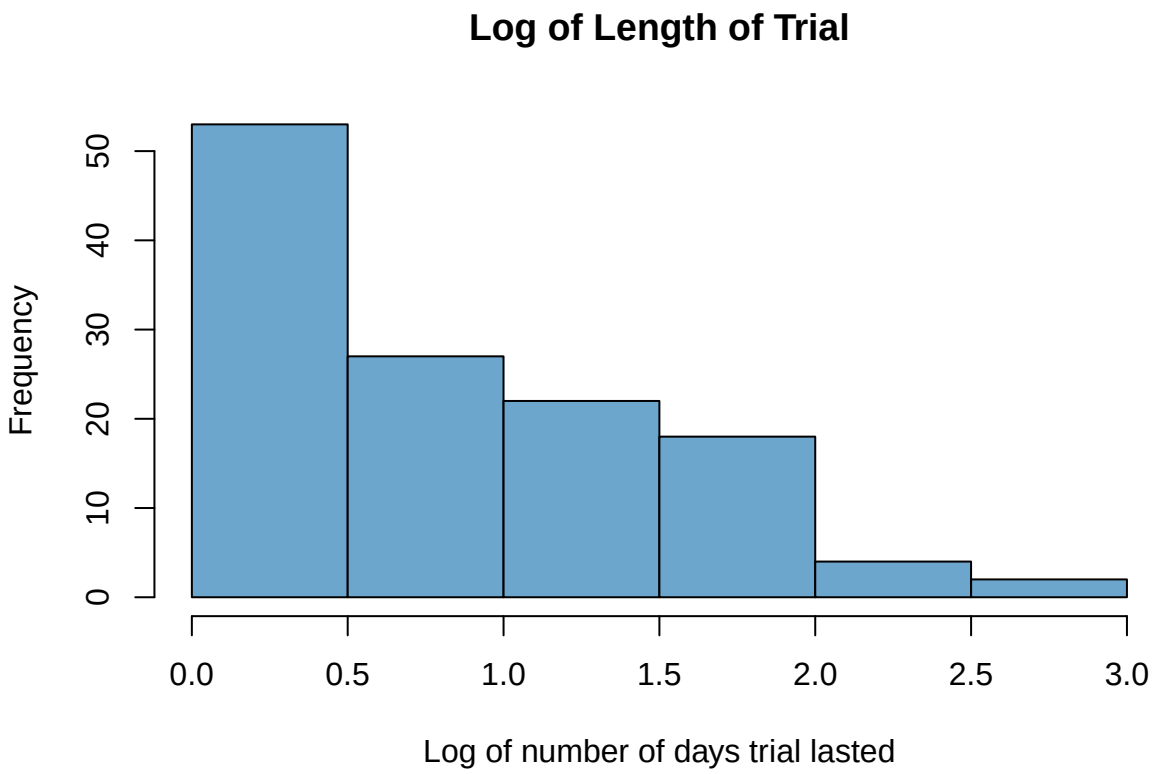
Observations

The Normal Q-Q plot shows systematic deviations from the line, so we cannot validate the Normality assumption for errors.

Transformations

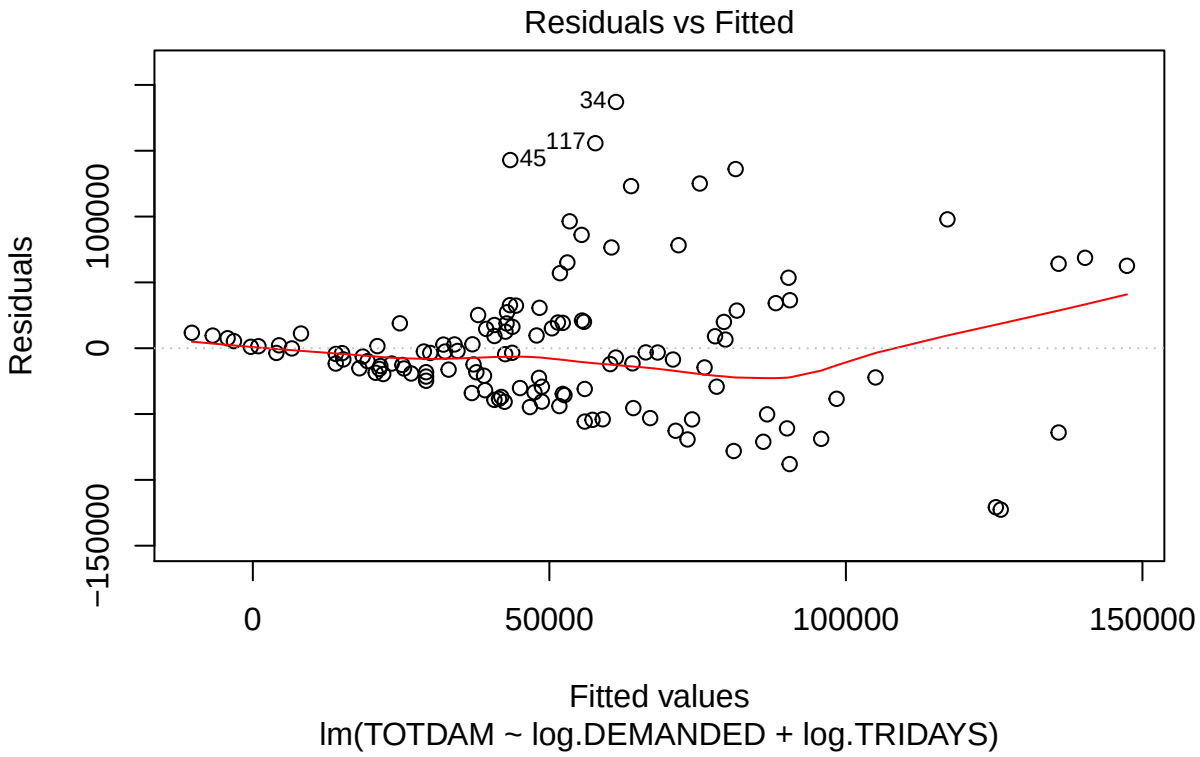
To validate the assumptions for errors, we perform two transformations. Since amount demanded and trial length showed the strongest skew, we will be taking their natural logs and producing updated residual and Normal Q-Q plots along with histograms.


Log of Amount Demanded
Frequency
0
10
20
30
40
8
10
12
14
16
Log of total amount requested by plaintiff (in $)



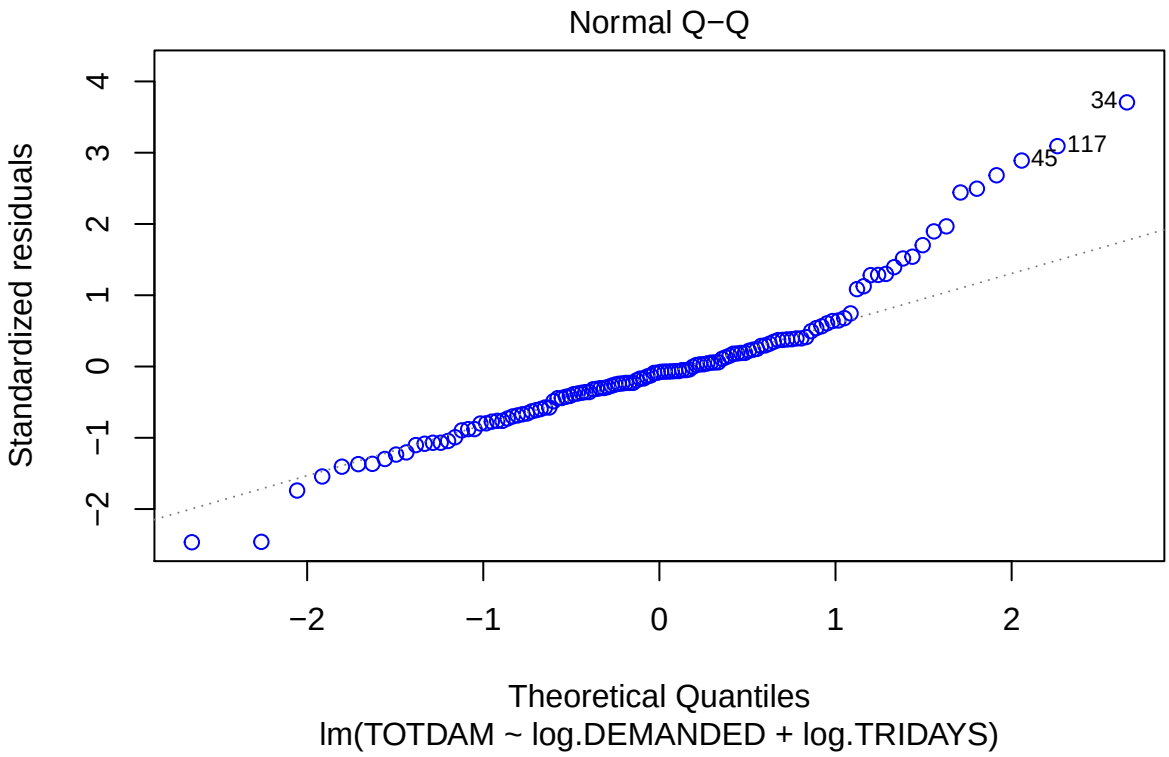
Observations

The skewness for both predictors has been significantly reduced. Log of amount demanded shows almost no skew and appears symmetric. While log of trial length still shows some skew, it has been reasonably mitigated.



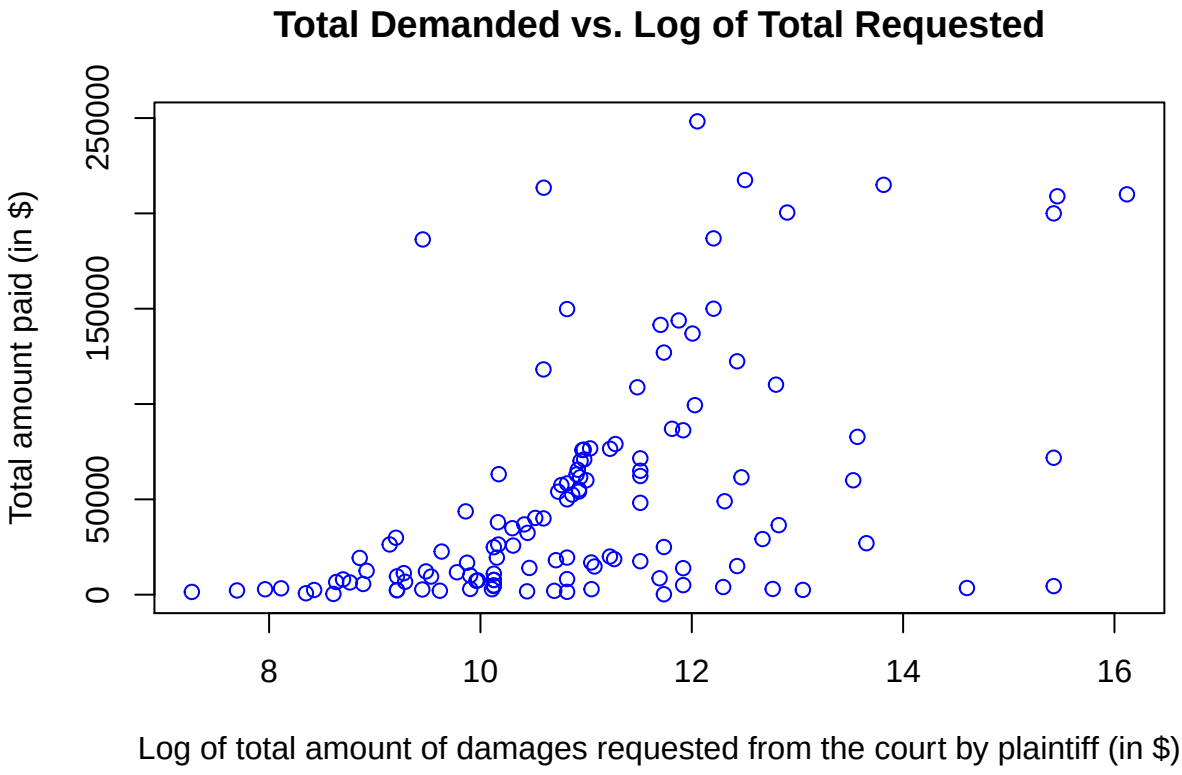
Observations

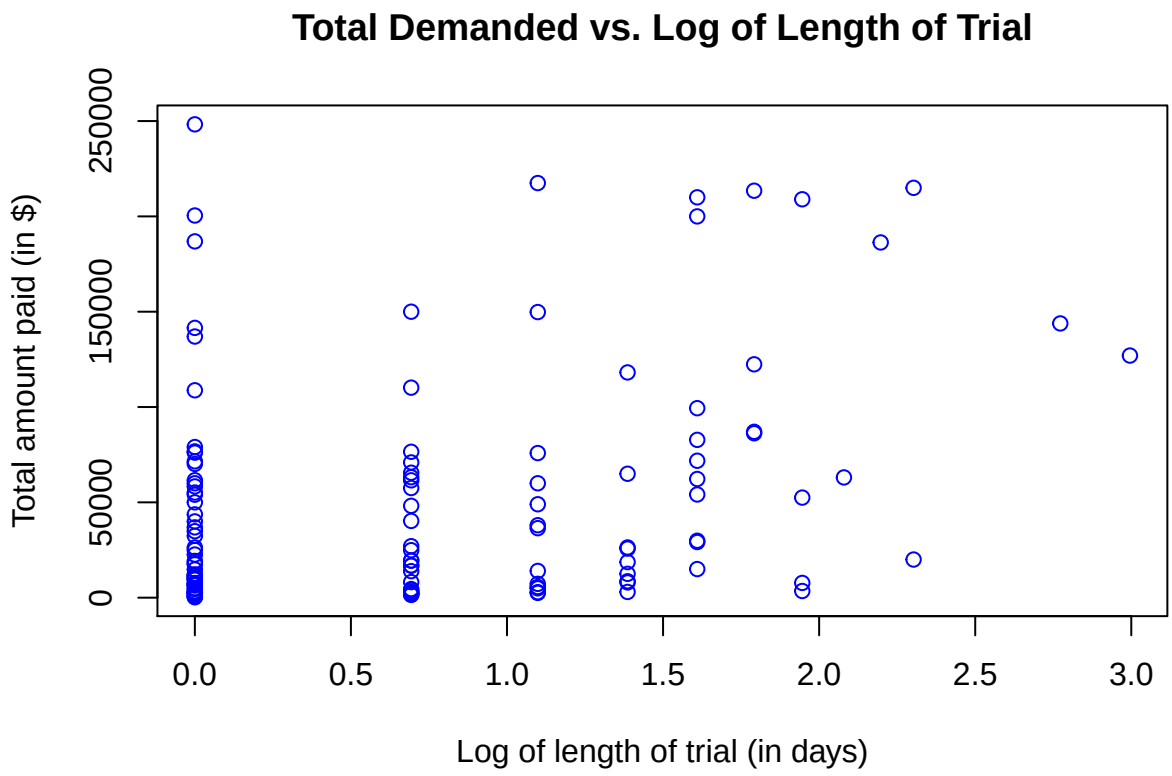
The pattern seen for the residual plot of the untransformed data has been reduced. Since residuals are shown to be generally scattered above and below the 0-line with little to no pattern, centered around 0, and constantly spread above and below the 0-line, we can reasonably assume that the errors are independent, have mean 0, and have constant standard deviations.



Observation

Overall, the standardized residuals of the transformed data deviate from the line far less than those of the untransformed data. Although the standardized residuals begin to deviate for theoretical quantiles greater than 2, we can still assume that the errors are normally distributed.

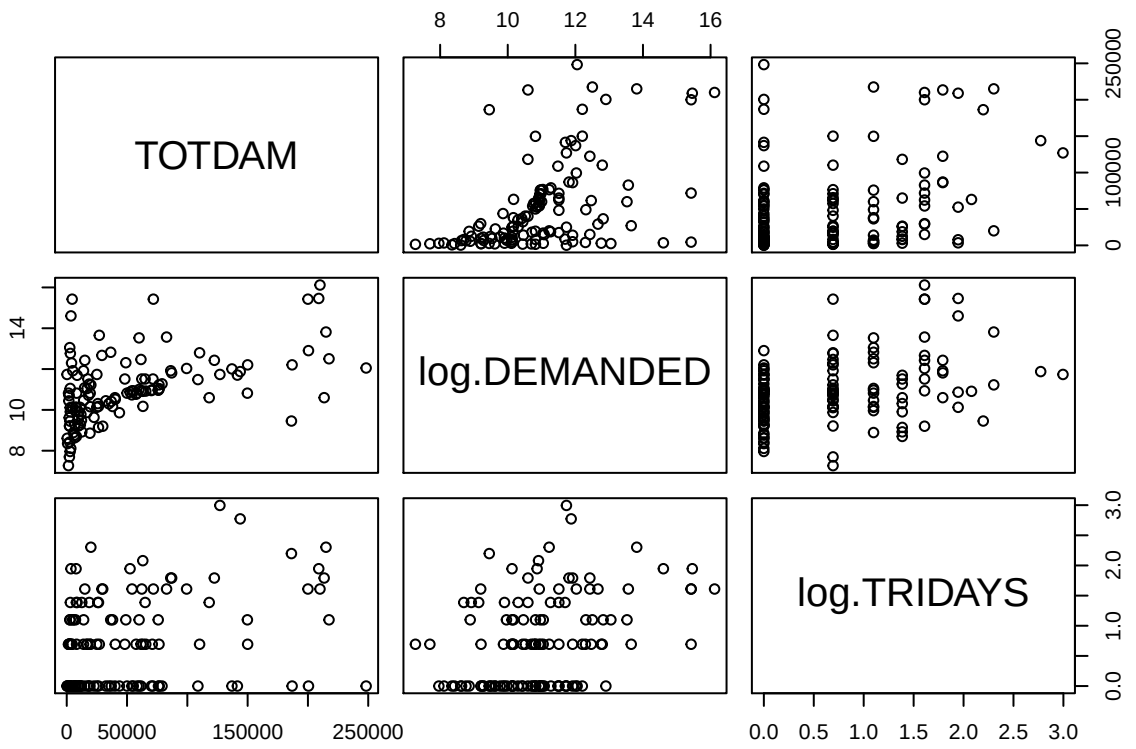

Total Demanded vs. Log of Total Requested
Total amount paid (in $)
0
50000
150000
250000
8
10
12
14
16
Log of total amount of damages requested from the court by plaintiff (in $)



Observations

Since both scatterplots show reasonable a linear association between total paid and its predictors, we can assume the linearity of the relationship.

Relationships Between Quantitative Variables




```
##                 TOTDAM log.DEMANDED log.TRIDAYS
## TOTDAM       1.0000000    0.5116685   0.3148573
## log.DEMANDED 0.5116685    1.0000000   0.3710529
## log.TRIDAYS  0.3148573    0.3710529   1.0000000
```

Observations

Notice that all modeled relationships above show some linear pattern. To test for multicollinearity, we will test for variance inflation factor.

```
## log.DEMANDED  log.TRIDAYS
##     1.159663     1.159663
```

Since no variables produced a vif greater than 2.5, we proceed without worry of strong multicollinearity.

Summary of Chosen Model

```
##
## Call:
## lm(formula = TOTDAM ~ log.DEMANDED + log.TRIDAYS + CLAIMTYPE +
##     CLAIMTYPE:log.DEMANDED + CLAIMTYPE:log.TRIDAYS, data = court)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -98464 -19450  -4999  11189 187596
##
## Coefficients: (1 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -228113      96397  -2.366 0.019728 *
## log.DEMANDED                        30600       8482   3.608 0.000468 ***
## log.TRIDAYS                        -60023      20797  -2.886 0.004703 **
## CLAIMTYPEMalpractice              -528088     233992  -2.257 0.026010 *
## CLAIMTYPEMotor                     201095     108570   1.852 0.066700 .
## CLAIMTYPEOther                     123956     107379   1.154 0.250868
## CLAIMTYPEPremises                    9499     159813   0.059 0.952714
## CLAIMTYPERental                     -7892     241885  -0.033 0.974031
## log.DEMANDED:CLAIMTYPEMalpractice    52547      20955   2.508 0.013627 *
## log.DEMANDED:CLAIMTYPEMotor         -26422       9705  -2.723 0.007545 **
## log.DEMANDED:CLAIMTYPEOther         -16923       9595  -1.764 0.080582 .
## log.DEMANDED:CLAIMTYPEPremises      -10376      14758  -0.703 0.483479
## log.DEMANDED:CLAIMTYPERental         -3422      23969  -0.143 0.886746
## log.TRIDAYS:CLAIMTYPEMalpractice        NA         NA      NA       NA
## log.TRIDAYS:CLAIMTYPEMotor           65064      23958   2.716 0.007692 **
## log.TRIDAYS:CLAIMTYPEOther           77602      22133   3.506 0.000661 ***
## log.TRIDAYS:CLAIMTYPEPremises       123788      48501   2.552 0.012088 *
## log.TRIDAYS:CLAIMTYPERental          59087      44343   1.333 0.185475
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45220 on 109 degrees of freedom
## Multiple R-squared:  0.5005, Adjusted R-squared:  0.4272
## F-statistic: 6.827 on 16 and 109 DF,  p-value: 1.538e-10
```

Observations

All necessary conditions for a multiple regression model were satisfied. For the incorporation of claim type, an interaction model was tested and kept since the majority of the interaction terms are significant. Furthermore, the model itself is significant with a P-value of 1.538e-10. From the multiple R-squared produced by our model, 50.05% of the variation in the total amount paid can be explained by the modeled relationship with the logs of all of its predictors.

Prediction

With our established model, we can now predict the amount paid to a plaintiff who demands 100,000, has a trial of five days long, and a malpractice claimtype.

Our model with the following values is as follows: (dummy variables equal to zero in this case are not included)

$$totaldemanded = \beta_0 + \beta_1(log(demanded)) + \beta_2(log(tridays)) + \beta_3(claimtypemalpractice) + \beta_8(log(demanded) * claimtypemalp$$

$$totaldemanded = -228113 + 30600(log(demanded)) - 60023(log(tridays)) - 528088 + 52547(log(demanded))$$

$$totaldemanded = -228113 + 30600(log(100000)) - 60023(log(5)) - 528088 + 52547(log(100000))$$

$$totaldemanded = 104460.9$$

We predict that the plaintiff will be paid 104460.9 dollars.

Discussion

Overall our model has demonstrated that the amount demanded by a plaintiff, the length of a trial, and the type of claim made all influence the final amount paid to the plaintiff by varying degrees. While our model is significant, there is room for improvement. To start, new variables that are more linearly related with the final amount could be introduced or used to replace predictors with weaker linear relationships with the response. There exists much more data for civil court cases that could include location, amount paid to plaintiff's lawyer(s), or demographic makeup of one or more parties. Finally, an Anova model could be used for the categorical variable claim type to help determine its relationship with total paid.