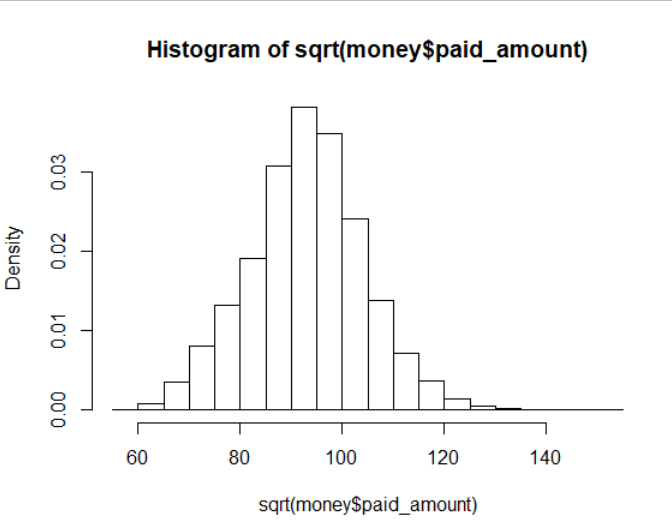
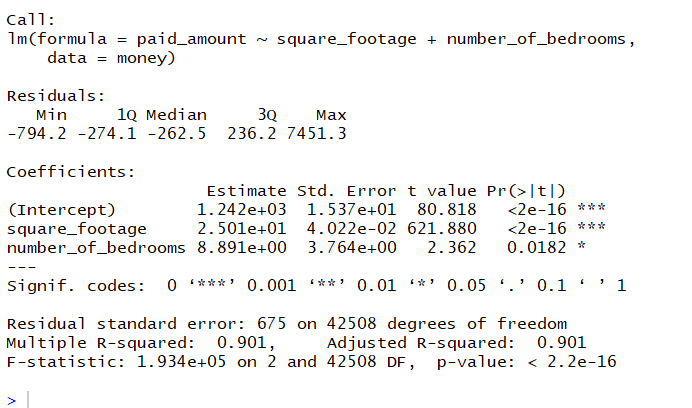
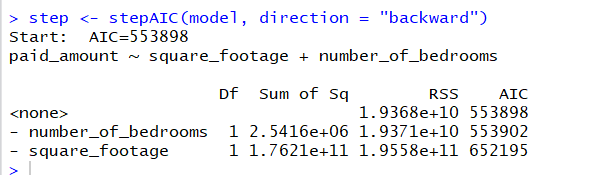
We attempt to estimate the potential loss generated by the false negatives. For each claim case denied, the coverage that could have been paid to the customer should the case wasn't denied is not available in this dataset. Fortunately, for each claim accepted, the coverage paid to the customer is accessible.

Using histogram, we were able to identify the distribution of coverage as normally distributed density and thus we choose a linear regression model as appropriate for our model to estimate the amount of money that could have been paid to a fraudulent customer, if he or she is misclassified as a normal customer.

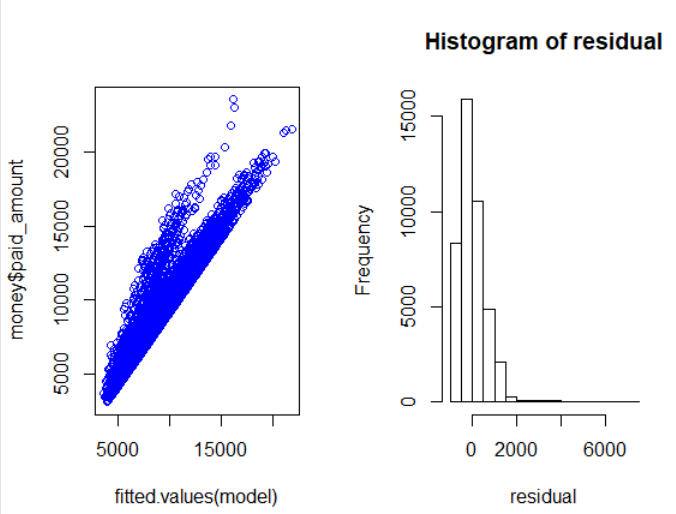


We obtained our best fitted variables by comparing summary tables of coverage against different variables and explorative analysis suggests that square footage of a home and numbers of bedrooms are the only significant predictors of coverage as per t-test statistics with p-values less than 0.05 also with the minimum value of AIC. Keeping other variables constant, we noticed that one unit increase in square footage of the house will result in an increase in coverage by 25.01.





There is a strong correlation (R square of 0.901) between fitted value (Yhat) and actual response (Y) as illustrated in the graph. Thus we would apply this model to predict coverage of a fraud case. The residuals are normally distributed despite slight right skewness.



The following boxplot indicates that the median of coverage for different numbers of bedrooms are pretty much the same, in particular there are lots of outliers above the fourth quantile which may represents that the coverage may depend on other factors that not given in the dataset, such as the reason why customer claims and so on…

