Predicting Cost of Property Loss Incurred by Tornado activity in the USA

ABSTRACT

I have illustrated a statistical model for predicting cost of property loss based on the tornado activity in United States during 1950 to 2014. The model is based on the data collected from The National Oceanic and Atmospheric Administration's (NOAA) Storm Prediction Center (SPC). The sustainability of positive amount of property loss in the dataset is estimated from a classification model using logistic regression on the training data. After classifying the observations in the dataset using logistic regression - predicting if there was property damage sustained – I have used multiple linear regression model. The purpose of this regression is to estimate the dollar amount of property damage endured, given we have already predicted positive property damage in the logistic regression. These relationships are broadly consistent with understanding financial models and have a high commercial value for the insurance companies. The insurer can predict the damage, which can occur due to the natural calamity and can estimate how much damage they will be responsible for, thus, it will help the company to calculate premium charges for the incoming customers in that area for a year.

DATA OVERVIEW

The Storm Prediction Center (SPC) maintains a data set of all reported tornadoes in the United States from 1 January 1950 to the present. The SPC dataset is the most reliable archive available for tornado and related studies. (We download the dataset from http://www.spc.noaa.gov/gis/svrgis/.) The dataset has 28 features of which 11 variables are numerical fields namely, Tornado Number, Injuries, Fatalities, Loss, Crop-loss, Length, Width, Latitude, Longitude, E-Latitude, and E-Longitude.

Data set variable description is provided in the attached document.

Data Exploration and Visualization is shown in the attached document.

VARIABLE SELECTION

After analyzing the complete dataset provided by NOAA, I have determined seven variables that would be helpful to predict the property loss.

Variable	Description
Property Loss	Outcome Variable: Amount of loss (in Millions of Dollars)
Storm Seasons	Categorical Variable: Tornado Season:1, Off-Season: 0
State Severity	Categorical Variable of Percentage of National Tornadoes in State: 1(<2%), 2(2-3%), 3(3-4%), 4(>4%)
F-Scale	Categorical variable of Fujita Scale that measures tornado severity. 0-5, 5 being most destructive

Injuries	Number of Injuries occurring due to the tornadoes
Fatalities	Number of Fatalities occurring due to the tornado
Length	Distance the tornado traveled (in miles)
Width	Diameter of the Tornado(yards)

Looking at the data, there was a significantly higher frequency of tornados in April through July, and significantly less in other months. This was a consistent pattern, which was similarly present since 1950. Thus, Storm Season is a binary variable indicating whether the tornado occurred during tornado season or not. The State variable was consolidated into four categories based on the percentage of all tornados that hit the state. For example, Texas contains roughly 9% of all tornados that occurred in the United States for our dataset. Given this, we categorized the states accordingly into the State Severity variable (explained in the data exploration document).

PROCEDURE OVERVIEW

In order to predict a dollar estimate for damages, two regression models were fit to the data. My approach is to use the nineteen years of the data for training and validating the models. My ultimate goal is to predict the approximate dollar amount of property loss from a given tornado. To estimate this using a linear regression, a logarithmic transformation of the independent variable Property Loss was necessary in order to ensure the assumptions of linear regression hold, specifically that the dependent variable exhibits a linear relation with each of the predictors. Due to the high frequency of zero-valued observations for property loss, taking the logarithm will cause computational problems. Thus, before fitting the linear regression I have classified the observations in the dataset by predicting whether or not there was a positive amount of property loss sustained. To perform this classification of the data I have fit a logistic regression on the training and validation portion of the dataset. Independent of the logistic regression, I have fit a linear regression model on the non-zero observations of property loss in 1996-2014 dataset. A k-fold validation process was run and the model with the lowest mean absolute error was chosen. The two models were tested to determine how well they were able predicted future values of property loss. I could then predict whether or not a particular storm resulted in positive property loss. The observations from the validation data for which we had predicted positive property loss were input into the linear regression model to predict the actual dollar amount of damage sustained by that tornado. This process resulted in the predicted estimate of property damage. Statistics of the results were collected and then assessed for the level of commercial utility they present. The subsequent sections illustrate the result of this procedure and conclusions drawn from those outcomes.

CLASSIFICATION MODEL TO PREDICT A TORNADO AS LOSS OR

NO-LOSS

Before I have fit the linear regression model, I have predicted whether or not property damage will result from a certain tornado. In my dataset, 37 percent of the tornados from 1950 to 2014 produced zero dollars of property loss. My subsequent linear regression model will use logarithmic transformation of Property Loss. However, since the logarithm of zero is undefined we must first use logistic regression to predict the binary outcome of Property Loss. The linear regression will thus be used to predict the amount of property damage sustained, given we have predicted positive property loss from the logistic regression model. The linear regression will thus be fit to only a little more than half of the observations, since we will remove all the non-positive values of the outcome variable before fitting.

Logistic Regression

The logistic regression is designed to classify the observations in the dataset into groups of whether or not the tornado had inflicted property damage. In my training dataset we converted our outcome variable of property damage, which is continuous, to being a binary outcome variable of either zero or one, where "1" indicates property loss and “0” indicates no property loss. Then I have fit this binary outcome variable to the following logistic regression model:

$$PropertyLoss=1=logit\\(\beta0+\beta1*Storm.Season+\beta2*FScale+\beta3*Injuries+\beta4*Fatalities+\beta\\5*State.Severity+\beta6*Width\text{h}+\beta7*Lengt\text{h})$$

Where Storm Season and State Severity are factor variables. Before fitting the model I have selected a random sample containing 30 percent of our dataset to be used for model validation. Fitting the full model, I have produced the logistic regression output with the associated confusion matrix:

```
=== Confusion Matrix ===

a        b        c <– classified as
8915     2365     a=L
2156     4547     b=N
```

Multiple Linear Regression

After classifying the observations in the dataset using logistic regression - predicting if there was property damage sustained – I have fit the multiple linear regression model. The purpose of this regression is to estimate the dollar amount of property damage endured, given we have already predicted positive property damage in the logistic regression. To fit the model I have removed the observations in our data set that have zero property damage. The relationship of Property Loss with most of the independent variables appears non-linear I took the logarithm of Property Loss. Since the logarithm of zero is undefined, the observations with zero property damage must be removed. However, since my linear regression model is attempting to estimate the amount of Property Loss, given positive amount of damage, the predictive significance of the model has not changed. State Severity and Storm Season are categorical non-ordinal variables, and F-scale is ordinal. The logarithmic transformation appears to correct much of the non- linearity between the predictors and the response. However the variables Injuries and Fatalities continued to appear to violate the linearity assumption. By taking the square root of both these predictors the non- linearity is corrected. I began by fitting the full model and then performing backwards elimination on the variables to find improvements in the error terms. The full model thus uses: FScale, sqrtInjuies, sqrtFatalities, State Severity, Storm Season, Length, and Width. Recall that State Severity is a factor variable with categories equal to State.Severity1-4, where State.Severity1 is the base category. Storm Season is a binary variable indicating if the tornado took place during storm season. The full model when fit to the whole training data set gives the following output:

Model: log$PropertyLoss$= $\beta0+\beta1*Storm.Season+\beta2*State.Severity+\beta3*FScale+\beta4*sqrtInjuries+\beta5*sqrtFatalities +\beta6*Width+\beta7* Length$