STAT 225 Project

Predicting Hurricane Deaths

Adi Arifovic, Oliver Baldwin Edwards, Leah Johnson

Contents

1 Background and Research Question

Our project focused on predicting the number of deaths caused by hurricanes based on variables such as windspeed, atmospheric pressure, and property damage. We hypothesized that the number of deaths would increase with an increase in maximum sustained windspeed and property damage, as intuitively it would make sense for more destructive hurricanes to also be more deadly.

2 Data and Methods

2.1 Data Collection

Our data came from the “hurricNamed“ dataset, from the “DAAG” R package. It contained data on 94 named hurricanes that made landfall in the United States mainland from 1950 to 2012. Within the dataset, we had information on the number of deaths per hurricane, the damage caused by the hurricane,

The data were sourced from several places, and were previously used in a research paper studying the relationship between the gender associated with the hurricane’s name, and the human damage it caused.

2.2 Data Exploration

We began with univariate data exploration. We made the decision to remove the categorical variables early on. One of them detailed the States affected by the hurricanes which we felt had too many levels to compare across, even when grouped. We also removed the year, date of first landfall, and the male/female name variable, as we were not interested in comparing across these. Our focus lay more with the actual physical attributes of the hurricane, and how these could be related to the number of deaths.

We removed one of the variables relating to base property damage, as it overlapped with a base damage variable adjusted to be in millions of 2014 US dollars. We kept this alternate variable instead, because it would be easier to compare this across a large time frame.

We looked at several kernel density estimates for each of the variables, running through multiple choices of kernel and bandwidth before selecting the one that best struck a balance between overfitting and underfitting.

We were particularly struck by the variable ‘deaths,’ - number of deaths - which was our outcome, as it was heavily right skewed. As such, we undertook a logarithmic transformation, which resulted in a more spread out distribution that we felt was more appropriate. This can be seen in Figure 1.

We faced a similar issue with the ‘NDAM2014’ - normalized damage - variable (KDEs shown in Figure 2), and chose to use a logarithmic transformation here as well, to better be able to fit a linear model later on.

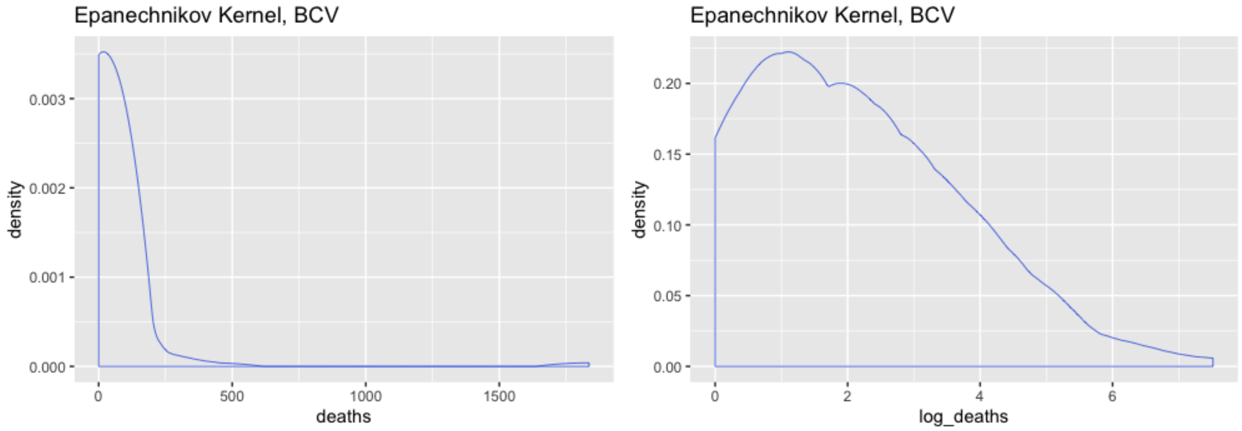

Figure 1: Kernel Density Estimates for ‘deaths’ Before and After Logarithmic Transformation

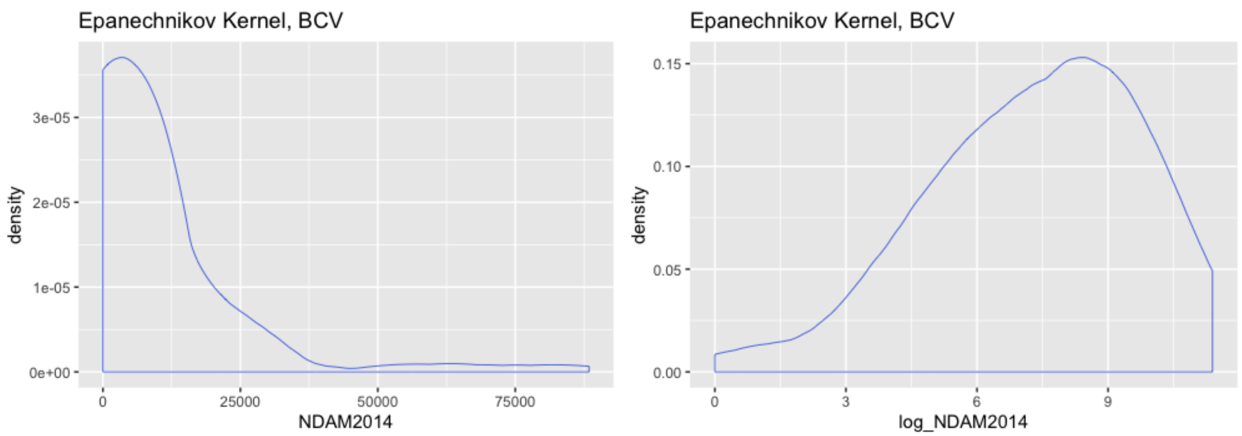

Figure 2: Kernel Density Estimates for 'NDAM' Before and After Logarithmic Transformation

Finally, we conducted correlation tests between the outcome, and all of our variables. In most cases, we were looking for a positive association. We used Kendall's tau rather than a Spearman or Pearson test, as it is a more robust test, and does not require monotonicity or linearity of the association as an assumption.

2.3 Models

Using best subsets, we ran through all possible combinations of linear models, and found optimum models for different numbers of terms. Following this, we used a series of nested F-tests in order to determine which predictors were significant in the model and which could be dropped, after accounting for the effects of others.

In order to check the conditions for the linear model, we examined a plot of residuals versus fitted values to check for linearity and equal variances. We then conducted a Kolmogorov Smirnov test to compare the residuals from our linear model to a normal distribution. We also plotted the empirical cdf for our residuals in order to visualize these compared to a normal cdf.

Following the Linear Model, we fit a JHM for comparison purposes, as the rank-based nature of this lends itself to a more robust model. We used a variation on a stepwise method, adding predictors into the model as well as conducting a series of drop in dispersion tests in order to make sure they were all significant.

Finally, we fit a Generalized Additive model. Since linear models are limited to showing linear relationships, we felt the use of a Generalized Additive Model might better showcase the relationships in our data, as well as reveal any that might have been overlooked by our OLS regression model, and the JHM model.

To fit the Generalized Additive Model, we looked at which of our variables had the strongest correlation with our outcome. We then used a variation on forward selection to come up with a model. We began by including the log-transformed normalized damage in the model in a variety of forms (linearly, using bsplines with varying degrees of freedom, using smoothing splines), and finding the model with the lowest AIC. This became the basis for the next model, in which we added the second highest correlated predictor in a similar manner, and so on.

This method gave us a model predicting log deaths with a smoothing spline on normalized damage, and a bspline with 7 degrees of freedom on landfall pressure. However, when we initially visualized this model, we found that a cubic polynomial for landfall pressure and a bspline with the default degrees of freedom for normalized damage actually gave us a better visual representation.

3 Results

Initially, during the data exploration stage, we carried out correlation tests between each of the variables and our outcome. In each case, the null hypothesis was

$H_0 : \tau = 0$

where τ was the Kendall correlation coefficient between log deaths and the variable in question. The specific alternative hypotheses and results along with estimations of the correlation coefficient can be found in Table 1.

While the results of the correlation tests were largely

Variable	Hypothesis Tested	$p-$value	Correlation Estimate
NDAM2014	$H_A : \tau > 0$	$8.073 \cdot 10^{-16}$	0.578
LF.times	$H_A : \tau > 0$	0.008542	0.209
LF.PressureMB	$H_A : \tau > 0$	1	-0.431
LF.PressureMB	$H_A : \tau < 0$	$1.915 \cdot 10^{-9}$	-0.431
LF.WindsMPH	$H_A : \tau > 0$	$7.606 \cdot 10^{-6}$	0.331
BaseDam2014	$H_A : \tau > 0$	$7.691 \cdot 10^{-15}$	0.558

Table 1: Correlation Test Results

as expected, we did find that when we tested for a positive correlation between landfall pressure (‘LF.PressureMB’) and log deaths (yellow row in Table 1), we got a $p-$value of around 1. This prompted us to run another test, this time looking for a negative correlation (green row in Table 1), for which we found significant supporting evidence.

After carrying out the best subsets selection method and series of nested F tests, we arrived at a single predictor model predicting log deaths using log ‘NDAM2014’ (normalized damage). The summary is shown in Table 2.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.6803	0.3937	-4.27	0.0000
NDAM2014	0.4973	0.0512	9.71	0.0000

Table 2: OLS Linear Regression Model Summary

This was a good model, with a relatively low AIC (301.11), and also a moderately high adjusted R^2 of 50.08%. We felt that our model explained a fairly reasonable amount of variation in the outcome, especially considering that it was a single predictor model.

In checking conditions for the linear model, we noted that Independence was reasonably assumed.

In order to check Linearity, and the Equal Variances condition, we looked at a plot of the residuals vs the fitted values of our linear model (Figure 3). This exhibited a fan shape, which indicates unequal variances, but we proceeded with caution.

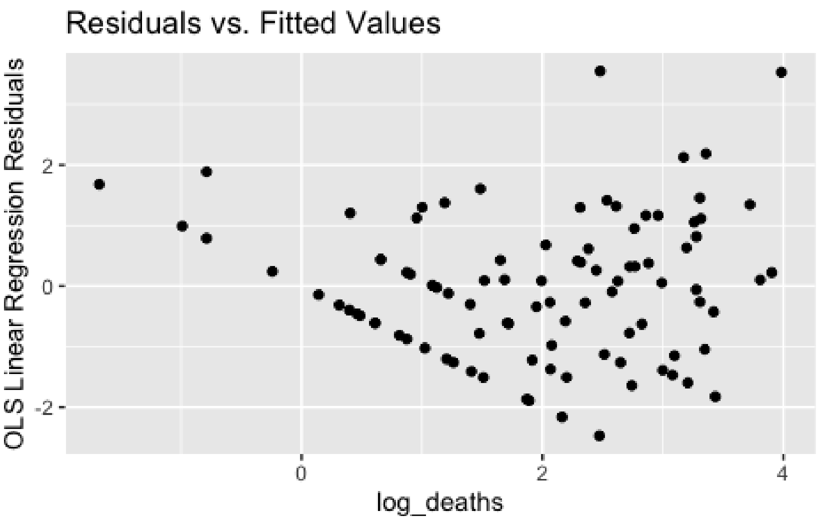

Figure 3: Plot of Residuals vs Fitted Values

In order to check that the normality condition held for this model, we undertook a Kolmogorov Smirnov test, to compare the distribution of the residuals from our model to a normal distribution we generated. Our null hypothesis was that there was no significant difference between the two, while our alternative hypothesis was that there was a significant difference.

$H_0 : F(t) = F^*(t)$

$H_A : F(t) \neq F^*(t)$ for at least one t

Where $F(t)$ refers to the estimated CDF of the distribution of residuals of our linear model, and $F^*(t)$ is the CDF of the normal distribution.

We failed to reject the null hypothesis at any reasonable significance level ($p = 0.7095$). The plot of the empirical cdf of residuals compared to a normal cdf supports this as well. As such, we were able to conclude that the normality condition was satisfied.

Figure 4 shows the empirical cdf of residuals (in blue) compared with a normal distribution (in red). We see that they almost overlap completely, which is in line with our assumption of normality in fitting the linear model.

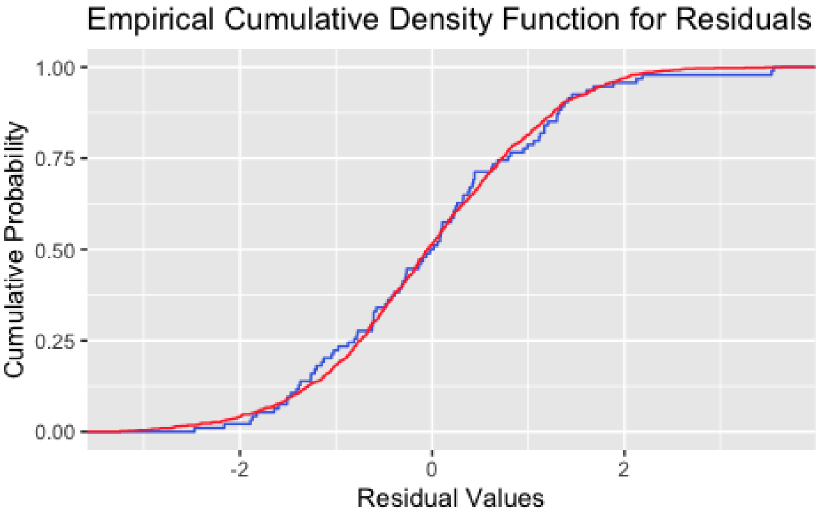

Figure 4: Empirical CDF of Residuals vs Normal CDF

Since we had an issue with the equal variances condition in the ordinary least squares regression model, we fitted a rank-based JHM model. Since this is a nonparametric procedure, it is more robust than the ordinary least squares estimation, and as such holds up better when conditions such as equal variances are not satisfied.

We manually implemented a step-wise selection procedure in order to find the best possible model to fit, and conducted drop in dispersion tests throughout the process to ensure the significance of each predictor in the model, after accounting for the effects of others.

Interestingly enough, this process resulted in a single predictor model, with the same predictor - logarithmically transformed 'NDAM2014' as in the OLS linear regression model. In this case, we have a multiple R^2 of 48.07%, which is decent.

Finally, we fit a Generalized additive model, in order to look for possible trends and relationships in the data our previous two models might have missed, being restricted by linearity. The Generalized Additive Model, being able to include smooths, has greater flexibility to fit the data, and as such is better able to exhibit more local trends.

Initially, we manually undertook a step-wise selection method similar to the one we used when fitting the JHM model. We included each predictor one by one, testing several models at each stage in order to find the one with the lowest AIC that also had a reasonable number of terms.

This was the GAM using a b-spline with the default number of degrees of freedom for log 'NDAM2014', and an s-spline with 5 as the df argument for 'LF.PressureMB'. We plotted this, but noticed that there was a poor fit for the 'LF.PressureMB' scatterplot. The fit was roughly quadratic or cubic to the eye, so we attempted fitting the 'LF.PressureMB' term with a polynomial. We ended up deciding on the cubic polynomial, based on a visual test. The final result was then the GAM predicting log deaths using a bspline with the default degrees of freedom on log 'NDAM2014', and a cubic polynomial for 'LF.PressureMB'.

An interesting thing to note here, is in our GAM we include a predictor in the model that is excluded in both our ordinary least squared and rank-based linear regression models. This predictor is 'LF.PressureMB', which we saw earlier was the only predictor to be negatively correlated with the outcome. As such, we think the GAM might provide us with a better picture of the data as a whole.

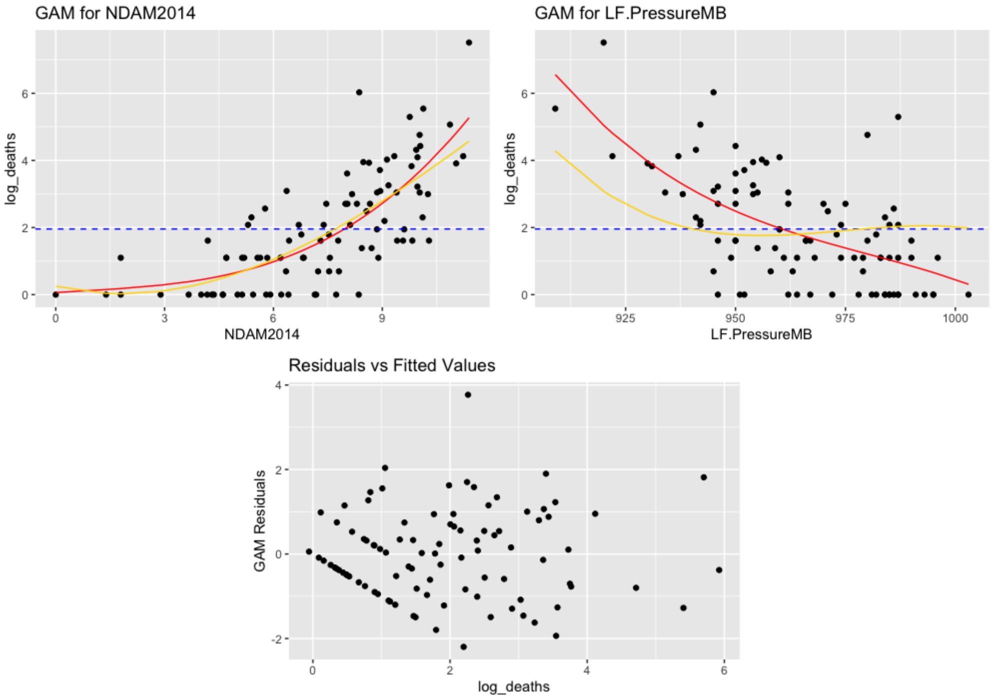

Figure 5: GAM Visualization

4 Conclusion

5 Limitations and Drawbacks

References