STAT225

Oliver Baldwin Edwards, Leah Johnson, Adi Arifovic

10 May, 2020

Predicting Hurricane Deaths

Group 2: Oliver Baldwin Edwards, Leah Johnson, Adi Arifovic

table of contents here

Purpose

We hope to predict the log number of hurricane deaths based on variables such as maximum sustained windspeed, atmospheric pressure, and property damage from a dataset containing data on 94 named hurricanes that made landfall in the US mainland from 1950 through 2012. We would expect the number of deaths to increase as the maximum sustained windspeed and property damage increases, since it would make sense for more destructive hurricanes to also be more deadly.

Data

Source

We plan to use the dataset "hurricNamed" from the "DAAG" R package containing data on 94 named hurricanes that made landfall in the US mainland from 1950 through 2012. It contains information on the number of deaths, the name of the hurricane, the year of the hurricane, the damage caused by the hurricane, etc. The data is sourced from multiple places and was used in a research paper claiming that hurricanes with female names did more human damage (after adjusting for the severity of the storm) than those with male names. Therefore, the data seems fairly reliable.

Response Variable

Our response variable is the log number of deaths that occurred due to a hurricane. The units are number of human deaths. We observe that the range of deaths are from 0 to 1836.

Explanatory Variables

The explanatory variables we will examine are LF.WindsMPH, LF.PressureMB, LF.times, BaseDam2014, NDAM2014, and deaths.

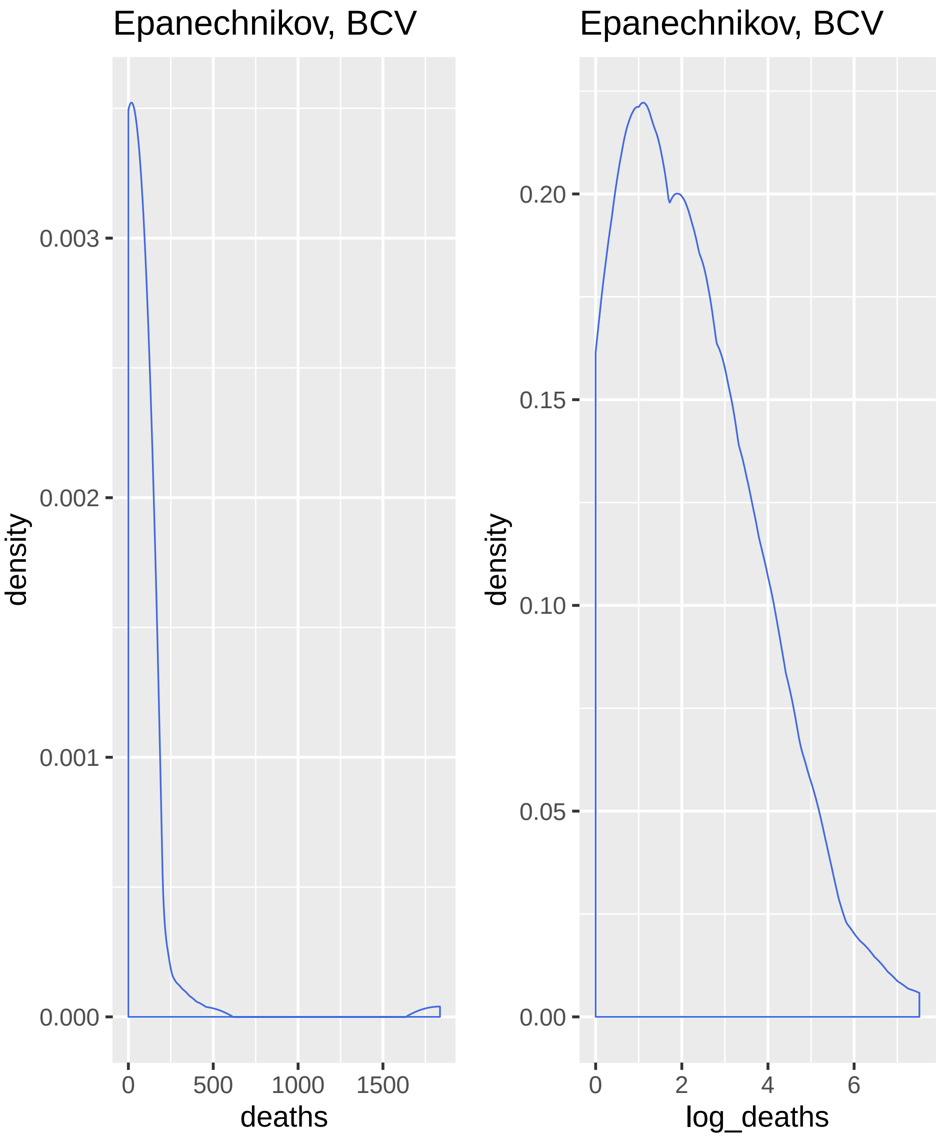
1. LF.WindsMPH describes the maximum sustained windspeed for each hurricane in miles per hour
2. LF.PressureMB describes the atmospheric pressure at landfall in millibars
3. LF.times describes the number of times the hurricane made landfall
4. BaseDam2014 describes the property damage caused by the hurricane in millions of 2014 US dollars
5. NDAM2014 describes the amount of damage the hurricane caused had it appeared in 2014 (no units given)
6. deaths describes the number of human deaths the hurricane caused

Exploratory Data Analysis

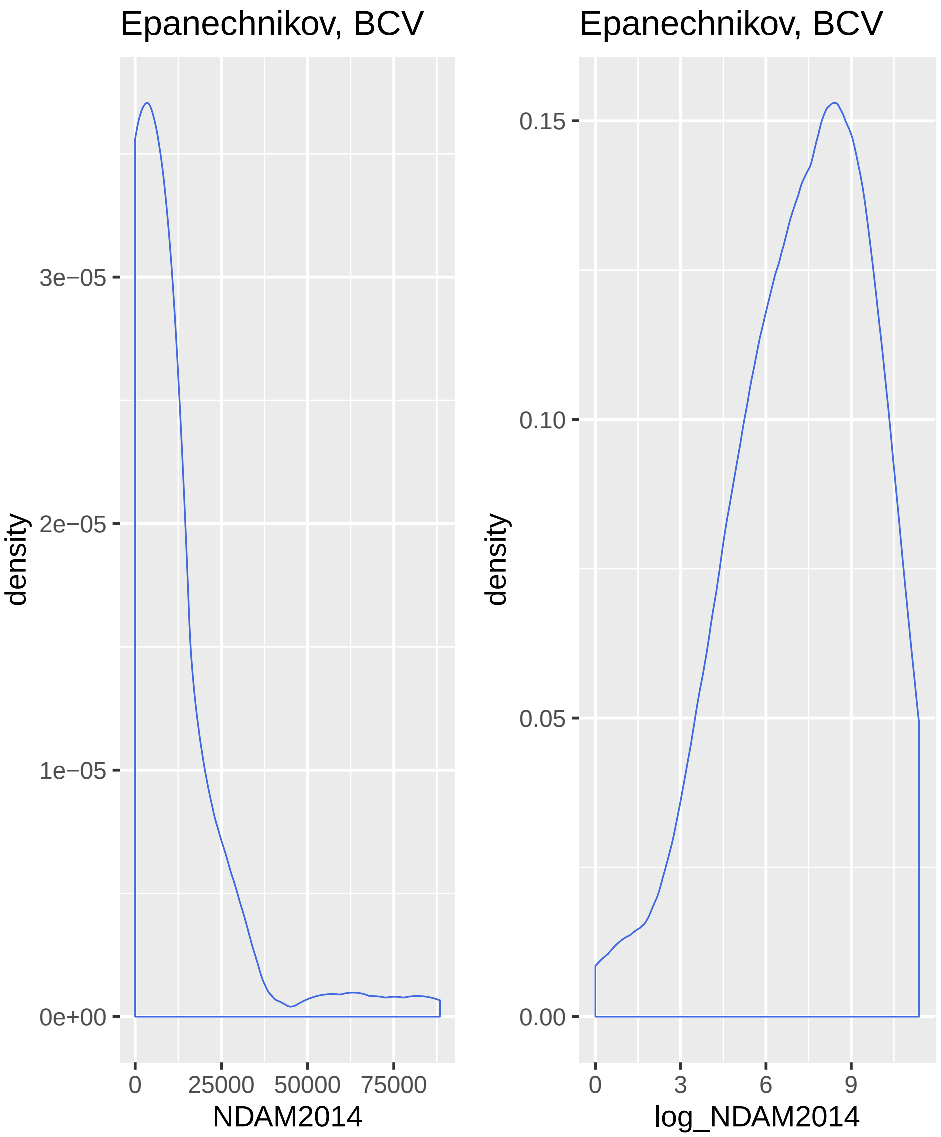
Kernel Density Estimates

The researchers first looked at kernel density estimates of our predictors, and found that `deaths`, `NDAM2014`, and `BaseDam2014` had very non-normal/skewed kernel density estimates. The researchers found that adding a `log()` transformation to `deaths`, `NDAM2014`, and `BaseDam2014` resulted in more normal kernel density estimates and minimized the effects of the outliers. (The application of a log transformation allows the researchers to fit linear regression models (OLS and JHM) to the data.) A few of the afforementioned KDEs can be seen below:

```
p1 <- ggplot(data = hurricane_logdeaths,
                         aes(x = hurricane_numeric$deaths)) +
                    geom_density(bw = "bcv", kernel = "epanechnikov", size = 0.3,
                                 color = "royalblue") +
                    ggtitle("Epanechnikov, BCV") +
                    xlab("deaths")


p2 <- ggplot(data = hurricane_logdeaths,
                         aes(x = hurricane_logdeaths$log_deaths)) +
                    geom_density(bw = "bcv", kernel = "epanechnikov", size = 0.3,
                                 color = "royalblue") +
                    ggtitle("Epanechnikov, BCV") +
                    xlab("log_deaths")

p3 <- ggplot(data = hurricane_logdeaths,
                         aes(x = hurricane_numeric$NDAM2014)) +
                    geom_density(bw = "bcv", kernel = "epanechnikov", size = 0.3,
                                 color = "royalblue") +
                    ggtitle("Epanechnikov, BCV") +
                    xlab("NDAM2014")

p4 <- ggplot(data = hurricane_logdeaths,
                         aes(x = hurricane_logdeaths$NDAM2014)) +
                    geom_density(bw = "bcv", kernel = "epanechnikov", size = 0.3,
                                 color = "royalblue") +
                    ggtitle("Epanechnikov, BCV") +
                    xlab("log_NDAM2014")

cowplot::plot_grid(p1, p2, nrow = 1)
```



```
cowplot::plot_grid(p3, p4, nrow = 1)
```



For `log_deaths`, the rectangular kernel overfits and the bandwidth selected by UCV is similarly unsuitable. The gaussian kernel looks appears to be oversmoothing. The triangular and epanechnikov kernels are similar, but the researchers decided to go with th epanechnikov kernel with bandwidth selected using BCV. The KDE looks fairly reasonable overall, and epanechnikov is generally a good choice of kernel.

Correlation Tests

The researchers proceeded to run correlation tests to investigate the potential for positive associations between different predictors and our response variable. Here we use Kendall's τ as it is robust against outliers. The researchers ran correlation tests with Kendall's τ at a significance level $\alpha = 0.05$ to test the following hypotheses: the null hypothesis of no association (independence) $H_0 : \tau \leq 0$ and the alternative hypothesis of positive association $H_A : \tau > 0$. (They also tested for negative associations with the set of hypotheses $H_0 : \tau \geq 0$ and $H_A : \tau < 0$.)

They found that log_NDAM2014, log_BaseDam2014, LF.times, and LF.WindsMPH were all significantly positively associated with log_deaths. The researchers also found that LF.PressureMB was significantly negatively associated with log_deaths. These aggegated results can be found in the table below: *(STILL NEED TO ADD THIS TABLE)*

Model Fitting

OLS

The researchers next proceeded to fit models to the data, beginning with ordinary least squares regression. *still need to list assumptions here for OLS?* The researchers found that the best OLS model was the model predicting log_deaths with simply log_NDAM2014. *Note that the change of log(BaseDam2014) has meant that none of our two predictor models are significant*

Checking OLS Assumptions

note we still need to plot the empirical CDFs here

The researcher's checked the assumption of normally distributed residuals for our above selected linear regression model. To do so, the researchers performed a Kolmogrov Smirnov test using a significance level $\alpha = 0.05$:

Hypotheses:

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

$H_A : F(t) \neq F^*(t)$ 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.

Assumptions:

1. Continuous: The test assumes that the theoretical distribution is continuous, which is reasonable.
2. Independence: This is reasonable, as there's no reason to believe the log deaths for one hurricane would affect another.
3. Parameters: This is also satisfied, we have mean = 0, and sd = 1 for the normal distribution, and neither of these is estimated.

Test Statistic and P-value:

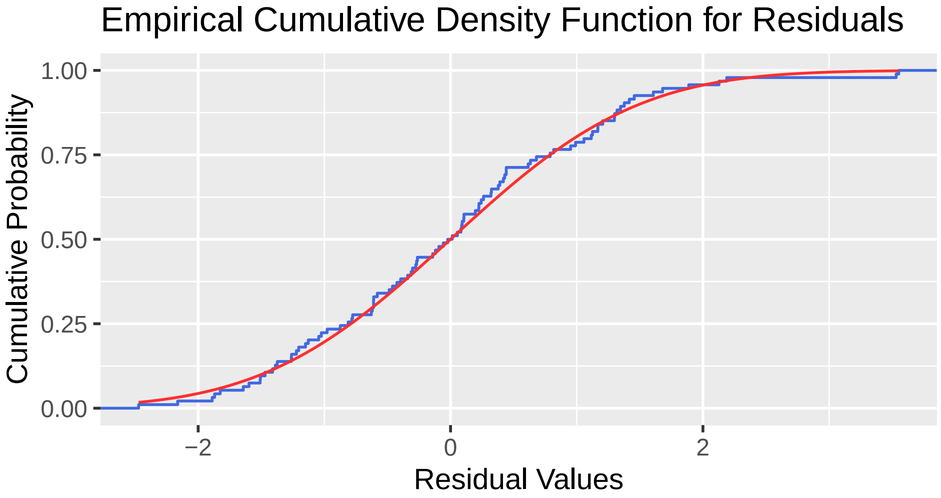
Decision:

We have a $p-$value of $0.7095 > 0.05$, so we fail to reject the null hypothesis at our significance level.

Conclusion:

We conclude that there is insufficient evidence to suggest that the estimated CDF of the distribution of the residuals in our linear model is significantly different to the CDF of the normal distribution. Thus, our assumptions for our linear model are met.

```r
df <- data.frame(values = resid(best_ols))

ggplot(df, aes(values)) + stat_ecdf(geom = "step", color = "royalblue") +
  stat_function(fun = pnorm, args = list(mean = mean(df$values),
                                         sd = sd(df$values)),
                                         color = "firebrick1") +
labs(title="Empirical Cumulative Density Function for Residuals",
     y = "Cumulative Probability", x="Residual Values")
```



JHM

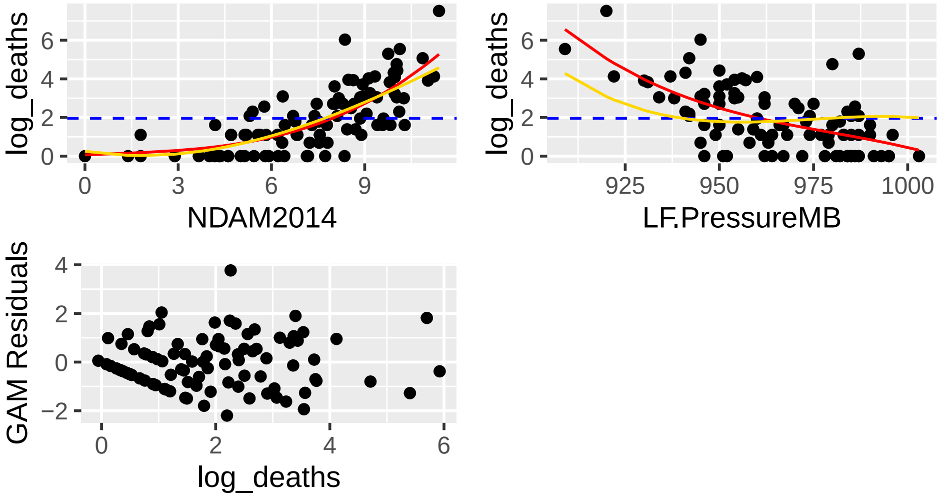
The researchers proceeded with a non-parametric approach to model fitting by using rank-based regression. They found that the best model predicting log_deaths was again one that used only log_NDAM2014. The researchers performed drop in dispersion tests when making this decision and found that the model containing only NDAM2014 was the best model.

GAM

Lastly, the researchers attempted to fit a generalized additive model (GAM) to the data. The researchers used manual forward selection to fit the best model. With an AIC of 290.8135, the researchers first found the best way to fit log_NDAM2014 was with a b-spline with the default number of degrees of freedom.

The researchers next found that the best way to fit LF.PressureMB was with an s-spline with 5 degrees of freedom (while keeping the previous b-spline for log_NDAM2014).

All other predictors increased the AIC, so the researchers decided on the GAM using a b-spline with the default number of degrees of freedom for log_NDAM2014, and an s-spline with 5 degrees of freedom LF.PressureMB. The researchers plotted this, but noticed that there was a poor fit for the LF.PressureMB scatterplot. The researchers noticed that the fit for LF.PressureMB appeared to be roughly quadratic, so attempted fitting the LF.PressureMB term with a polynomial. The final GAM can then be seen below: *note we still need to overlay the OLS and JHM lines from above I think*



we need to include motivation for adding back in the Presure variable since we took it out with our OLS and JHM models

Conclusion

We found that best way to predict the log number of deaths from a US hurricane that made landfall between 1950 and 2012 was to use a Generalized Additive Model using a b-spline with the default degrees of freedom on log 'NDAM2014', and a cubic polynomial for 'LF.PressureMB'. This makes sense—and matches our original hypothesis—since we would intuitively expect the number of deaths to increase as the severity (in this case the amount of hurricane damage) increases. We ultimately chose the GAM as our final model over the linear and rank based models we found due to the its lower AIC, its better graphical representation, and its ability to fit to local trends (since our predictors did not appear to be globally linear).

Limitations

Limitations to this report include shortcomings of the dataset, reliance on normalized metrics, and the numerous log transformations performed. While the dataset itself is comes from a reliable source, there is limited numerical data available for each hurricane. We started with a small number of total numeric predictors (six), and only two of which were related to the actual weather data of the hurricane (windspeed and pressure). We might have been able to obtain a more accurate model predicting the log number of deaths if we had access to more predictors pertaining to the severity of each hurricane, such as the radius size. Additionally, one of our predictors was normalized hurricane damage, which is necessary in order to compare hurricanes across a wide time range. However, this is a potential limitation as any inaccuracies with this metric would mean inaccuracies in our report. It is of note that this predictor was obtained from this article (add source here, link is https://ascelibrary.org/doi/abs/10.1061/(ASCE)1527-6988(2008)9:1(29)) but we still wish to note this potential limitation. Lastly, we rely heavily on the log transformations performed on numerous predictors throughout our report. A potential limitation is that we did not find the best possible log transformation for each predictor.