Group 2: STAT225 Final Project Proposal

Oliver Baldwin Edwards, Leah Johnson, Adi Arifovic

Monday, April 20, 2020

1) Group

Group 2

2) Members

Oliver Baldwin Edwards, Leah Johnson, Adi Arifovic

3) Title

Predicting Hurricane Deaths

4) Purpose

We hope to predict the 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.

5) Data

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.

Potential shortcomings of the dataset include limited data on weather-related variables for each hurricane, as it only contains maximum sustained windspeed, atmospheric pressure at landfall, and number of landfalls. Additionally, the "NDAM2014" column contains data on hurricane damage had the hurricane appeared in 2014 and it is unclear how these estimates were calculated (also no units are given on these observations).

```r
# load in full dataset
full_hurricane_df <- hurricNamed
glimpse(full_hurricane_df)
```

```
## Rows: 94
## Columns: 12
## $ Name           <chr> "Easy", "King", "Able", "Barbara", "Florence", "Caro...
## $ Year           <int> 1950, 1950, 1952, 1953, 1953, 1954, 1954, 1954, 1955...
## $ LF.WindsMPH    <int> 120, 130, 85, 85, 85, 120, 120, 145, 120, 85, 120, 1...
## $ LF.PressureMB  <int> 958, 955, 985, 987, 985, 960, 954, 938, 962, 987, 96...
## $ LF.times       <int> 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1...
## $ BaseDamage     <dbl> 3.3000, 28.0000, 2.7500, 1.0000, 0.2000, 460.2275, 4...
## $ NDAM2014       <dbl> 1870, 6030, 170, 65, 18, 21375, 3520, 28500, 2270, 1...
## $ AffectedStates <chr> "FL", "FL", "SC", "NC", "FL", "NC,NY,CT,RI", "MA,ME"...
## $ firstLF        <date> 1950-09-04, 1950-10-17, 1952-08-30, 1953-08-13, 195...
## $ deaths         <int> 2, 4, 3, 1, 0, 60, 20, 20, 0, 200, 7, 15, 416, 1, 0,...
## $ mf             <fct> f, m, m, f, f, f, f, f, f, f, m, f, f, f, f, f, f, f...
## $ BaseDam2014    <dbl> 32.419419, 275.073859, 24.569434, 8.867416, 1.773483...
```

6) Population

Each observational unit contains data on one named hurricane that made landfall in the US between 1950 and 2012. Since we are trying to predict the number of deaths that occur from a hurricane based on this data, our predictions would generalize to hurricanes that occurred outside of the time range from 1950 to 2012 in the US (noting that we might not be able to generalize across different time periods).

7) Response Variable(s)

Our response variable is the 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.

8) Explanatory Variables

We will examine LF.WindsMPH, LF.PressureMB, LF.times, BaseDam2014, NDAM2014, and deaths.

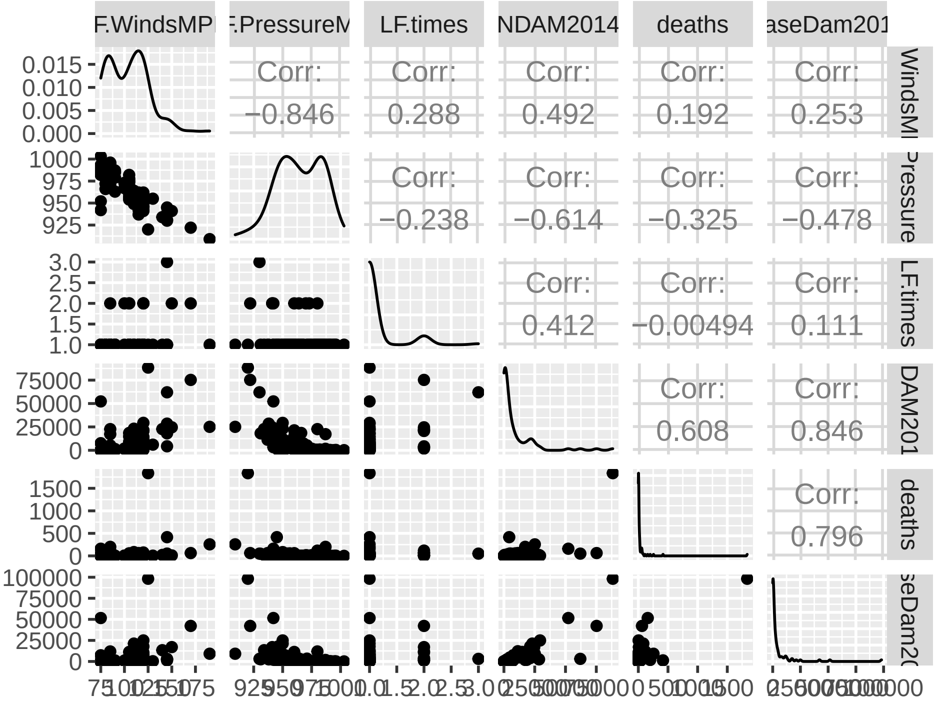
1. LF.WindsMPH describes the maximum sustained windspeed for each hurricane in miles per hour
2. LF.PressureMB describes the atmospheric pressure at landfall in millibars
3. LF.times describes the number of times the hurricane made landfall
4. BaseDam2014 describes the property damage caused by the hurricane 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

9) Exploratory Analysis

The below ggpairs is for the full, unaltered dataset:

```r
# remove non numeric and BaseDamage for ggpairs call
hurricane_pairs <- full_hurricane_df %>%
  select(-c(Name, AffectedStates, firstLF, mf, BaseDamage, Year))

ggpairs(hurricane_pairs)
```



A model predicting deaths by pressure and windspeed from the unaltered dataset:

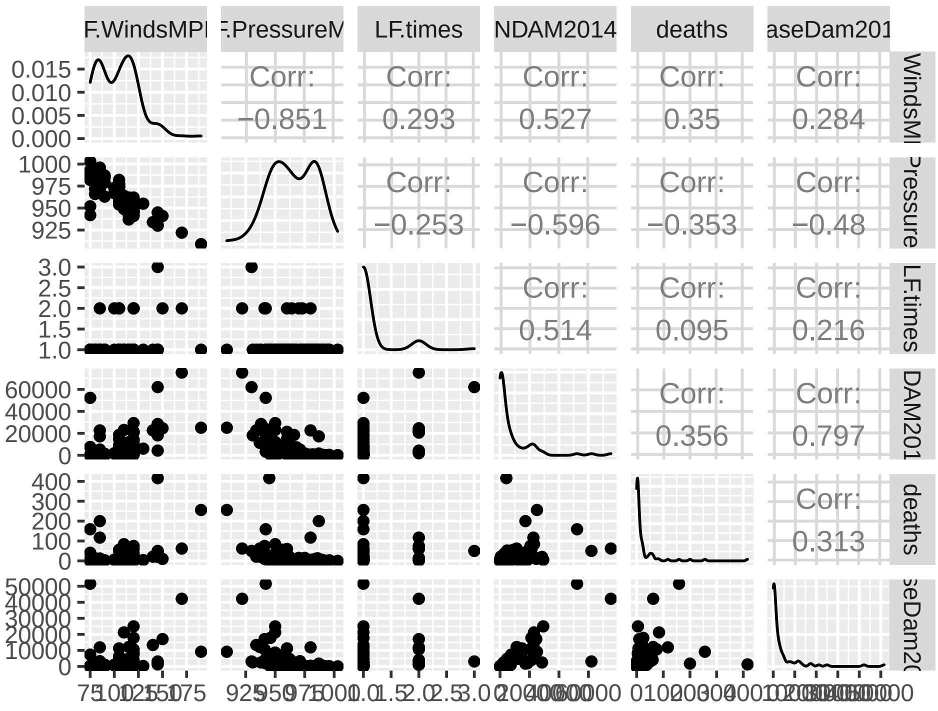
```r
# predicting deaths by pressure and windspeed with unadjusted death column
msummary(lm(deaths ~ LF.PressureMB + LF.WindsMPH, data = hurricane_pairs))
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5738.784   1893.651   3.031  0.00318 **
## LF.PressureMB    -5.632      1.815  -3.103  0.00255 **
## LF.WindsMPH      -2.512      1.594  -1.576  0.11856
##
## Residual standard error: 184.4 on 91 degrees of freedom
## Multiple R-squared:  0.1292, Adjusted R-squared:   0.11
## F-statistic: 6.749 on 2 and 91 DF,  p-value: 0.00185
```

Now we run a ggpairs call removing the one death outlier:

```r
# try removing death outlier
remove_outlier <- hurricane_pairs %>%
  filter(deaths != 1836)

ggpairs(remove_outlier)
```



And run a model using the dataset with the removed death outlier:

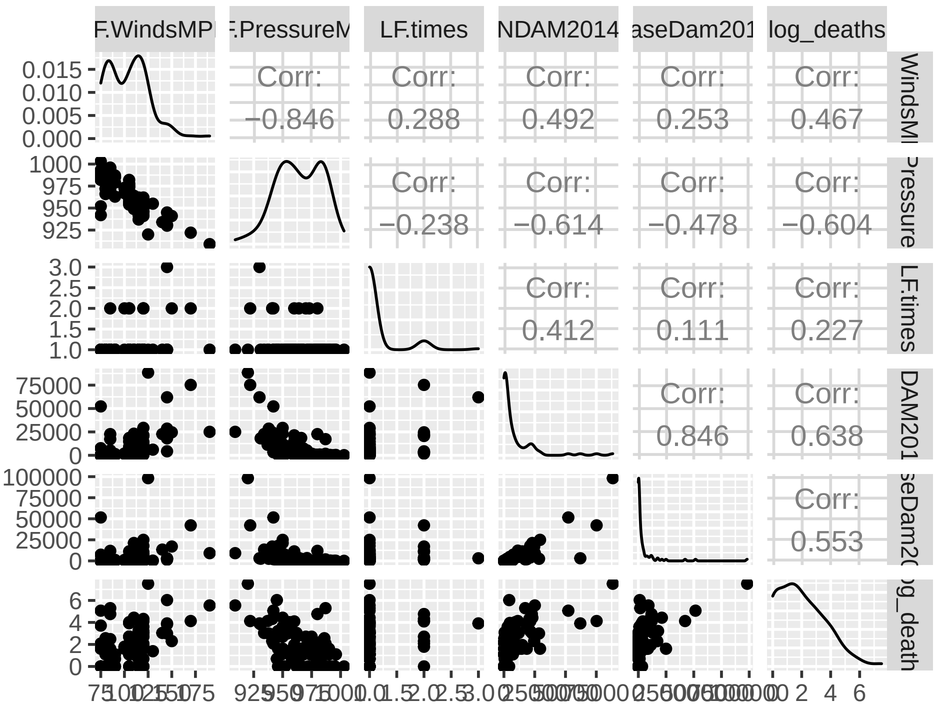
```
# predicting deaths by pressure and windspeed with removed death outlier
msummary(lm(deaths ~ LF.PressureMB + LF.WindsMPH, data = remove_outlier))
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    552.7578   583.0183   0.948    0.346
## LF.PressureMB   -0.5971     0.5594  -1.067    0.289
## LF.WindsMPH      0.4615     0.4799   0.962    0.339
##
## Residual standard error: 54.37 on 90 degrees of freedom
## Multiple R-squared:  0.1332, Adjusted R-squared:  0.1139
## F-statistic: 6.915 on 2 and 90 DF,  p-value: 0.001609
```

Lastly we apply a log scale to the deaths column and run a ggpairs call:

```
# try applying a log transformation to deaths
hurricane_logdeaths <- hurricane_pairs %>%
  mutate(log_deaths = ifelse(deaths == 0, 0, log(deaths))) %>%
  select(-deaths)

ggpairs(hurricane_logdeaths)
```



And start examining a few different models using this dataset with log(deaths):

```
# different models to predict log deaths
pressure_windspeed <- msummary(lm(log_deaths ~ LF.PressureMB + LF.WindsMPH,
                                  data = hurricane_logdeaths)) ;pressure_windspeed
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.56484   13.70080   4.567 1.55e-05 ***
## LF.PressureMB -0.06163    0.01313  -4.693 9.45e-06 ***
## LF.WindsMPH   -0.01120    0.01154  -0.971    0.334
##
## Residual standard error: 1.334 on 91 degrees of freedom
## Multiple R-squared:  0.3708, Adjusted R-squared:  0.357
## F-statistic: 26.81 on 2 and 91 DF,  p-value: 6.994e-10
```

```
pressure_ndam <- msummary(lm(log_deaths ~ LF.PressureMB + NDAM2014,
                                  data = hurricane_logdeaths)) ;pressure_ndam
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.916e+01  7.837e+00   3.721 0.000343 ***
## LF.PressureMB -2.862e-02  8.070e-03  -3.546 0.000620 ***
## NDAM2014       4.668e-05  1.040e-05   4.489 2.09e-05 ***
##
## Residual standard error: 1.213 on 91 degrees of freedom
## Multiple R-squared:  0.4795, Adjusted R-squared:  0.4681
## F-statistic: 41.92 on 2 and 91 DF,  p-value: 1.248e-13
```

```
pressure_basedam <- msummary(lm(log_deaths ~ LF.PressureMB + BaseDam2014,
                                  data = hurricane_logdeaths)) ;pressure_basedam
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.747e+01  7.187e+00   5.213 1.15e-06 ***
## LF.PressureMB -3.705e-02  7.424e-03  -4.991 2.87e-06 ***
## BaseDam2014    4.514e-05  1.161e-05   3.888 0.000192 ***
```

```
##
## Residual standard error: 1.242 on 91 degrees of freedom
## Multiple R-squared:  0.4548, Adjusted R-squared:  0.4428
## F-statistic: 37.96 on 2 and 91 DF,  p-value: 1.029e-12
```

We see at least a couple of interesting relationships between deaths / log(deaths) and predictor variables from the above ggpairs calls and models.