Analysis of the Impact of US Storms - 1950 to 2011

August 21, 2018

Synopsis

This analysis looks at the impact of US Storms on Population Health (Deaths and Injuries) and on Financial Cost (Property and Crop Damage). Time is an important consideration due to geographical changes in farming, property type/location and indeed on weather patterns.

Raw Empirical analysis shows typcially Tornado and Flood style events are most impactful on Health and Property, but more recently hot weather impacts are more prevalent such as Heat and Wildfire.

Data Processing

Data provided have Injuries, Deaths, Crop Damages and Propoerty Damages (measured in $/millions) and a variety of other location data.

Data collection and FAQ notes are available below: https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2Fpd01016005curr.pdf

https://d396qusza40orc.cloudfront.net/repdata%2Fpe-er2_doc%2FNCDC%20Storm%20Events-FAQ%20Page.pdf

The data can be downloaded here:
https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2

The data are obtained and extracted below with some exploration:

```
setwd("C:/Nick/07 R/6JohnHopkins/5 Reproducible Research/Assignment2")
stormData<-read.csv("repdata%2Fdata%2FStormData.csv.bz2")
n<-nrow(stormData)  #902,297 records
t<-data.frame(table(stormData$EVTYPE))
nt<-nrow(t)   #985 event types, need to group
head(stormData) #need to select and process some fields too
```

```
##   STATE__           BGN_DATE BGN_TIME TIME_ZONE COUNTY COUNTYNAME STATE
## 1       1  4/18/1950 0:00:00     0130       CST     97     MOBILE    AL
## 2       1  4/18/1950 0:00:00     0145       CST      3    BALDWIN    AL
## 3       1  2/20/1951 0:00:00     1600       CST     57    FAYETTE    AL
## 4       1   6/8/1951 0:00:00     0900       CST     89    MADISON    AL
## 5       1 11/15/1951 0:00:00     1500       CST     43    CULLMAN    AL
## 6       1 11/15/1951 0:00:00     2000       CST     77 LAUDERDALE    AL
##    EVTYPE BGN_RANGE BGN_AZI BGN_LOCATI END_DATE END_TIME COUNTY_END
## 1 TORNADO         0                                               0
## 2 TORNADO         0                                               0
## 3 TORNADO         0                                               0
## 4 TORNADO         0                                               0
## 5 TORNADO         0                                               0
## 6 TORNADO         0                                               0
##   COUNTYENDN END_RANGE END_AZI END_LOCATI LENGTH WIDTH F MAG FATALITIES
## 1         NA         0                      14.0   100 3   0          0
## 2         NA         0                       2.0   150 2   0          0
## 3         NA         0                       0.1   123 2   0          0
## 4         NA         0                       0.0   100 2   0          0
```

```
## 5         NA         0                      0.0   150 2   0          0
## 6         NA         0                      1.5   177 2   0          0
##   INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP WFO STATEOFFIC ZONENAMES
## 1       15    25.0          K       0
## 2        0     2.5          K       0
## 3        2    25.0          K       0
## 4        2     2.5          K       0
## 5        2     2.5          K       0
## 6        6     2.5          K       0
##   LATITUDE LONGITUDE LATITUDE_E LONGITUDE_ REMARKS REFNUM
## 1     3040      8812       3051       8806              1
## 2     3042      8755          0          0              2
## 3     3340      8742          0          0              3
## 4     3458      8626          0          0              4
## 5     3412      8642          0          0              5
## 6     3450      8748          0          0              6
```

Some fields will need pre-processing including
- Multiplying Crop and Property values which are in mixed precision (thousands, millions, billions)
- Creating modellable Dates from the machine date format (event year, banded-10 event year, event Month|Year)
- Creating the "propCost"" variable as the sum of Crop and Property damages to allow single analysis
- Creating the "healthCost"" variable as the amalgam of Injury and Death variables to allow single analysis. For the purposes of an initial view, the spurious multiple of 10x has been applied to the Deaths to give them relatively more importance in the results as compared to Injuries. This is not a statement of fact but a necessary choice for a single analysis

```
stormData$CROPDMG2 <- ifelse(toupper(stormData$CROPDMGEXP)=="K", stormData$CROPDMG/1000, ifelse(toupper
stormData$PROPDMG2 <- ifelse(toupper(stormData$PROPDMGEXP)=="K", stormData$PROPDMG/1000, ifelse(toupper
stormData$axDate <- as.Date(stormData$BGN_DATE,'%m/%d/%Y')
stormData$axYear <- as.numeric(format(stormData$axDate, '%Y'))
stormData$axYearBand<-stormData$axYear-(stormData$axYear-10*floor(stormData$axYear/10)) #mod10 function
stormData$axMonth <- paste(format(stormData$axDate, '%m'),format(stormData$axDate, '%Y'),sep=" ")
stormData$healthCost <- (stormData$FATALITIES*10+stormData$INJURIES)
stormData$propCost <- (stormData$PROPDMG2+stormData$CROPDMG2)
```

The EVTYPE variable is very rich but contains maybe 900+ individual event types, some of which confound and would make finding a concrete conclusion tricky.

The below code would output the frequency against the words in the event description in order to identify some fundamental groups. These could be clustered using some tree basis and/or a text-mining package. In this case, some words were selected as below to exclude and then the main events were grouped into some reasonable functional groups. The variables and the code to analyse the raw data can be reproduced below.

The first step is to gain an intial list of words and remove anything not-particularly useful in the top 100 words. Once complete, present the frquency again and identify words to group:

```
#clean the EVTYPE field into groups
t1<-toupper(stormData$EVTYPE)
t1 <- strsplit(t1, " ", fixed = TRUE); t1 <- unlist(t1);
#t1[1:100,]   #identify some words to drop.  This has been excluded due to size but could easily be run
strip<-c("FLASH","HEAVY","HIGH","WILD/FOREST","MARINE","WEATHER","CLOUD","STRONG","URBAN/SML","EXTREME"
t2 <- t1[!t1 %in% strip]
t2 <- table(t2);
t2<-as.data.frame(t2);
names(t2)=c("WORD", "FREQ");
t2 <- t2[order(-t2$FREQ),]
```

```
#t2[1:100,] #identify synonyms from the largest word frequencies to help map the fields together.  This
```

The functional groups can be created by searching for the words in the full dataset. The dataset is separated to ensure a distinct mapping is produced.

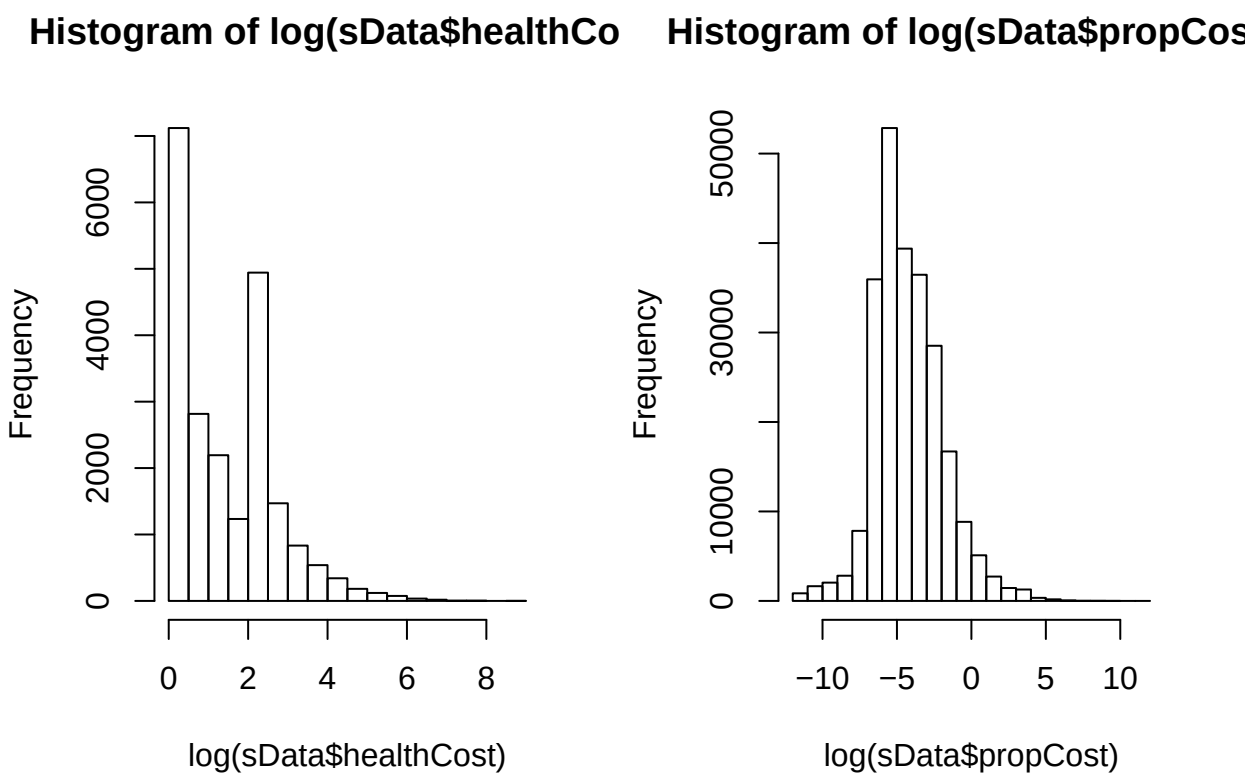
The "event" key is added before the set is appended and slimmed down to relevant modeling variables.

A quick summary is provided to show the initial view of impactful datasets by Health impact and Crop Damage.

```
#now create the final mapping on the main dataset
c1<-c("GUSTY|WIND|WINDS|WINDSS|HURRICANE")
c2<-c("HAIL|WIND/HAIL|WINDS/HAIL")
c3<-c("DOWNBURST|SQUALLS|MICROBURST|LIGHTING|THUNDERSTORM|TSTM|LIGHTNING|STORM|RAIN|PRECIPITATION|RAINS
c4<-c("FLOOD|FLOODING|FLD|FLOOD/FLOOD|FLOOD/RAIN/WINDS|FLOODIN|FLOODS|FLOODING/FLOOD|SEICHE")
c5<-c("TORNADO|WATERSPOUT|FUNNEL|FUNNELS|TYPHOON|WATERSPOUTS|WATERSPOUT/TORNADO|WATERSPOUT/|WATERSPOUT-
c6<-c("SNOW|BLIZZARD|SNOW/BLOWING|SNOW/FREEZING|SNOW/HIGH|SNOW/ICE|SNOW/SLEET|SNOW/SLEET/FREEZING|RAIN/S
c7<-c("COLD|ICE|CHILL|COLD/WIND|FROST/FREEZE|WINTER|FREEZING|COOL|ICY|HYPOTHERMIA/EXPOSURE|FREEZE|WINTR
c8<-c("WILDFIRE|FIRE|FIRES|WILDFIRES")
c9<-c("HIGH TEMPERATURE|HIGH TEMPERATURES|HOT|WARM|WARMTH|HEAT")
c10<-c("DROUGHT|DROUGHT/EXCESSIVE|DRYNESS")
c11<-c("WAVES|CURRENTS|TIDE|SURF/HIGH|WAVE|LANDSLUMP|TIDES|SURF|CURRENTS/HEAVY|TIDAL|SWELLS|TSUNAMI")
c12<-c("LANDSLIDE|MUDSLIDES|LANDSLIDES|MUDSLIDE")
c13<-c("AVALANCHE|AVALANCE|AVALANCH")
c14<-c("FOG")

sD1<-stormData[grepl(c1, stormData$EVTYPE),];res<-stormData[!grepl(c1, stormData$EVTYPE),]
sD2<-res[grepl(c2, res$EVTYPE),];res<-res[!grepl(c2, res$EVTYPE),]
sD3<-res[grepl(c3, res$EVTYPE),];res<-res[!grepl(c3, res$EVTYPE),]
sD4<-res[grepl(c4, res$EVTYPE),];res<-res[!grepl(c4, res$EVTYPE),]
sD5<-res[grepl(c5, res$EVTYPE),];res<-res[!grepl(c5, res$EVTYPE),]
sD6<-res[grepl(c6, res$EVTYPE),];res<-res[!grepl(c6, res$EVTYPE),]
sD7<-res[grepl(c7, res$EVTYPE),];res<-res[!grepl(c7, res$EVTYPE),]
sD8<-res[grepl(c8, res$EVTYPE),];res<-res[!grepl(c8, res$EVTYPE),]
sD9<-res[grepl(c9, res$EVTYPE),];res<-res[!grepl(c9, res$EVTYPE),]
sD10<-res[grepl(c10, res$EVTYPE),];res<-res[!grepl(c10, res$EVTYPE),]
sD11<-res[grepl(c11, res$EVTYPE),];res<-res[!grepl(c11, res$EVTYPE),]
sD12<-res[grepl(c12, res$EVTYPE),];res<-res[!grepl(c12, res$EVTYPE),]
sD13<-res[grepl(c13, res$EVTYPE),];res<-res[!grepl(c13, res$EVTYPE),]
sD14<-res[grepl(c14, res$EVTYPE),];sD15<-res[!grepl(c14, res$EVTYPE),]


#put the indicator and re-group (we separated to avoid overlapping mappings)
sD1$event<-1;sD2$event<-2;sD3$event<-3;sD4$event<-4;sD5$event<-5;sD6$event<-6;
sD7$event<-7;sD8$event<-8;sD9$event<-9;sD10$event<-10;sD11$event<-11;
sD12$event<-12;sD13$event<-13;sD14$event<-14;sD15$event<-15
sData<-rbind(sD1, sD2, sD3, sD4, sD5, sD6, sD7, sD8, sD9, sD10, sD11, sD12, sD13, sD14, sD15)
sData<-sData[,c("STATE__","axYear","axYearBand","axMonth","healthCost","propCost","event","INJURIES","F

#costs for property and health.
#tough to value a human life verses human injury, but in principle death is a greater tragedy
#i've applied some spurious multiple of 10x to help support the evealuation
s1<-aggregate(sData$healthCost, by=list(Category=sData$event), FUN=sum)
s2<-aggregate(sData$propCost, by=list(Category=sData$event), FUN=sum)
```

Initial plots of the data show it is quite spread which is typical of naturally occuring data. Instead, I will output the log charts which look more even. This suggests analysis/modelling should be done in log-space.

```
#initial plots showed the data is quite spread whcih is typical of naturally occuring data
#par(mfrow=c(2,1))
#hist(sData$healthCost)
#hist(sData$propCost)

#move to a log plot shows a better picture which highlights that the models should be on that
#basis too
par(mfrow=c(1,2))
hist(log(sData$healthCost))
hist(log(sData$propCost))
```



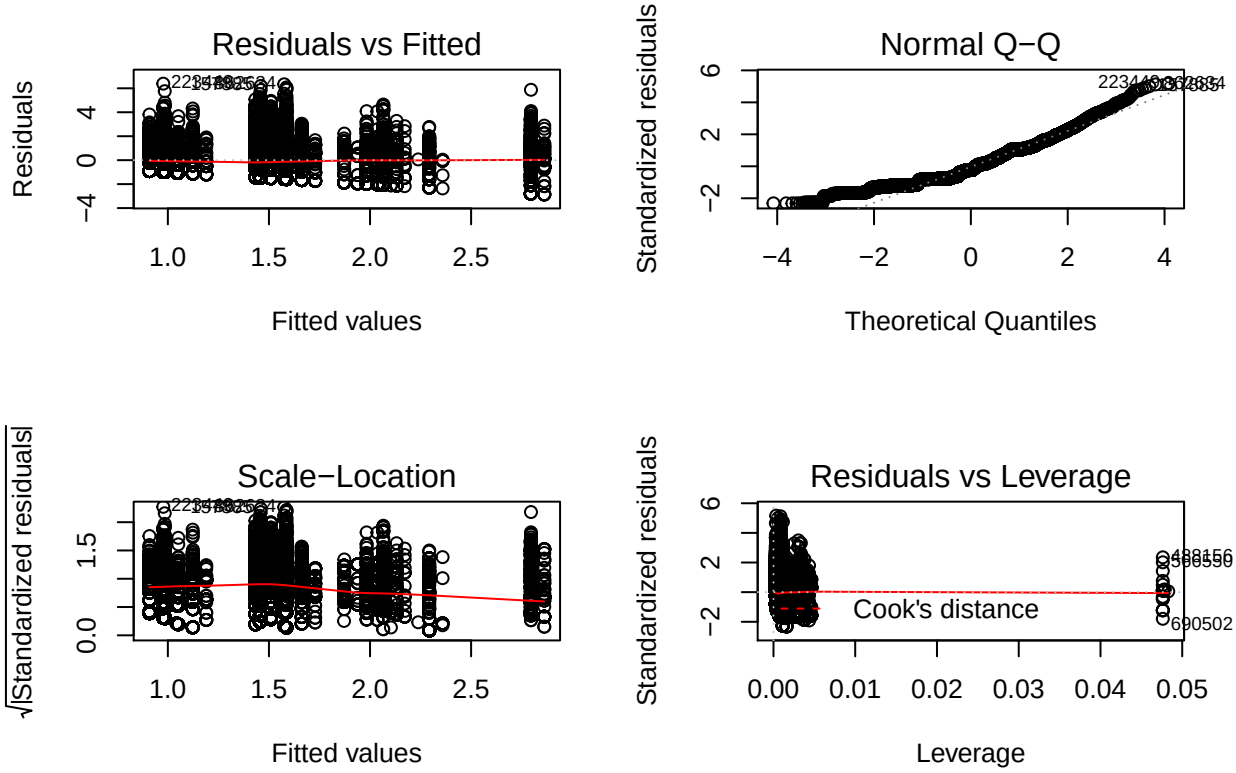
Now the data is in good shape, excluding zero responses the models are applied. Reviewing the output and the outliers iteratively, there were additional groupings required to obtain results. There are still some outliers but given the extreme nature of events I believe this is necessary.

The final models show mostly credible event groups (certainly the higher risk groups were significant) and the banded event year (axYearBanded) look reasonable.

```
x1<-sData$healthCost==0
x2<-sData$propCost==0
#only 1 obs in Drought (event==10) so group with "other" 15
temp1<-sData[!x1,]
temp1$event<-ifelse(temp1$event==10|temp1$event==14|temp1$event==2, 15, temp1$event)
m4<-lm(log(healthCost)~factor(axYearBand)+factor(event),data=temp1)
summary(m4)
```

```
##
## Call:
## lm(formula = log(healthCost) ~ factor(axYearBand) + factor(event),
##     data = temp1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8619 -0.9773 -0.2842  0.8979  6.3866
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.06313    0.04582  23.205  < 2e-16 ***
## factor(axYearBand)1960 -0.13190    0.05121  -2.576  0.01000 *
## factor(axYearBand)1970 -0.12746    0.04765  -2.675  0.00748 **
## factor(axYearBand)1980 -0.15406    0.04742  -3.249  0.00116 **
## factor(axYearBand)1990 -0.07711    0.04463  -1.728  0.08406 .
## factor(axYearBand)2000 -0.07844    0.04489  -1.747  0.08059 .
## factor(axYearBand)2010 -0.01138    0.05196  -0.219  0.82661
## factor(event)3         -0.00873    0.02670  -0.327  0.74365
## factor(event)4          1.07966    0.03707  29.121  < 2e-16 ***
## factor(event)5          0.52386    0.02670  19.616  < 2e-16 ***
## factor(event)6          0.99722    0.07382  13.508  < 2e-16 ***
## factor(event)7          1.30714    0.07733  16.904  < 2e-16 ***
## factor(event)8          0.13813    0.07035   1.964  0.04959 *
## factor(event)9          1.81011    0.04406  41.085  < 2e-16 ***
## factor(event)11         1.05162    0.06471  16.250  < 2e-16 ***
## factor(event)12         1.18677    0.27077   4.383 1.18e-05 ***
## factor(event)13         0.88677    0.08209  10.803  < 2e-16 ***
## factor(event)15         0.67691    0.04447  15.221  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.238 on 21911 degrees of freedom
## Multiple R-squared:  0.1266, Adjusted R-squared:  0.1259
## F-statistic: 186.8 on 17 and 21911 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(m4)
```

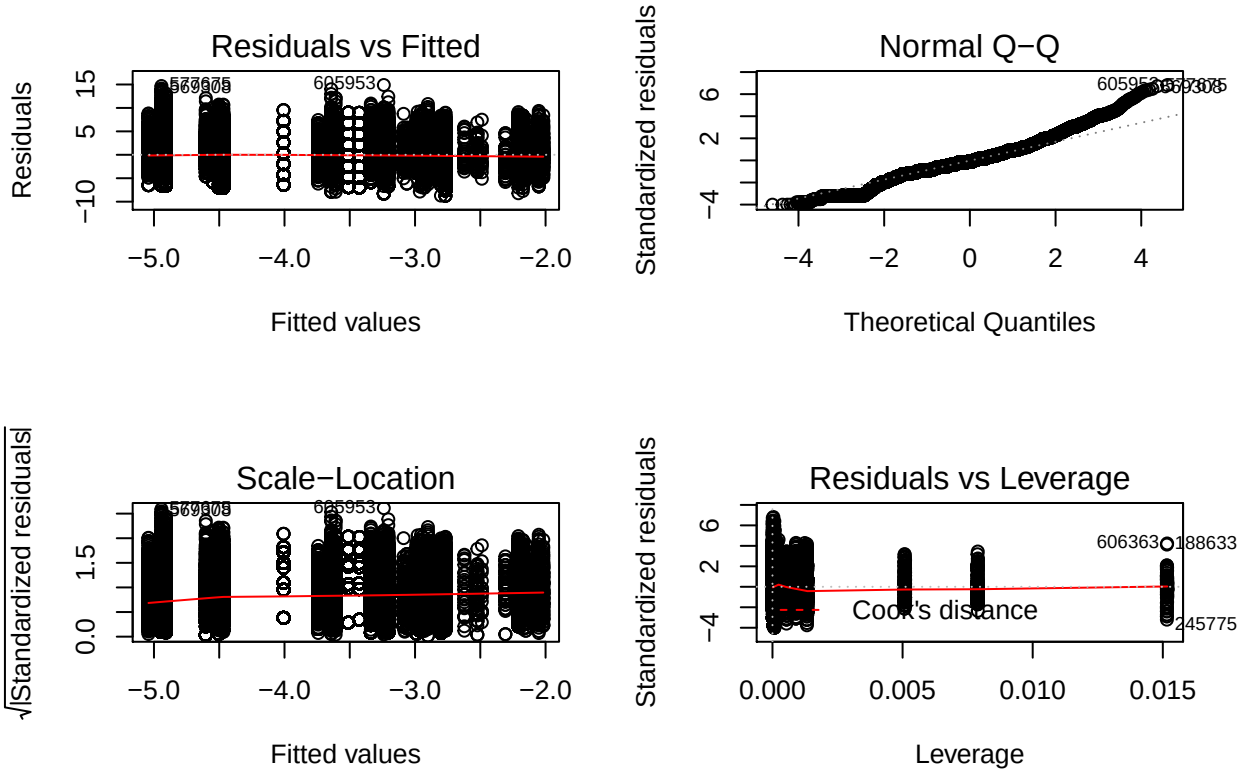



```
#outlier in 1950 only now  #634122 (Drought variable is most likely)
temp2<-sData[!x2,]
temp2$event<-ifelse(temp2$event==13|temp2$event==10|temp2$event==14, 15, temp2$event)
#group avalanche, fog 14 and Avalanche 13 with other 15
m5<-lm(log(propCost)~factor(axYearBand)+factor(event),data=temp2)
summary(m5)
```

```
##
## Call:
## lm(formula = log(propCost) ~ factor(axYearBand) + factor(event),
##     data = temp2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7456 -1.3060 -0.2562  1.2194 14.8920
##
## Coefficients:
##                           Estimate Std. Error  t value Pr(>|t|)
## (Intercept)               -6.68798    0.03817 -175.210  < 2e-16 ***
## factor(axYearBand)1960     0.53953    0.04496   12.000  < 2e-16 ***
## factor(axYearBand)1970     1.03558    0.04247   24.382  < 2e-16 ***
## factor(axYearBand)1980     1.11994    0.04274   26.205  < 2e-16 ***
## factor(axYearBand)1990     1.77966    0.03810   46.706  < 2e-16 ***
## factor(axYearBand)2000     1.74534    0.03802   45.904  < 2e-16 ***
## factor(axYearBand)2010     1.64586    0.03923   41.955  < 2e-16 ***
## factor(event)2             0.43948    0.01490   29.495  < 2e-16 ***
```

```
## factor(event)3          1.30035    0.01916   67.852  < 2e-16 ***
## factor(event)4          1.70357    0.01354  125.777  < 2e-16 ***
## factor(event)5          2.14102    0.01829  117.035  < 2e-16 ***
## factor(event)6          1.97931    0.05082   38.945  < 2e-16 ***
## factor(event)7          1.95435    0.07686   25.428  < 2e-16 ***
## factor(event)8          2.89073    0.06580   43.935  < 2e-16 ***
## factor(event)9          1.96270    0.26897    7.297 2.95e-13 ***
## factor(event)11         2.42003    0.19393   12.479  < 2e-16 ***
## factor(event)12         1.99821    0.15577   12.828  < 2e-16 ***
## factor(event)15         2.73430    0.07959   34.356  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.184 on 245013 degrees of freedom
## Multiple R-squared:  0.1195, Adjusted R-squared:  0.1195
## F-statistic:  1957 on 17 and 245013 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(m5)
```



Results

Q1. Population Health

Empirically {WIND, HAIL, THUNDERSTORM} are the primpary impacts on the Health Impact score I created. When the period of time is taken into account (using log-linear-model m4) these are less important

and {HEAT, COLD, LANDSLIDE} become the most impactful events.

Q2. Property Damage

Once again, the empirical results differ from the model with {FLOOD, WIND, THUNDERSTORM} as the major imapcts on a one way. With the time effect (from log-linear model m5) we see that {WILDFIRE, OCEAN, TORNADO} are the most impactful.

General Summary

Heat and Wildfire are the more impactful weather events in the US in recent years. The analysis is fairly basic at this stage but it suggests to clear courses of action: 1. Working with the data collection teams to create a better functional mapping of the event types (potentially with more simple/clearer groupings to aid analysis or a more enhanced text-mining approach) 2. Consideration of climate change impacts on the event types with the most risk to see if there are socio-demographic impacts and resourcing decisions that could be considered.