Report 2

Wei Zhong

4/3/2017

Introduction

A tropical cyclone is a rotating, organized system of clouds and thunderstorms that originates over tropical or subtropical waters and has a closed low-level circulation. Tropical cyclones rotate counterclockwise in the Northern Hemisphere. They are classified as four types inlcuding tropical depression, tropical storms, hurricane and major hurricane. More specifically, tropical depression is a tropical cyclone with maximum sustained winds of 38 mph (33 knots) or less. Tropical Storm means a tropical cyclone with maximum sustained winds of 39 to 73 mph (34 to 63 knots). And hurricane is a tropical cyclone with maximum sustained winds of 74 mph (64 knots) or higher. In the western North Pacific, hurricanes are called typhoons; similar storms in the Indian Ocean and South Pacific Ocean are called cyclones. Finally, major hurricane is a tropical cyclone with maximum sustained winds of 111 mph (96 knots) or higher, corresponding to a Category 3, 4 or 5 on the Saffir-Simpson Hurricane Wind Scale.[1]

Traditionally, areas of tropical cyclone formation are divided into seven basins. These include the north Atlantic Ocean, the eastern and western parts of the northern Pacific Ocean, the southwestern Pacific, the southwestern and southeastern Indian Oceans, and the northern Indian Ocean (Arabian Sea and Bay of Bengal). The western Pacific is the most active and the north Indian the least active. An average of 86 tropical cyclones of tropical storm intensity form annually worldwide, with 47 reaching hurricane or typhoon strength, and 20 becoming intense tropical cyclones.[2]

Specifically, the Atlantic hurricane season runs from June 1st to November 30th, and the Eastern Pacific hurricane season runs from May 15th to November 30th. The Atlantic basin includes the Atlantic Ocean, Caribbean Sea, and Gulf of Mexico. The Eastern Pacific basin extends to 140°W.[3]

The El Niño-Southern Oscillation (ENSO) is a naturally occurring phenomenon that involves fluctuating ocean temperatures in the equatorial Pacific. The warmer waters essentially slosh, or oscillate, back and forth across the Pacific, much like water in a bath tub. For North America and much of the globe, the phenomenon is known as a dominant force causing variations in regional climate patterns. The pattern generally fluctuates between two states: warmer than normal central and eastern equatorial Pacific SSTs (El Niño) and cooler than normal central and eastern equatorial Pacific SSTs (La Niña). Often, sea surface temperatures (SSTs) are used to identify this oscillation, but it is important to understand that changes in sub-surface ocean temperatures are the first to respond to an oncoming change in the ENSO phase.[4]

Additionally, the ONI is based on SST departures from average in the Niño 3.4 region, and is a principal measure for monitoring, assessing, and predicting ENSO. Defined as the three-month running-mean SST departures in the Niño 3.4 region. Departures are based on a set of improved homogeneous historical SST analyses. It is one index that helps to place current events into a historical perspective.[5]

Research Question

Based on the data we have, first, I would like to examine what factors might explain the total number of tropical cyclones. Second, I tend to analyze what might predict the probability that a named tropical cyclone becomes a major hurricane.

[1] http://www.nhc.noaa.gov/climo/
[2] https://en.wikipedia.org/wiki/Tropical_cyclone_basins
[3] http://www.nhc.noaa.gov/climo/
[4] http://climate.ncsu.edu/climate/patterns/ENSO.html
[5] http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/lanina/enso_evolution-status-fcsts-web.pdf

Data

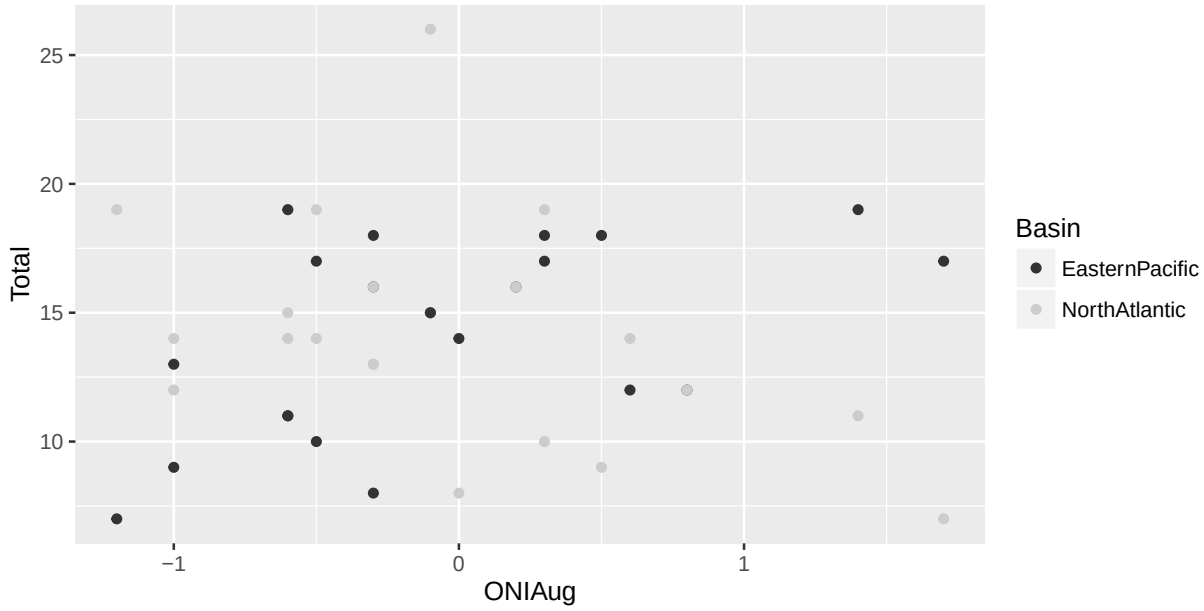
The cyclone data includes 44 observations and 7 variables including season from 1995 to 2016, basin including north Atlantic and aastern Pacific oceans, the number of tropical storms, the number of hurricane, the number of major hurricane, the total number of tropical storms and ONI values. In the following, table 1 shows the summary of the cyclone data. Table 2 displays marginal totals (summed over all seasons) of each category for each basin. Also Figure 1 represents a plot of total number of storms versus August ONI, for each basin separately.

Table 1: Statistical summary of the cyclone data

Statistic	N	Mean	St. Dev.	Min	Max
Season	44	2,005.500	6.418	1,995	2,016
TropicalStorm	44	6.682	2.639	1	12
Hurricane	44	4.023	1.861	1	8
MajorHurricane	44	3.545	2.151	0	10
Total	44	14.250	3.989	7	26
ONIAug	44	−0.059	0.733	−1.200	1.700

Table 2: Marginal totals of each category for each basin (summed over all seasons)

	Category	TropicalStorm	Hurricane	MajorHurricane	Total
1	EasternPacific	142	88	82	312
2	NorthAtlantic	152	89	74	315



Analysis of Total Named Storms

In this section, I fit a poisson loglinear regression for the total number of named storms, with a term for season (numeric variable), basin (indicator variable), August ONI and the interaction between basin and August ONI.

The form of the equation for the linear predictor is

$$\log E[\text{Total}_i|x_i] = \alpha + \beta_1 \text{Season} + \beta_2 \text{Basin} + \beta_3 \text{ONI} + \beta_4 \text{Basin} \times \text{ONI}$$

where, x_i is the i-th observation for the four variables including season, basin, ONI and the interaction between basin and ONI, α is the intercept and β_i is coefficient for each predictor.

The results of the loglinear model are shown in Table 3. For the goodness of fit, I use residual deviance, which is a measure of how well the model fits the data if the model fits well, the observed values will be close to their predicted means, causing the deviance to be small. According to the result of poisson loglinear regression, the residual deviance is 35.653 with 39 degrees of freedom. To calculate the p-value for the deviance goodness of fit test I simply calculate the probability to the right of the deviance value for the chi-squared distribution on 39 degrees of freedom. And the resulting p-value is 0.6233406. Thus we do not have strong evidence to reject the null hypothesis. In other words, the null hypothesis (i.e. the model) is not rejected. The fitted values are not significantly different from the observed values.

For the north Atlantic basin, the multiplicative effect of a one-unit increase in August ONI on the mean number of named storms can be calculated as

$$e^{0.1694707 - 0.3693889} = 0.8187977$$

For the eastern Pacific basin, the multiplicative effect of a one-unit increase in August ONI on the mean number of named storms can be calculated as

$$e^{0.1694707} = 1.184678$$

By comparing the multiplicative effects for two basins, we can conclude: for eastern Pacific basin, 1-unit increase in ONI yields an 18.5% increase in the estimated mean number of the named storms, while for north Atlantic basin, 1-unit increase in ONI yield 18.1% less in the estimated mean.

Table 3: Results of loglinear regression

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-16.0180	12.6393	-1.27	0.2050
Season	0.0093	0.0063	1.48	0.1396
factor(Basin)NorthAtlantic	-0.0139	0.0809	-0.17	0.8641
ONIAug	0.1695	0.0754	2.25	0.0246
factor(Basin)NorthAtlantic:ONIAug	-0.3694	0.1116	-3.31	0.0009

Analysis of Proportion of Major Hurricanes

I start with the model that has full three-way interaction between season, basin and ONI. Applying backward elimination, I get a model with season, basin, ONI, interaction between season and basin and interaction between basin and ONI. The results are given in Table 4.

Based on the estimates of this model, I estimate the probability that a named storm becomes a major hurricane for the 2017 season under different ONI values for each basin. The results are shown as in Table 5.

Further Analyese

For loglinear regression:

For each basin, I form a Wald 95% confidence interval for the multiplicative effect of a 1-unit increase in August ONI on the mean number of named storms. From the estimated asymptotic covariance matrix which is shown in Table 6, we have standard errors for ONI estimates for each basin.

Table 4: Result of logistic regression using backward elimination

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-52.6027	41.4799	-1.27	0.2047
Season	0.0257	0.0207	1.24	0.2140
factor(Basin)NorthAtlantic	132.1446	60.2681	2.19	0.0283
ONIAug	0.3757	0.1756	2.14	0.0324
Season:factor(Basin)NorthAtlantic	-0.0660	0.0300	-2.20	0.0282
factor(Basin)NorthAtlantic:ONIAug	-0.5037	0.2726	-1.85	0.0646

Table 5: The probability that a named storm becomes a major hurricane for 2017 season

	Season	Basin	ONIAug	Probability
1	2017	EasternPacific	-1.50	0.21
2	2017	EasternPacific	0.00	0.31
3	2017	EasternPacific	1.50	0.44
4	2017	NorthAtlantic	-1.50	0.18
5	2017	NorthAtlantic	0.00	0.16
6	2017	NorthAtlantic	1.50	0.13

For north Atlantic basin, the standard error is $\sqrt{0.005682844 + 2 \times (-0.005676686) + 0.01245255} = 0.08235301$. And for eastern Pacific basin, the standard error is $\sqrt{0.005682844} = 0.07538464$. Therefore, the corresponding transformed 95% Wald CI for the multiplicative effect of a 1-unit increase in August ONI on the mean number of named storms for north Atlantic basin is:

$$\exp(-0.1999182 \pm 1.96 \times 0.08235301) = (0.6967489, 0.9622256)$$

And corresponding transformed 95% Wald CI for the multiplicative effect of a 1-unit increase in August ONI on the mean number of named storms for eastern Pacific basin is:

$$\exp(0.1694707 \pm 1.96 \times 0.07538464) = (1.021954, 1.373311)$$

For logistic Regression:

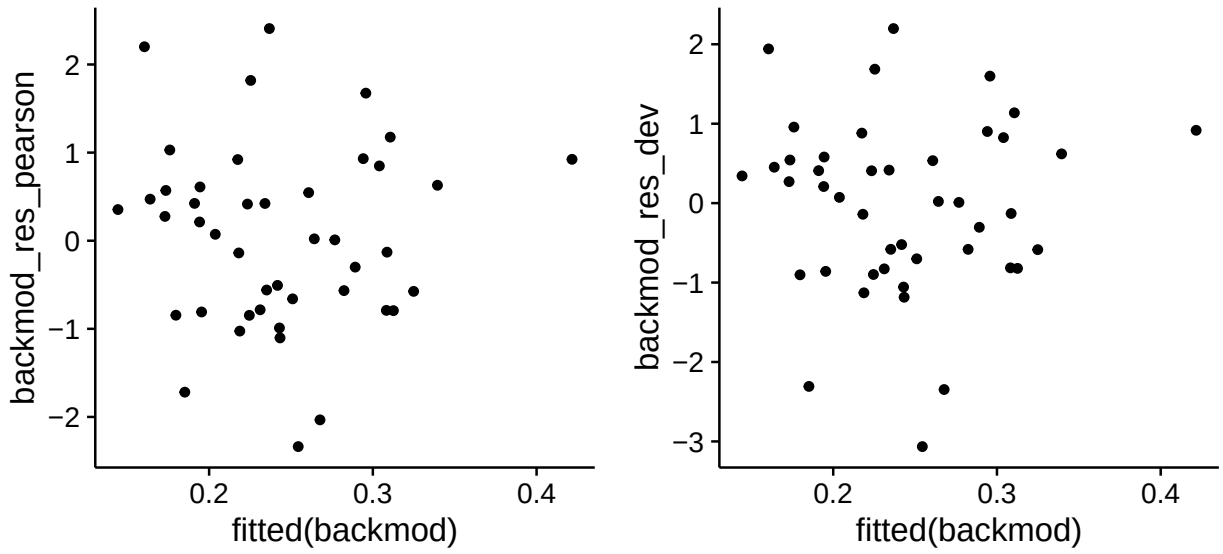
To assess the fit of the final model (after backward elimination), I make plots of the Pearson residuals and of deviance residuals. The plots don't show any pattern for both two types of residuals. Additionally, the chi squared p-value associated with the residual deviance 51.890 is 0.06594003, which suggests no evidence of lack of fit.

To assess the potential influential observations, I calculate the absolute values of standardized residuals of the model, and find the observation 7, 17, 33, 37 have a absoluate value of standardized residuals larger than 2. Thus based on this criteria, the influential observations are 7, 17, 33, 37.

Alternatively, there is no observations with absolute values of standardized residuals larger than 3. Thus we could conclude there is no influential observations for the criteria that standardized residuals in absolute values are larger than 3.

Table 6: Estimated asymptotic covariance matrix

	(Intercept)	Season	NorthAtlantic	ONIAug	NorthAtlantic:ONIAug
(Intercept)	159.751196	-0.079639	-0.001236	0.018772	0.033078
Season	-0.079639	0.000040	-0.000001	-0.000009	-0.000016
NorthAtlantic	-0.001236	-0.000001	0.006551	0.000199	0.000856
ONIAug	0.018772	-0.000009	0.000199	0.005683	-0.005677
NorthAtlantic:ONIAug	0.033078	-0.000016	0.000856	-0.005677	0.012453



Conclusion

To answer the first research question, the results loglinear model suggest that August ONI and interaction between ONI and basin statistically significant predict the total number of tropical storms. More specifically, the correlation between ONI values and total number of storms depends on different basin. Specially, higher ONI of storms in eastern Pacific basin would predict a higher amount of cyclones compared to north Atlantic basin.

For the second question, through backward elimination we obtain a model with variables of season, basin, ONI, interaction between season and basin and interaction between basin and ONI. Among these selected variables, basin, ONI, interaction between season and basin and interaction between basin and ONI give significant prediction about the probability. Moreover, the goodness of fit test suggests no evidence of lack of fit of this model. Furthermore, through different scenario of ONI values, we could infer that for eastern Pacific, a higher ONI leads to a higher probability that a named storm becomes a major hurricane, while in north Atlantic a higher ONI is associated with a lower probability given a same season in both cases.

Appendix

```r
rm(list = ls())
getwd()
setwd("/Users/Rocio/Library/Mobile Documents/com~apple~CloudDocs/STAT_426")
cyclone <- read.table("cyclone.txt", header = T)

names(cyclone)
dim(cyclone)

# Describe and summarize the variables.
# Include a table of marginal totals (summed over all seasons)
# of each category for each basin, and comment. Also, for each basin separately,
# produce a plot of total number of storms versus August ONI.



#install.packages("stargazer")
library(stargazer)
stargazer(cyclone, summary = T)

attach(cyclone)
dta_by_season <- aggregate(cyclone[, 3:6], by = list(Category = Season), FUN = sum)
dta_by_basin <- aggregate(cyclone[, 3:6], by = list(Category = Basin), FUN = sum)

library(xtable)
xtable(dta_by_basin)
xtable(dta_by_season)

library(ggplot2)
g <- ggplot(data = cyclone, aes(x = ONIAug, y = Total, colour = Basin))
g + geom_point() + scale_colour_grey()

## Fit a loglinear regression model for the total number of named storms,
## with a term for season (linear in year - not categorical),
## basin (indicator variable), August ONI, and interaction between basin and August ONI.
cycfit <- glm(Total ~ Season + factor(Basin) + ONIAug + ONIAug*factor(Basin),
              family = poisson, data = cyclone)
summary(cycfit)

stargazer(cycfit, title = "Results of loglinear regression")



## Goodness-of-Fit Test
deviance(cycfit)
df.residual(cycfit)
1 - pchisq(deviance(cycfit), df.residual(cycfit)) ## p-value, cannot reject the null hypothesis


## For log(mu), basin NorthAtlantic intercept will be
intercept_na <- coef(cycfit)[1] + coef(cycfit)[3]
## slope for ONIAug for basin NorthAtlantic
```

```r
slope_aug_na <- coef(cycfit)[4] + coef(cycfit)[5]


## For log(mu), basinEasternPacific intercept is
coef(cycfit)[1]
## slope for ONIAug for basin EasternPacific
coef(cycfit)[4]

## multiplicative effect of a one-unit increase in August ONI
## on the mean number of named storms, for basin NorthAtlantic
exp(coef(cycfit)[4] + coef(cycfit)[5])

## multiplicative effect of a one-unit increase in August ONI
## on the mean number of named storms, for basin EasternPacific
exp(coef(cycfit)[4])


## Consider logistic regression for the probability that a named tropical cyclone
## becomes a major hurricane
## 3 way interaction between season, basin and ONIAug
cycfull <- glm(cbind(MajorHurricane, Total - MajorHurricane) ~ Season * factor(Basin) * ONIAug,
               family = binomial, data = cyclone)

summary(cycfull)

## backward elimination

backmod <- step(cycfull)
summary(backmod)

## coef for basin EasternPacific
coef_ep <- c(coef(backmod)[1], coef(backmod)[2], coef(backmod)[4])

## coefs for basin NorthAtlantic
coef_na <- c(coef(backmod)[1] + coef(backmod)[3], coef(backmod)[2] + coef(backmod)[5],
  coef(backmod)[4] + coef(backmod)[6])

## create a table for 2017 different ONI AUG values
pred_data <- data.frame(matrix(0, 6, 3))
colnames(pred_data) <- c("Season", "Basin", "ONIAug")
pred_data$Season <- rep(2017, 6)
pred_data$Basin <- c(rep("EasternPacific", 3), rep("NorthAtlantic", 3))
pred_data$ONIAug <- rep(c(-1.5, 0, 1.5), 2)
pred_data$y <- predict(backmod, pred_data, type = "response")
pred_data


## double check
ep_fun<- function(oniaug, year){
  result <- coef_ep[1] + coef_ep[2]*year + coef_ep[3]*oniaug
  return(result)
}
```

```r
na_fun <- function(oniaug, year){
  result <- coef_na[1] + coef_na[2]*year + coef_na[3]*oniaug
  return(result)
}

exp(unname(ep_fun(c(-1.5, 0, 1.5), 2017)))/ (1 + exp(unname(ep_fun(c(-1.5, 0, 1.5), 2017))))
exp(unname(na_fun(c(-1.5, 0, 1.5), 2017)))/ (1 + exp(unname(na_fun(c(-1.5, 0, 1.5), 2017))))

## a) For the loglinear regression: For each basin,
## form a (transformed) Wald 95% confidence interval
## for the multiplicative effect of a one-unit increase in August ONI
## on the mean number of named storms.

# for basin EaternPacific:
beta_vcov = vcov(cycfit)
ci_ep = exp(c(cycfit$coefficients[4] - 1.96*sqrt(beta_vcov[4,4]),
              cycfit$coefficients[4] + 1.96*sqrt(beta_vcov[4,4])))

# for basin NorthAtlantic:
slope_na_se = sqrt(beta_vcov[4,4] + 2*beta_vcov[4,5] + beta_vcov[5,5])
ci_na = exp(c(slope_aug_na - 1.96 * slope_na_se,
              slope_aug_na + 1.96 * slope_na_se))

ci_ep
ci_na

## b) For the logistic regression:
## Assess the fit of your final model (after backward elimination)
## using appropriate residuals.
backmod_res_pearson <- residuals(backmod, "pearson")
backmod_res_dev <- residuals(backmod)

sum(residuals(backmod, "pearson")^2) ##
deviance(backmod) ## deviance
1- pchisq(deviance(backmod), df.residual(backmod))
## p-value is large indicating no evidence of lack of fit.
which(abs(rstandard(backmod, type  = "pearson")) > 2)
which(abs(rstandard(backmod, type  = "pearson")) > 3)

library(grid)
library(gridExtra)
p1 <- ggplot(data = cyclone, aes(x = fitted(backmod), y = backmod_res_pearson)) + geom_point()
p2 <- ggplot(data = cyclone, aes(x = fitted(backmod), y = backmod_res_dev)) + geom_point()
grid.arrange(p1, p2, ncol = 2)
```