CASE STUDY: CYCLONES

DEFINITION

Cyclones are defined as

``an atmospheric system in which the barometric pressure diminishes progressively to a minimum at the centre and toward which the winds blow spirally inward from all sides, resulting in a lifting of the air and eventually in clouds and precipitation...".

Hurricanes are cyclones, that originate in the tropics with windspeeds beyond 64 knots (= 74 mph, 113 km/h)
SOME BACKGROUND ON HURRICANES

(1) Average insurance claims per year $1 billion
(2) Extreme “violent” hurricanes can exceed $10 billion
(3) Example hurricane “Andrew” in 1992:
GENERAL GOAL OF THE STUDY

(4) Understand properties of cyclones based on the recorded variables
(5) Predict:
   a. Track of cyclones
   b. Probability of landfall

SOURCE OF THE DATA

• US National Hurricane Center
• Various web sites on tropical storms:
  o US National Hurricane Center: http://www.nhc.noaa.gov/
  o Tropical Storm Page: http://www.solarefa.hawaii.edu/Tropical/tropical.html
  o More data at: http://www.bbsr.edu/rpi/research/demaria/demaria4.html
  o ...
DATA DESCRIPTION

• Number of variables: 18
  o Date 3 (Year, Month, Day)
  o Name, Number 2
  o Location, X (longitude), Y (latitude) 2
  o Categorical 2
    ▪ Stormtype
    ▪ Landfall
  o Continuous 9
    ▪ Speeds
    ▪ Angles
    ▪ Distance

• Number of Cases: 1819
  But only 334 different storms
  (Number of observations per storm: min 1 (42=12.5%), median 5, max 24)
## SUGGESTED APPROACHES

<table>
<thead>
<tr>
<th>Approach</th>
<th>Reason</th>
<th>Type of Question addressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculate summaries of all variables</td>
<td>Extract scale, location and range information</td>
<td>What is the average windspeed of a hurricane in this dataset?</td>
</tr>
<tr>
<td>Draw distributions of variables</td>
<td>Understand asymmetry and outlier of the variables</td>
<td>Which variables are useful for a statistical model?</td>
</tr>
<tr>
<td>Plot interactions of variables</td>
<td>Understand interaction structure of the data</td>
<td>Which variables contribute information to a model?</td>
</tr>
<tr>
<td>Draw maps with hurricane locations</td>
<td>Understand geographical distribution</td>
<td>Where do hurricanes occur?</td>
</tr>
<tr>
<td>Draw tracks of hurricanes</td>
<td>Look for similar track types</td>
<td>What does a typical track of a hurricane look like?</td>
</tr>
<tr>
<td>Are there different types of tracks shapes?</td>
<td>Plot geographical distribution of variables – or – link information of other variables into the scatterplot</td>
<td>What is the interaction of location with all the variables</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>---------------------------------------------------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>– Check accuracy of the data –</td>
<td>Try to predict a landfall from single measurements</td>
<td>What is the probability that this hurricane will hit land?</td>
</tr>
</tbody>
</table>

| Set up statistical model for landfall | | | |
ACTUAL APPROACHES

(1) Summaries

a. Year:

> summary(Year)

Min.   1st Qu.  Median   Mean   3rd Qu.  Max.

> barplot(table(Year))
b. Month

```r
> summary(Month)
  Min. 1st Qu.  Median    Mean   3rd Qu.   Max.
  1.00   8.00   9.00   8.769    9.00   12.00

> barplot(table(Month))
```
c. Name

> summary(Name)

<table>
<thead>
<tr>
<th>Name</th>
<th>NOT</th>
<th>CHARLIE</th>
<th>SUBTROP</th>
<th>EDITH</th>
<th>BETSY</th>
<th>ANNA</th>
<th>ABLE</th>
<th>DOG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>289</td>
<td>36</td>
<td>35</td>
<td>33</td>
<td>32</td>
<td>30</td>
<td>29</td>
<td>29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>CAROL</th>
<th>GLADYS</th>
<th>GRETA</th>
<th>INGA</th>
<th>BECKY</th>
<th>ELLA</th>
<th>FLORA</th>
<th>FRANCES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>28</td>
<td>28</td>
<td>28</td>
<td>28</td>
<td>27</td>
<td>27</td>
<td>27</td>
<td>27</td>
</tr>
</tbody>
</table>

...
d. Location

```r
> plot(X, Y, col=3-Landfall)
```
Some example paths for “Donna” and “Anna”:
> plot(X, Y, col=3-Landfall)
> lines(X[Name=="DONNA"], Y[Name=="DONNA"], col=2, lwd=3)
> lines(X[Stormnumber==275], Y[Stormnumber==275], col=3, lwd=3)
e. Stormtype

```r
> table(Stormtype)
Stormtype
   1   3   5
1674 76 69
```
f. Landfall

\[ > \text{table(Landfall)} \]

<table>
<thead>
<tr>
<th>Landfall</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1431</td>
<td>388</td>
</tr>
</tbody>
</table>

\[ > \frac{388}{1431} \]

0.2711391

\[ > \text{barplot(table(Landfall))} \]

\[ > \text{table(LandfallN)} \]

<table>
<thead>
<tr>
<th>LandfallN</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>234</td>
<td>100</td>
</tr>
</tbody>
</table>

\[ > \frac{234}{100} \]

0.4273504

\[ > \text{barplot(table(LandfallN))} \]
g. Speeds

> hist(Windspeed, col=3)
> hist(Windspeed[Landfall==T], add=T, col=2)
h. Angles
i. Distance
RELATIONS BETWEEN VARIABLES

a. Checked all variables against Landfall, but …

… highlighted subgroup is hard to compare with total in a histogram.

1. alternative: Boxplots
2. alternative: Spinograms

Example: Windspeed

Histogram  Spinogram
Trackangle

Histogram

Spinogram
b. Dependencies of the derived variables

(1) Inward speed and parallel speed are derived from track speed

Trackspeed

Inwarspeed vs. Parallelspeed
(2) Coast angle and distance angle are derived from track angle
(3) Distance is related to longitude and latitude

Distance

Latitude vs. Longitude
c. Looking at 2-way interactions between continuous variables
d. Interactions in more than 2 dimensions: 2-d Tour
e. A SIMPLE MODEL FOR LANDFALL

```r
> m1 <- lm(Landfall ~ 1)
> summary(m1)
Call:
  lm(formula = Landfall ~ 1)

... 
Coefficients:

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| 0.213304 | 0.009607   | 22.2    | <2e-16 *** |

... 
> add1(m1, cyclones[c(1:3,9:16)])

Single term additions

Model:
Landfall ~ 1

<table>
<thead>
<tr>
<th>Term</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td></td>
<td>305.2</td>
<td>-3244.8</td>
</tr>
<tr>
<td>X</td>
<td>1</td>
<td>19.0</td>
<td>286.2</td>
<td>-3359.7</td>
</tr>
<tr>
<td>Y</td>
<td>1</td>
<td>17.6</td>
<td>287.6</td>
<td>-3351.0</td>
</tr>
<tr>
<td>Windspeed</td>
<td>1</td>
<td>1.2</td>
<td>304.1</td>
<td>-3249.8</td>
</tr>
<tr>
<td>Trackspeed</td>
<td>1</td>
<td>6.0</td>
<td>299.3</td>
<td>-3278.8</td>
</tr>
<tr>
<td>Trackangle</td>
<td>1</td>
<td>2.4</td>
<td>302.9</td>
<td>-3256.9</td>
</tr>
<tr>
<td>Distance</td>
<td>1</td>
<td>3.2</td>
<td>302.0</td>
<td>-3262.1</td>
</tr>
<tr>
<td>Coastangle</td>
<td>1</td>
<td>19.9</td>
<td>285.3</td>
<td>-3365.6</td>
</tr>
<tr>
<td>Distanceangle</td>
<td>1</td>
<td>13.8</td>
<td>291.4</td>
<td>-3327.1</td>
</tr>
<tr>
<td>Inwardspeed</td>
<td>1</td>
<td>16.7</td>
<td>288.5</td>
<td>-3345.3</td>
</tr>
<tr>
<td>Parallelspeed</td>
<td>1</td>
<td>16.8</td>
<td>288.5</td>
<td>-3345.6</td>
</tr>
</tbody>
</table>
```
This stepwise regression yields the model:

\[ \text{Landfall} \sim \text{Coastangle} + \text{Inwardspeed} + X + \text{Windspeed} + Y \]

with a \( R^2 \) of 13.5%

Prediction: \( \text{hist(predict(m1, cyclones))} \)
Confusion matrix:

```r
> table(round(0.164+predict(m1, cyclones)))
   0  1
 1430 389
> table(round(0.164+predict(m1, cyclones)), Landfall)
Landfall
      0  1
 0  1215 215
 1  216 173
> mosaicplot(table(round(0.164+predict(m1, cyclones)), Landfall))
```
INVESTIGATION OF THE RESIDUALS

• Geographic mapping of:

  false positives

  false negatives
Approach: Partition the data into

- Gulf region
- Atlantic region
- North-east Atlantic region
- (removing outlier “Ginger”)
R-Code:

```r
> group<-rep(1, length(X))
> group[(Y>=-0.5825*X-22.5641) & X > -80]<-2
> group[(Y>=-0.4806*X+7.8837) | (Distanceangle>90 & group == 2 & Landfall == 0)]<-3
> group[Stormnumber==310]<-4
> barplot(table(group), col=2:5)
> plot(X,Y, col=group+1)
```
f. REESTIMATING TWO SEPARATE LOGISTIC MODELS

**Logistic Regression:**

ordinary linear model:

\[ Y = aX + b + e \]

Problem with Y:
Y is dichotomous, but the linear fit will give estimates on \(-\cdot\) to \(+\cdot\)

Solution:
Introduce sigmoid link-function to map data from \(-\cdot, +\cdot\) [to [0, 1].

\[ \ln(p/(1-p)) = aX + b + \epsilon \quad \text{(logit-link)} \]

Remarks:
Other continuous link function from \(-\cdot, +\cdot\) [to [0, 1] are used as well.

Parameter estimate no longer works with simple solution of linear equation, but needs iterative optimization methods to find a solution.

\(R^2\) can not be extracted as in the Ordinary Least Square case
i) Estimate for the Gulf region:

```r
> g1<-glm(Landfall ~ 1, subset=(group==1), family="binomial")
> add1(g1, cyclones[c(1:3,9:16)])
... 

Landfall ~ Trackangle + Y + Inwardspeed + Coastangle

> table(round(0.07+predict(g1, cyclones, type="response")[group==1]))

     0  1
 253 179

> table(round(0.07+predict(g1, cyclones, type="response")[group==1]), Landfall[group==1])

     0  1
 0 191  62
 1  62 117

R^2: 20.8%```
ii) Estimate for the Atlantic region:

```r
> a1 <- glm(Landfall ~ 1, subset=(group==2), family="binomial")
> add1(a1, cyclones[c(1:3, 9:16)])
```

Landfall ~ Coastangle + Parallelspeed + Windspeed + Y + X

```r
> table(round(0.1809+predict(a1, cyclones, type="response")[group==2]))
```

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>656</td>
<td>185</td>
</tr>
</tbody>
</table>

```r
> table(round(0.1809+predict(a1, cyclones, type="response")[group==2]), Landfall[group==2])
```

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>559</td>
<td>106</td>
</tr>
<tr>
<td>1</td>
<td>106</td>
<td>79</td>
</tr>
</tbody>
</table>

$R^2$: 11.9%
iii) Combined estimate for both regions:

<table>
<thead>
<tr>
<th>Landfall</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>741</td>
<td>168</td>
</tr>
<tr>
<td>1</td>
<td>168</td>
<td>196</td>
</tr>
</tbody>
</table>

R²: 17.9%
g. IMPROVE MODEL BY USING THE TRACK STRUCTURE

Idea:
Use the predictions along a track to “vote” for landfall

Classify a complete storm as landfall, iff the “majority” of its points are classified as landfall. Use linear weights to emphasize “latest” measures.

i) only Gulf and Atlantic

<table>
<thead>
<tr>
<th>Landfall</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>817</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>92</td>
<td>224</td>
</tr>
</tbody>
</table>

$R^2$: 38.1%

ii) all data

<table>
<thead>
<tr>
<th>Landfall</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1334</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>97</td>
<td>248</td>
</tr>
</tbody>
</table>

$R^2$: 45.4%
tracks of “Donna” and “Anna”
prediction tracks of “Donna” and “Anna”
h. INVESTIGATION OF THE NEW RESIDUALS

i) False negatives 140/1474 = 9%

No apparent structure visible.
ii) False positives $97/345 = 28$

The following storms apparently hit Cuba or Mexico:
- Mexico: 138, 155, 177, 246, 298
- Cuba: 48, 75, 83, 89, 192, 257, 282, 340, 370, 379

These storm account for 47 measurements.

--> corrected false positives: $50/345 = 14\%$