EDF 5401 Midterm, Part 1: Hurricanes.

2013-10-15

These data come from the [Data and Story Library](https://dasl.datadescription.com/datafile/hurricanes-2015/?_sf_s=hurricane&_sfm_cases=4+59943)

Most weather models note at relationship between the barimetric pressure and the peak wind speeds. A secondary question is, as the average temperature rises, is that relationship changing.

library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.3 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.0  
✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
✔ purrr 1.0.2   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(DescTools)

# Part 1 Hurricanes

## Exploratory Analysis

### Load the data

Load the data. Force category to be an ordered category.

hurric <- read\_delim("hurricanes-2015.txt")

Rows: 226 Columns: 5  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: "\t"  
chr (1): Name  
dbl (4): Year, Max.Wind.Speed(kts), Central.Pressure(mb), Category  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

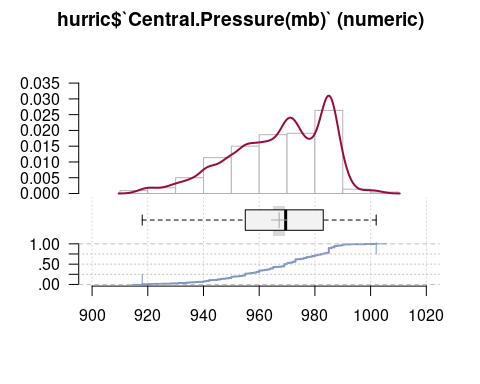
hurric$Category <- ordered(hurric$Category)  
summary(hurric)

Name Year Max.Wind.Speed(kts) Central.Pressure(mb)  
 Length:226 Min. :1851 Min. : 65.00 Min. : 918.0   
 Class :character 1st Qu.:1882 1st Qu.: 70.00 1st Qu.: 955.0   
 Mode :character Median :1910 Median : 85.00 Median : 969.5   
 Mean :1939 Mean : 88.78 Mean : 967.2   
 3rd Qu.:2006 3rd Qu.:100.00 3rd Qu.: 983.0   
 Max. :2015 Max. :150.00 Max. :1002.0   
 NA's :6   
 Category   
 1 :92   
 2 :55   
 3 :49   
 4 :17   
 5 : 3   
 NA's:10

### One-dimensional analyses

Desc(hurric$`Central.Pressure(mb)`)

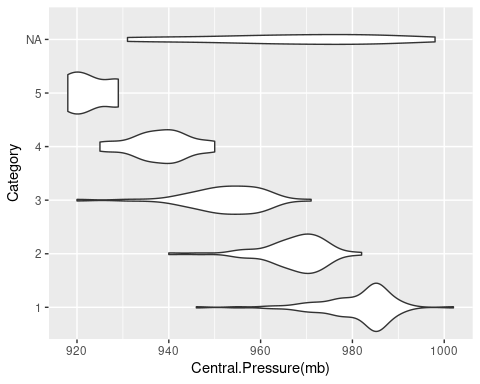
------------------------------------------------------------------------------   
hurric$`Central.Pressure(mb)` (numeric)  
  
 length n NAs unique 0s mean meanCI'  
 226 220 6 66 0 967.16 964.90  
 97.3% 2.7% 0.0% 969.43  
   
 .05 .10 .25 median .75 .90 .95  
 935.95 942.00 955.00 969.50 983.00 986.00 988.00  
   
 range sd vcoef mad IQR skew kurt  
 84.00 17.04 0.02 20.76 28.00 -0.59 -0.30  
   
lowest : 918.0, 920.0, 922.0, 925.0, 928.0  
highest: 990.0 (4), 991.0, 993.0, 998.0, 1'002.0  
  
heap(?): remarkable frequency (11.8%) for the mode(s) (= 985)  
  
' 95%-CI (classic)



Look at differences in pressure by category.

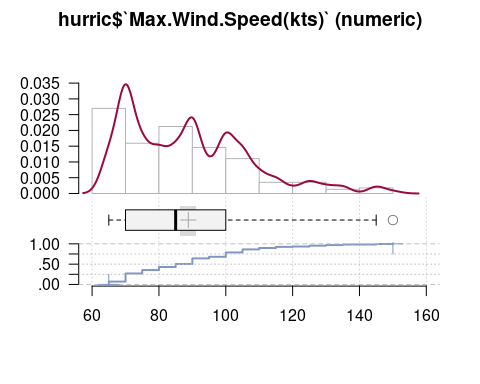
ggplot(hurric, aes(x=`Central.Pressure(mb)`,y=Category)) + geom\_violin()

Warning: Removed 6 rows containing non-finite values (`stat\_ydensity()`).



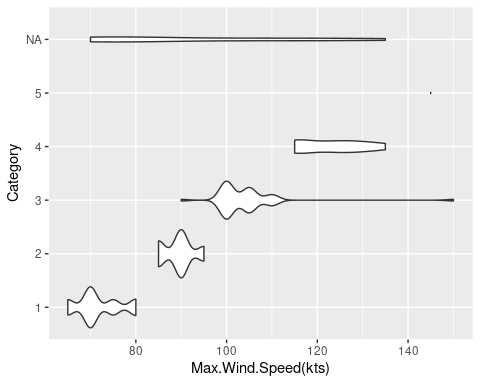
Desc(hurric$`Max.Wind.Speed(kts)`)

------------------------------------------------------------------------------   
hurric$`Max.Wind.Speed(kts)` (numeric)  
  
 length n NAs unique 0s mean meanCI'  
 226 226 0 17 0 88.78 86.33  
 100.0% 0.0% 0.0% 91.24  
   
 .05 .10 .25 median .75 .90 .95  
 65.00 70.00 70.00 85.00 100.00 112.50 125.00  
   
 range sd vcoef mad IQR skew kurt  
 85.00 18.73 0.21 22.24 30.00 0.86 0.39  
   
lowest : 65.0 (16), 70.0 (45), 75.0 (19), 80.0 (17), 85.0 (17)  
highest: 125.0 (5), 130.0 (3), 135.0 (3), 145.0 (3), 150.0  
  
heap(?): remarkable frequency (19.9%) for the mode(s) (= 70)  
  
' 95%-CI (classic)



Look at differences in maximum speed by category. (Note category is largely defined by wind speed.)

ggplot(hurric, aes(x=`Max.Wind.Speed(kts)`,y=Category)) + geom\_violin()



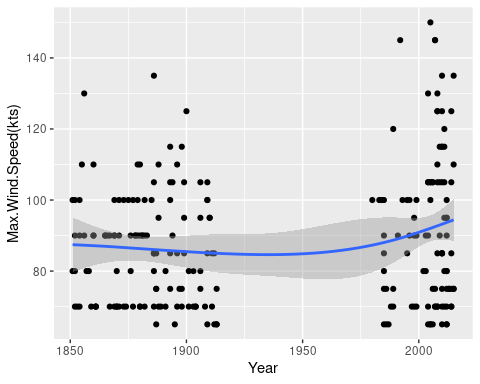
### Relationships with time

round(cor(hurric[,2:4],use="complete.obs"),3)

Year Max.Wind.Speed(kts) Central.Pressure(mb)  
Year 1.000 0.131 -0.147  
Max.Wind.Speed(kts) 0.131 1.000 -0.898  
Central.Pressure(mb) -0.147 -0.898 1.000

ggplot(hurric,aes(x=Year,y=`Max.Wind.Speed(kts)`)) +   
 geom\_point() + geom\_smooth()

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'



Hmm. Note big gap in data between 1925 and 1975. Maybe before/after climate change? Note 1950 appears to be a cut point.

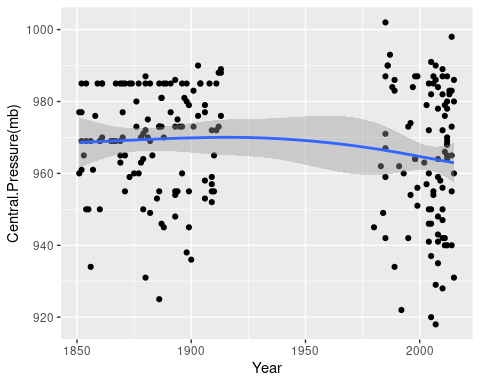
hurric <- mutate(hurric,recent=Year>1950)

ggplot(hurric,aes(x=Year,y=`Central.Pressure(mb)`)) +   
 geom\_point() + geom\_smooth()

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning: Removed 6 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 6 rows containing missing values (`geom\_point()`).



## Scatterplots

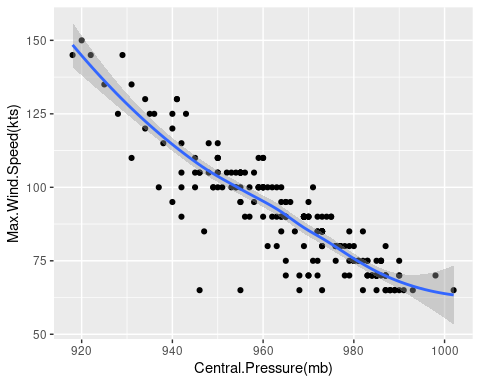
### XY

ggplot(hurric,aes(x=`Central.Pressure(mb)`,y=`Max.Wind.Speed(kts)`)) +  
 geom\_point() + geom\_smooth()

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning: Removed 6 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 6 rows containing missing values (`geom\_point()`).



### Jittered

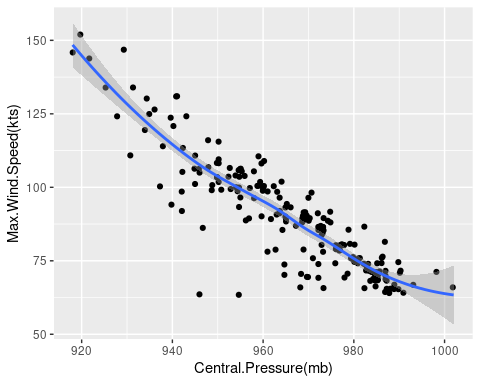
Hmm. Points lying on top of each other, try some jittering.

ggplot(hurric,aes(x=`Central.Pressure(mb)`,y=`Max.Wind.Speed(kts)`)) +  
 geom\_point(position="jitter") + geom\_smooth()

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning: Removed 6 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 6 rows containing missing values (`geom\_point()`).



### XY by Recent

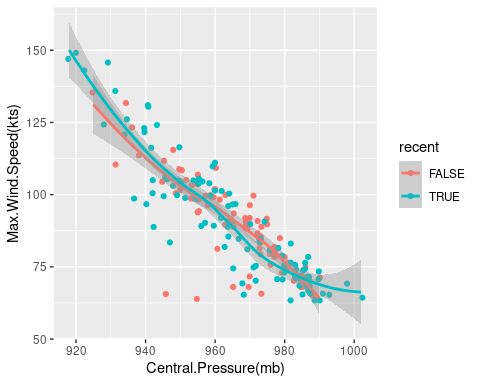
Color by recent to see if the current and recent groups are similar or not.

ggplot(hurric,aes(x=`Central.Pressure(mb)`,y=`Max.Wind.Speed(kts)`,color=recent)) +  
 geom\_point(position="jitter") + geom\_smooth()

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning: Removed 6 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 6 rows containing missing values (`geom\_point()`).



### Outliers

There seem to be a couple of ouliers. Lets try to find them.

hout <- which(hurric$`Central.Pressure(mb)`<960 &   
 hurric$`Max.Wind.Speed(kts)` < 75)  
hurric[hout,]

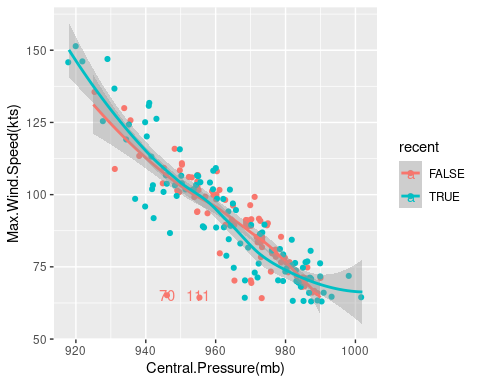
# A tibble: 2 × 6  
 Name Year `Max.Wind.Speed(kts)` `Central.Pressure(mb)` Category recent  
 <chr> <dbl> <dbl> <dbl> <ord> <lgl>   
1 ----- 1887 65 946 1 FALSE   
2 ----- 1909 65 955 1 FALSE

ggplot(hurric,aes(x=`Central.Pressure(mb)`,y=`Max.Wind.Speed(kts)`,  
 color=recent)) +  
 geom\_point(position="jitter") + geom\_smooth() +  
 geom\_text(data=hurric[hout,],aes(label=hout))

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning: Removed 6 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 6 rows containing missing values (`geom\_point()`).

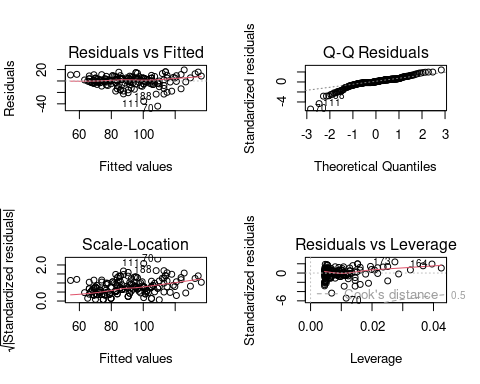


## Build the Regression Model

lm\_hurric <- lm(`Max.Wind.Speed(kts)` ~ `Central.Pressure(mb)`, data=hurric)  
summary(lm\_hurric)

Call:  
lm(formula = `Max.Wind.Speed(kts)` ~ `Central.Pressure(mb)`,   
 data = hurric)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-44.063 -2.145 0.182 4.459 19.365   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1031.24439 31.37284 32.87 <2e-16 \*\*\*  
`Central.Pressure(mb)` -0.97482 0.03243 -30.06 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 8.18 on 218 degrees of freedom  
 (6 observations deleted due to missingness)  
Multiple R-squared: 0.8056, Adjusted R-squared: 0.8047   
F-statistic: 903.4 on 1 and 218 DF, p-value: < 2.2e-16

oldpar <- par(mfrow=c(2,2))  
plot(lm\_hurric)



par(oldpar)

### Try some Y transforms

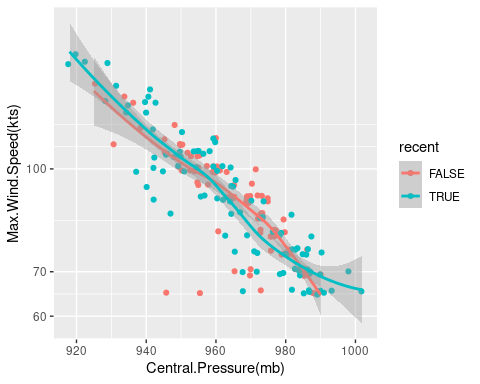
Log

ggplot(hurric,aes(x=`Central.Pressure(mb)`,y=`Max.Wind.Speed(kts)`,color=recent)) +  
 geom\_point(position="jitter") + geom\_smooth() + scale\_y\_log10()

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning: Removed 6 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 6 rows containing missing values (`geom\_point()`).



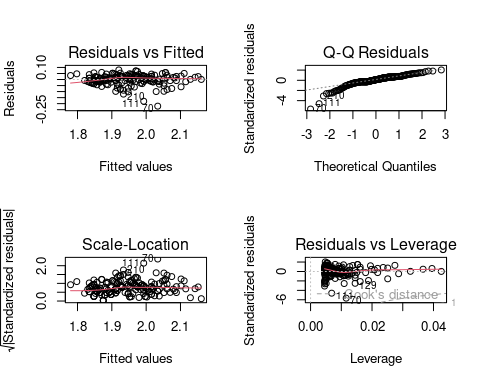
Log model

llm\_hurric <- lm(log(`Max.Wind.Speed(kts)`,10) ~ `Central.Pressure(mb)`, data=hurric)  
summary(llm\_hurric)

Call:  
lm(formula = log(`Max.Wind.Speed(kts)`, 10) ~ `Central.Pressure(mb)`,   
 data = hurric)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.220809 -0.011784 0.004988 0.024811 0.079637   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 6.3232468 0.1506825 41.96 <2e-16 \*\*\*  
`Central.Pressure(mb)` -0.0045344 0.0001558 -29.11 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.03929 on 218 degrees of freedom  
 (6 observations deleted due to missingness)  
Multiple R-squared: 0.7954, Adjusted R-squared: 0.7944   
F-statistic: 847.3 on 1 and 218 DF, p-value: < 2.2e-16

Log Diagnostics

oldpar <- par(mfrow=c(2,2))  
plot(llm\_hurric)



par(oldpar)

Flipping a coin, I’m doing the rest of the analyses on the linear scale.

### Outliers

Calculate dfbetas for identified outliers

dfbetas(lm\_hurric)[hout,]

(Intercept) `Central.Pressure(mb)`  
70 -0.4973275 0.4904759  
111 -0.2239674 0.2186286

Run the regression without the outliers.

lm\_hurric\_no <- lm(`Max.Wind.Speed(kts)` ~ `Central.Pressure(mb)`,   
 data=hurric, subset=-hout)  
summary(lm\_hurric\_no)

Call:  
lm(formula = `Max.Wind.Speed(kts)` ~ `Central.Pressure(mb)`,   
 data = hurric, subset = -hout)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-23.8948 -2.6321 0.4889 4.0445 18.1659   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1052.70051 27.92752 37.69 <2e-16 \*\*\*  
`Central.Pressure(mb)` -0.99663 0.02887 -34.52 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 7.246 on 216 degrees of freedom  
 (6 observations deleted due to missingness)  
Multiple R-squared: 0.8466, Adjusted R-squared: 0.8459   
F-statistic: 1192 on 1 and 216 DF, p-value: < 2.2e-16

## Run separately for old and recent data.

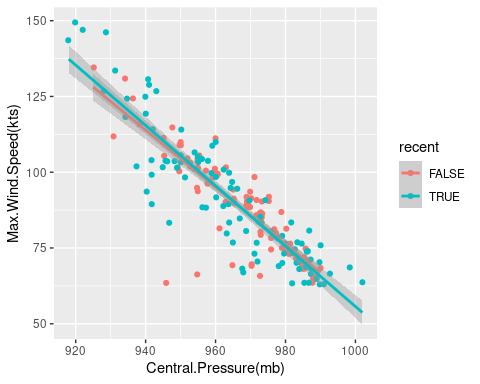
Redo the plot with method="lm" to visualize different lines.

ggplot(hurric,aes(x=`Central.Pressure(mb)`,y=`Max.Wind.Speed(kts)`,color=recent)) +  
 geom\_point(position="jitter") + geom\_smooth(method="lm")

`geom\_smooth()` using formula = 'y ~ x'

Warning: Removed 6 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 6 rows containing missing values (`geom\_point()`).



### 19th C, Early 20th

lm\_hurric\_19 <- lm(`Max.Wind.Speed(kts)` ~ `Central.Pressure(mb)`,   
 data=hurric, subset=!recent)  
summary(lm\_hurric\_19)

Call:  
lm(formula = `Max.Wind.Speed(kts)` ~ `Central.Pressure(mb)`,   
 data = hurric, subset = !recent)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-43.276 -1.539 0.273 3.389 15.273   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 999.38405 46.70347 21.40 <2e-16 \*\*\*  
`Central.Pressure(mb)` -0.94197 0.04817 -19.56 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 7.497 on 120 degrees of freedom  
Multiple R-squared: 0.7612, Adjusted R-squared: 0.7592   
F-statistic: 382.4 on 1 and 120 DF, p-value: < 2.2e-16

## Save slope and SE for later processing.  
hurric\_slope\_19 <- summary(lm\_hurric\_19)$coefficients[2,1:2]

### Late 20th, Early 21st

lm\_hurric\_20 <- lm(`Max.Wind.Speed(kts)` ~ `Central.Pressure(mb)`,   
 data=hurric, subset=recent)  
summary(lm\_hurric\_20)

Call:  
lm(formula = `Max.Wind.Speed(kts)` ~ `Central.Pressure(mb)`,   
 data = hurric, subset = recent)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-23.4643 -4.4765 0.8828 5.1694 18.6154   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1051.2682 44.6541 23.54 <2e-16 \*\*\*  
`Central.Pressure(mb)` -0.9956 0.0463 -21.50 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 9.014 on 96 degrees of freedom  
 (6 observations deleted due to missingness)  
Multiple R-squared: 0.8281, Adjusted R-squared: 0.8263   
F-statistic: 462.4 on 1 and 96 DF, p-value: < 2.2e-16

## Save slope and SE for later processing.  
hurric\_slope\_20 <- summary(lm\_hurric\_20)$coefficients[2,1:2]

Compare slopes in a table:

rbind(early=hurric\_slope\_19,  
 late=hurric\_slope\_20)

Estimate Std. Error  
early -0.9419745 0.04816806  
late -0.9955690 0.04629930

Standard error for the difference is

sqrt(hurric\_slope\_19[2]^2+hurric\_slope\_20[2]^2)

Std. Error   
0.06681158