STAT021 Statistical Methods II

Lecture 12 SLR Outliers and Influential Points

Lu Chen
Swarthmore College
10/11/2018

Simple Linear Regression

CHOOSE

- Exploratory data analysis; Model: $Y = \beta_0 + \beta_1 X + \epsilon$ where $\epsilon \overset{iid}{\sim} N(0, \sigma)$

FIT

- Maximum likelihood estimation (MLE)

ASSESS model

- Inference for the intercept and slope; ANOVA and R^2

ASSESS error

- Check conditions and transformations; Outliers and influential points

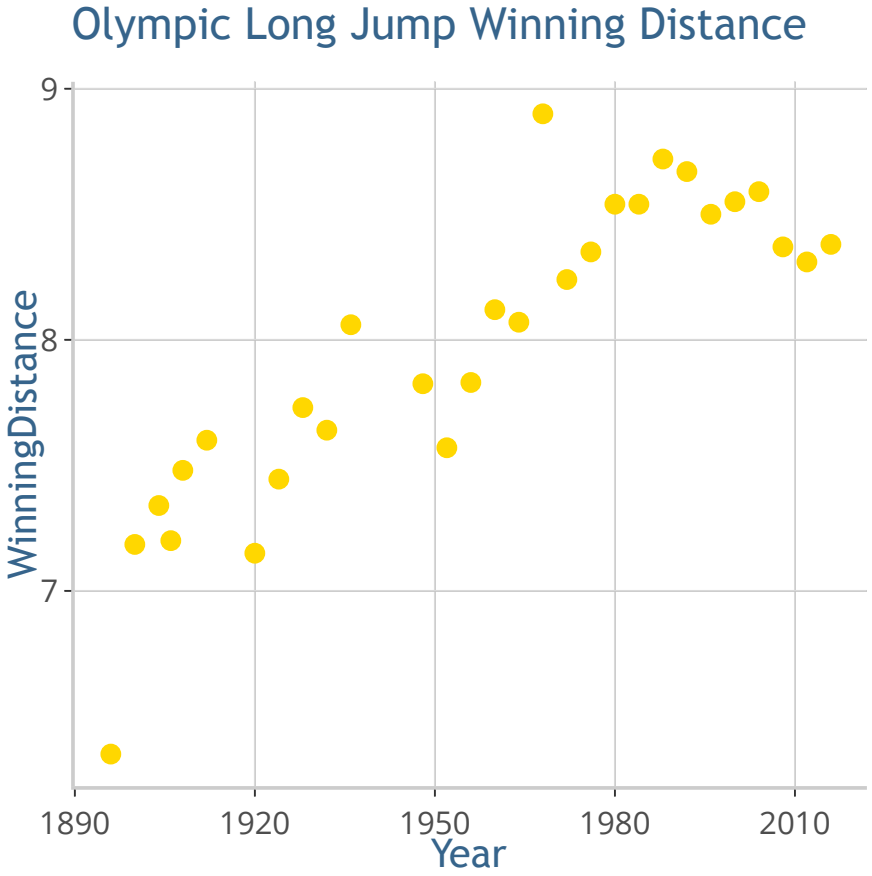
USE

- Predictions

Outline

- Example 1: Men's Olympic long jump winning distance
- Example 2: Presidential election in 2000
- Three diagnostic statistics
  - **Leverage**
  - **Standardized and studentized residuals**
  - **Cook's Distance**
- Three statistics in one diagnostic plot
- Applying to previous class examples
- Some notes
- Midterm examination

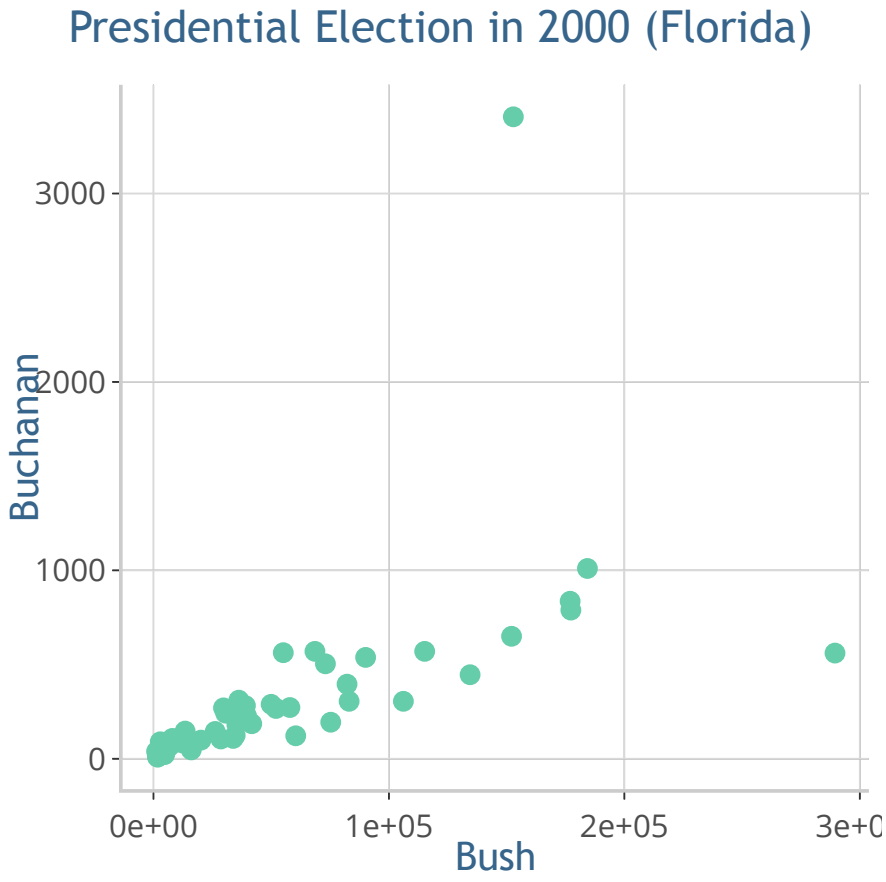
Example 1: Men's Olympic long jump



The winning men's Olympic long jump distance *WinningDistance* vs. *Year* for the $n = 29$ Olympics held during 1896-2016.

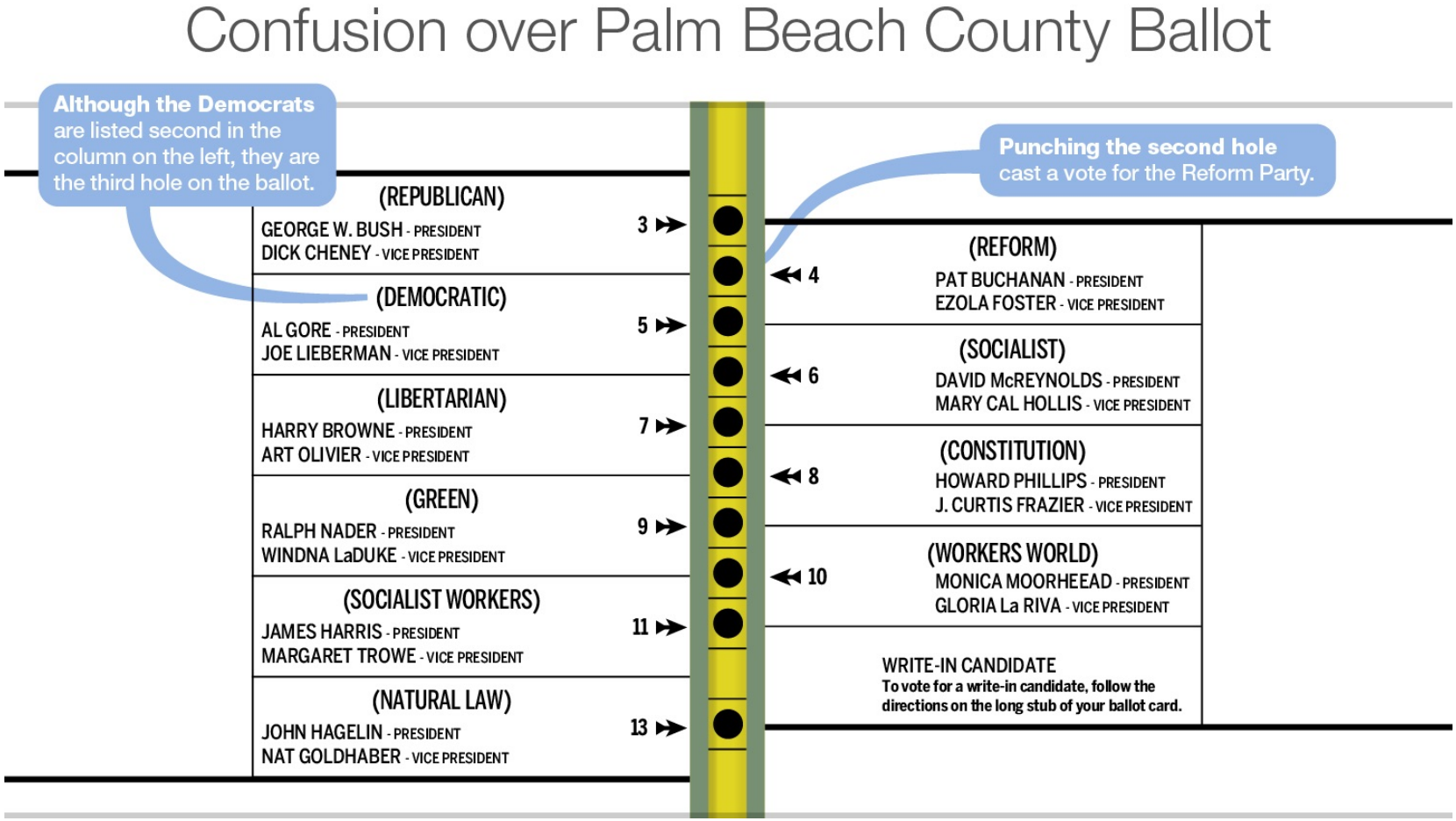
- Anything interesting?
- During the 1968 Olympics, Bob Beamon (USA) shocked the track and field world by jumping 8.9 meters (29'2.5"), breaking the world record by 55 cm (nearly 2 ft).
- This is still the current Olympic record.
- It has been estimated that the tail wind and altitude in Mexico may have improved Beamom's long jump distance by 31 cm (12.2 in).

Example 2: Presidential election in 2000



- Nationally, George W. Bush (Republican) received 47.9% of the popular vote, Al Gore (Democratic) received 48.4%, with the electoral votes from Florida determining the outcome.
- In Florida, Bush won by just 537 votes over Gore (48.847% to 48.838%) out of almost 6 million votes cast.
- This is a scatterplot of votes for Reform Party candidate Pat Buchanan vs. Bush in the 67 counties of Florida.
- Palm Beach County, one of the 67 counties in FL, used a unique "butterfly ballot", which resulted in an unusually high number of votes for Pat Buchanan.

Butterfly ballot



Identify outliers and influential points

Three diagnosis statistics for identifying outliers and influential points

Leverage

- Identifies influential points, data points that have a great impact on the regression line

Standardized and studentized residuals

- Identify outliers, data points with unusually large residuals

Cook's Distance

- A combination of leverage and standardized/studentized residuals
- Identifies unusual points - outliers, influential points, or both

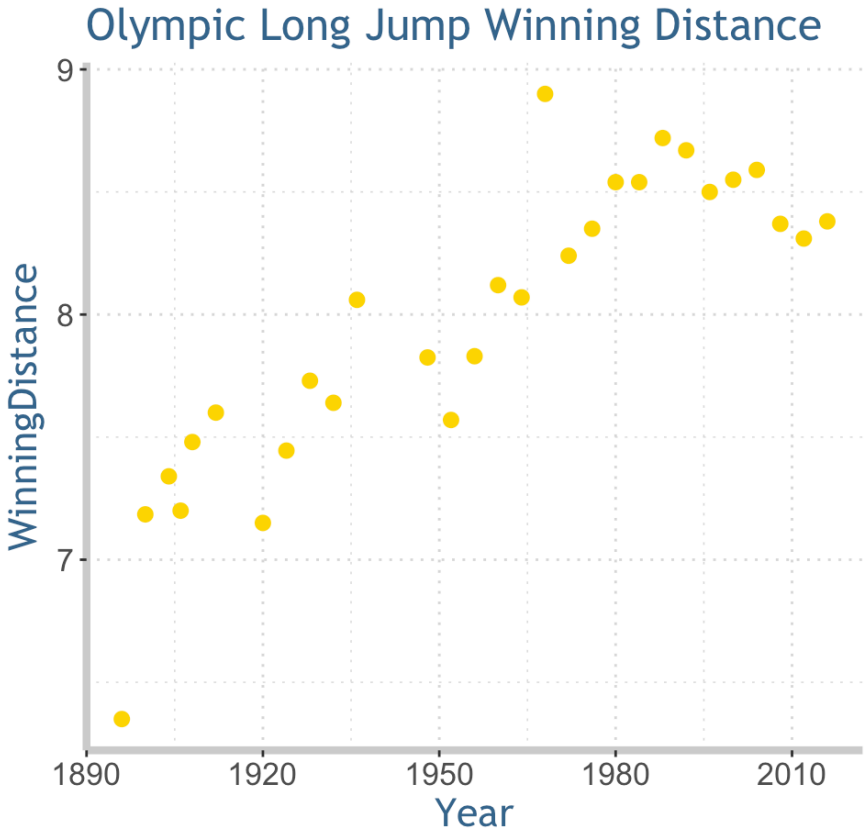
Leverage

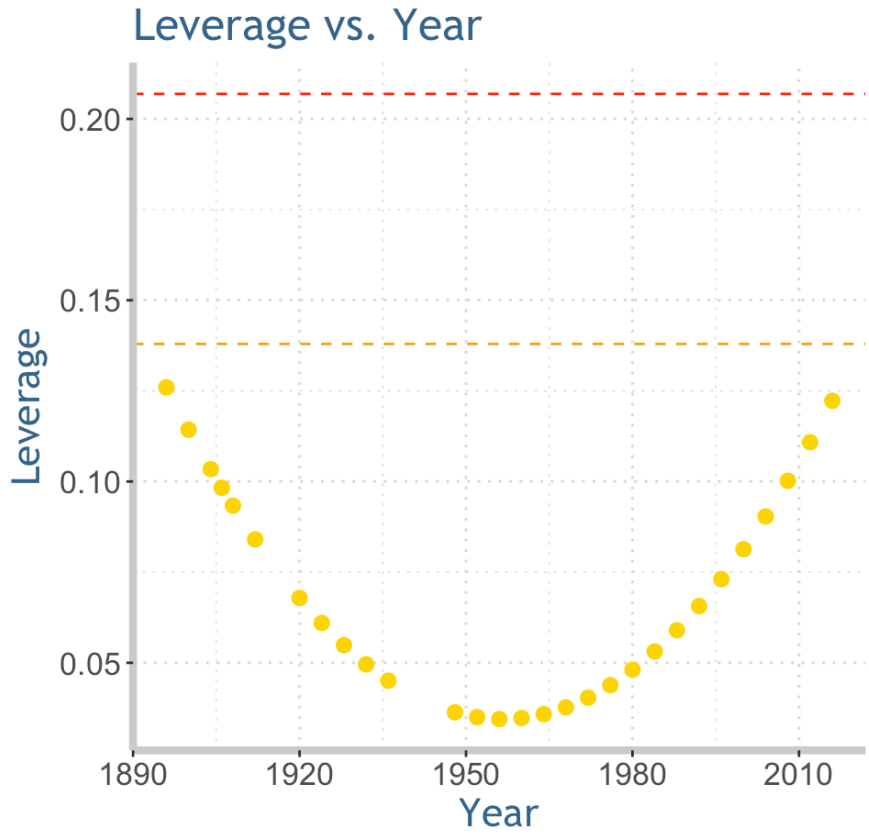
For a simple linear regression on n data points, the **leverage** of any point (x_i, y_i) is

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

- Generally, points farther from the mean value of the predictor $(\bar{x})$ have greater potential to influence the slope of a fitted regression line.
- Points with higher leverage have a greater capacity to pull the regression line in their direction.
- For SLR
  - Points with $h_i > 4/n$ are considered to have somewhat high leverages.
  - Points with $h_i > 6/n$ are considered to have especially high leverages.
- Leverage does NOT depend on the response variable.
- Points with high leverage do NOT necessarily have large influence on the regression line.

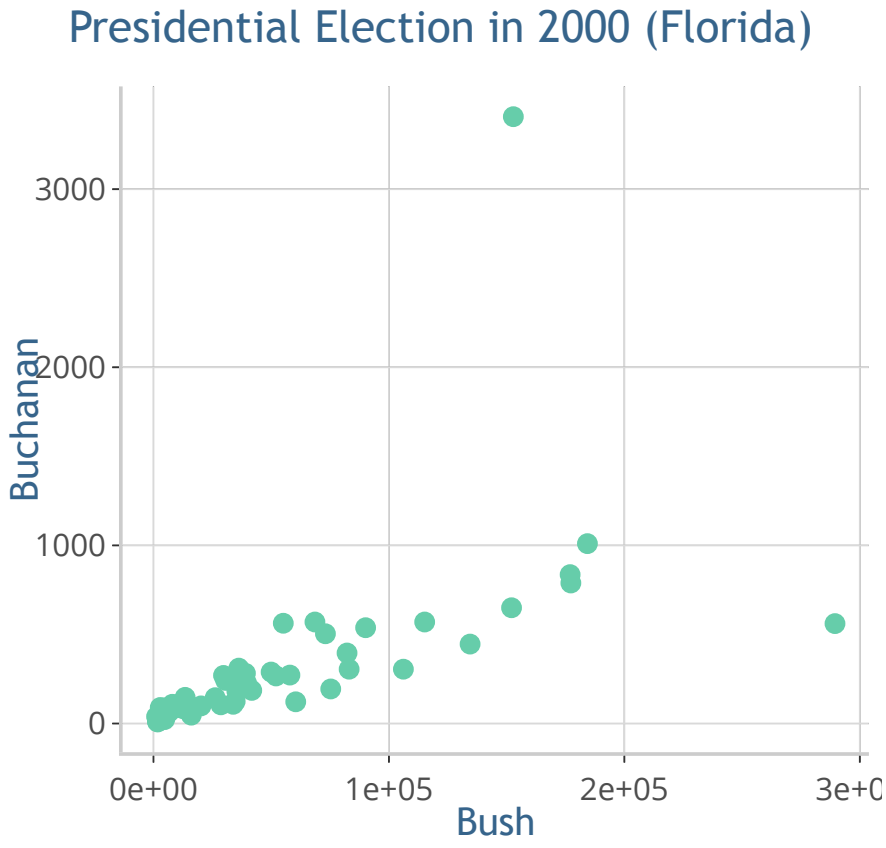
Leverage - Example 1

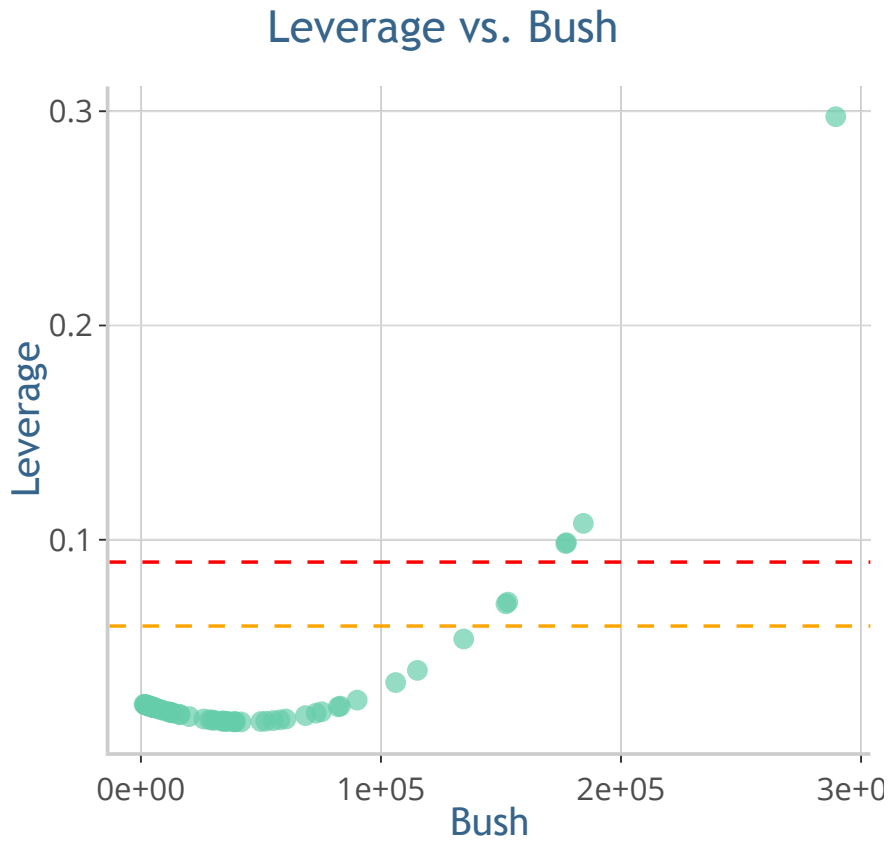




- None of the points has leverage values exceeding the two cutoffs $4/n = 4/29 = 0.138$ and $6/n = 6/29 = 0.207$.

Leverage - Example 2



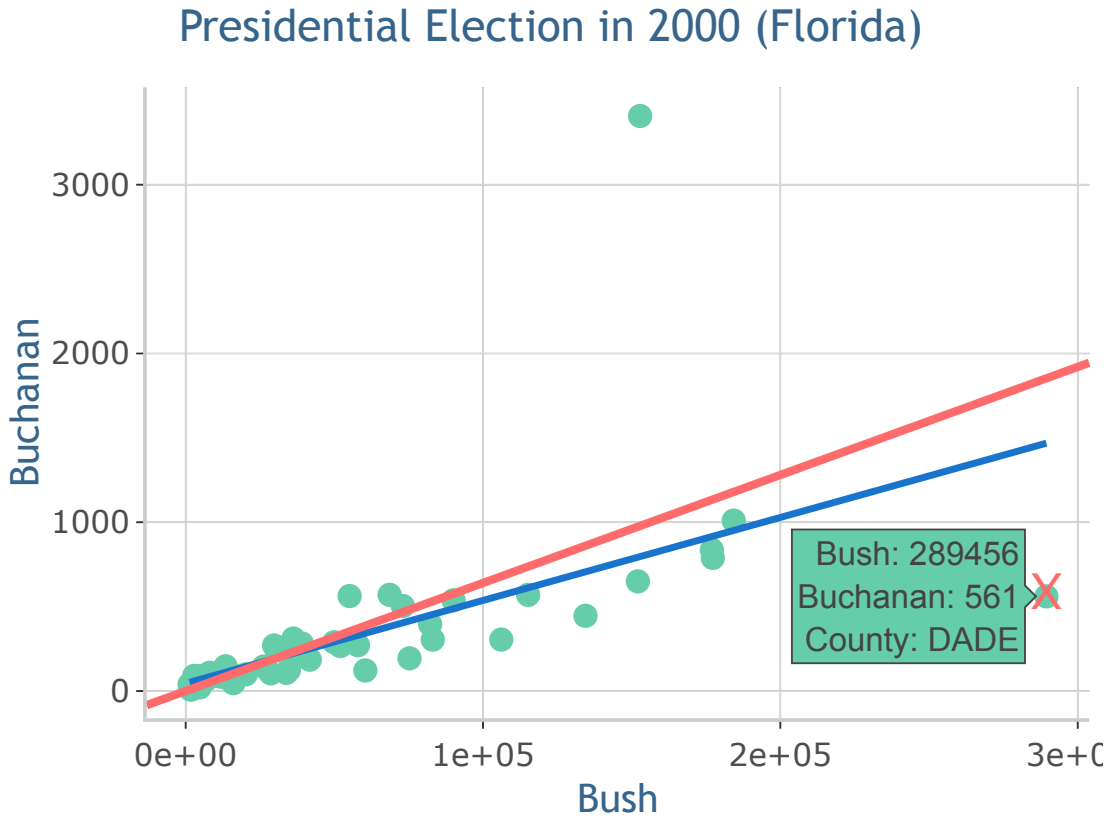


- 6 points have leverages values greater than $4/n = 4/67 = 0.060$, among which, 4 exceeds $6/n = 6/67 = 0.090$.

Leverage - Example 2

Regression line

- ○ With all data
- ◉ Without Dade
- ○ Without Palm Beach
- ○ Without Dade and Palm Beach



- Link

Standardized and Studentized residuals

The **standardized residual** for the i^{th} data point in a regression model can be computed using

$$\text{stdres}_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

where $\hat{\sigma}$ is the residual standard deviation and h_i is the leverage for the i^{th} point.

For a **studentized (or delete-t) residual**, we replace $\hat{\sigma}$ with the residual standard deviation, $\hat{\sigma}_{(i)}$, from fitting the model with the i^{th} point omitted:

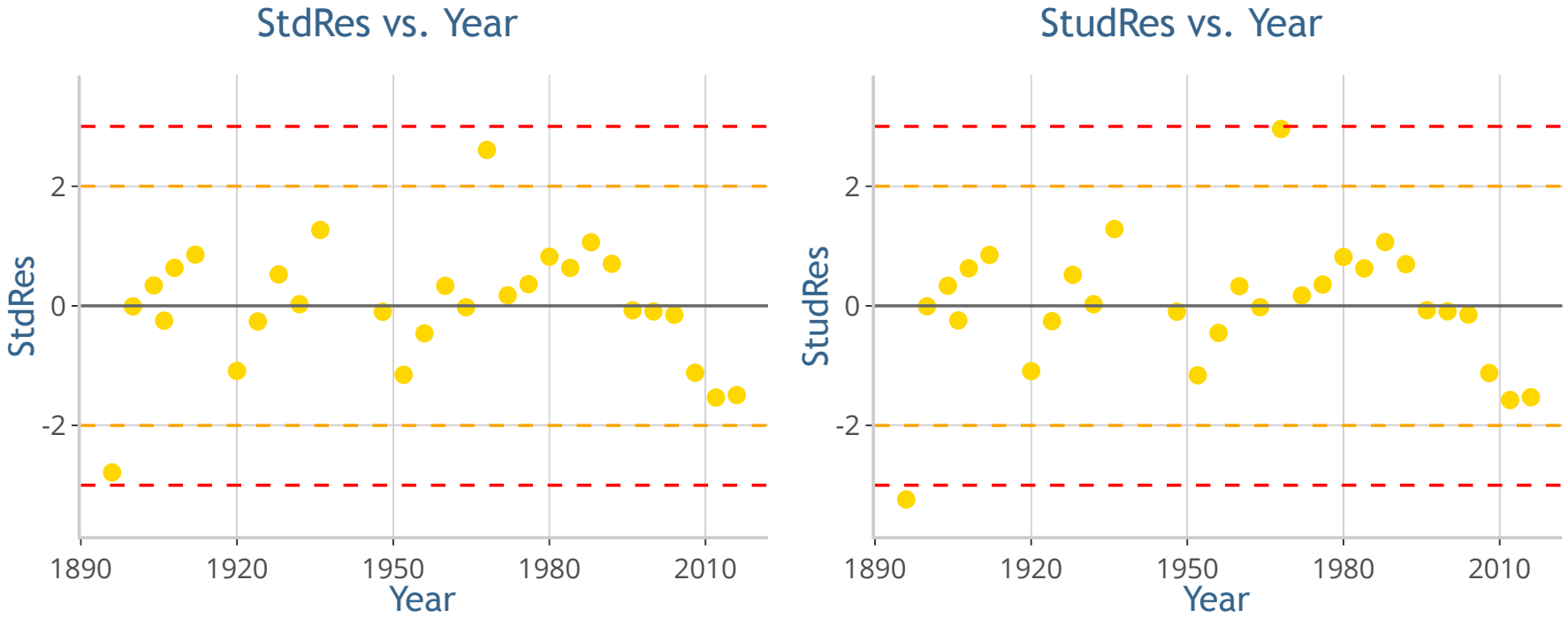
$$\text{studres}_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_i}}$$

Standardized and Studentized residuals

$$\text{stdres}_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1-h_i}}, \quad \text{studres}_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}}$$

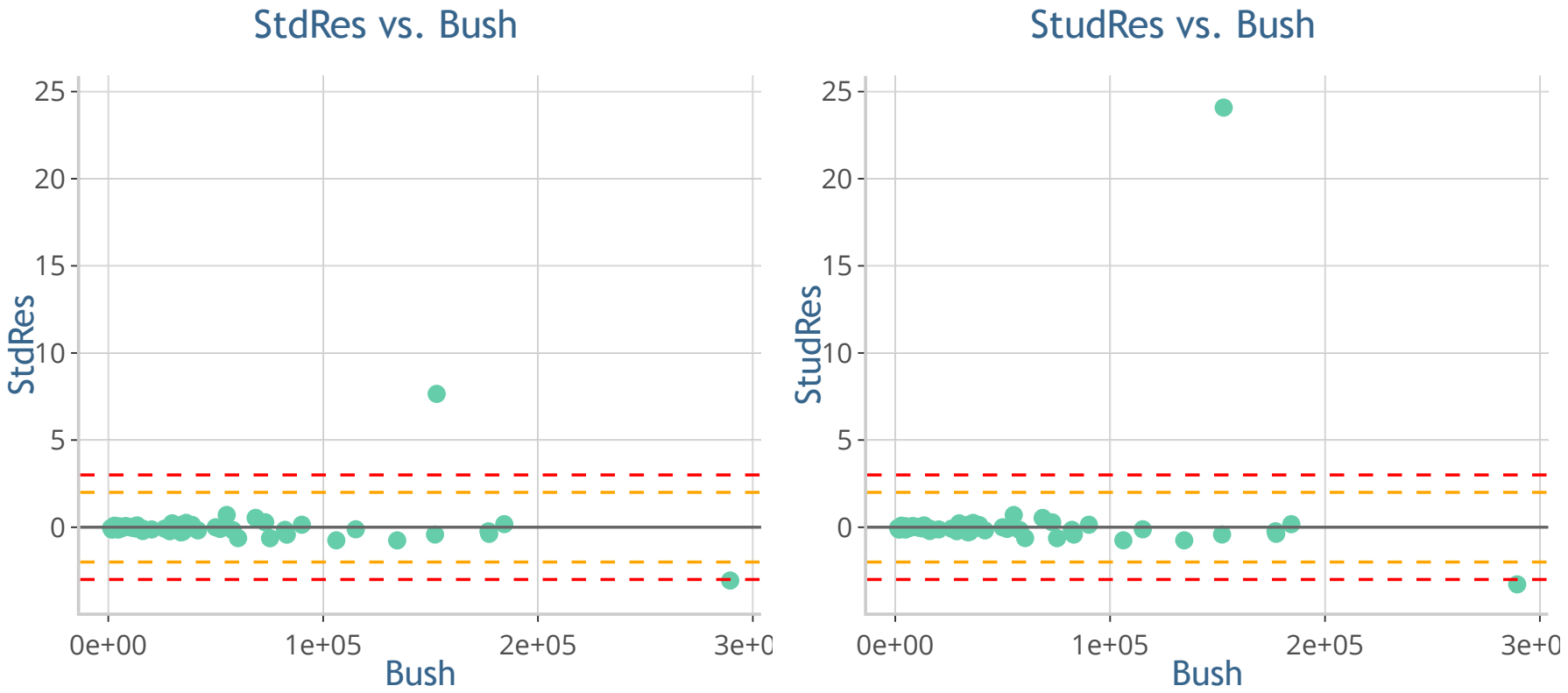
- The adjustment in the standard deviation for the studentized residual helps avoid a situation where a very influential data case has a big impact on the regression fit, thus artificially making its residual smaller.
- Under the usual conditions for the regression model, the standardized or studentized residuals follow a t-distribution.
- We identify data points with standardized or studentized residuals
  - beyond ± 2 as moderate outliers
  - beyond ± 3 as more serious outliers.

StdRes and StudRes - Example 1



- The point for 1896 has a large negative residual - moderate to serious outlier.
- The point for 1968 has a large positive residual - moderate outlier.
- The studentized residuals usually magnifies the residual values of points with large standardized residuals but do not influence other points much.

StdRes and StudRes - Example 2



- The point for the Palm Beach county has a very large positive residual - extremely serious outlier.
- The point for the Dade county has a large negative residual - serious outlier.

Cook's Distance

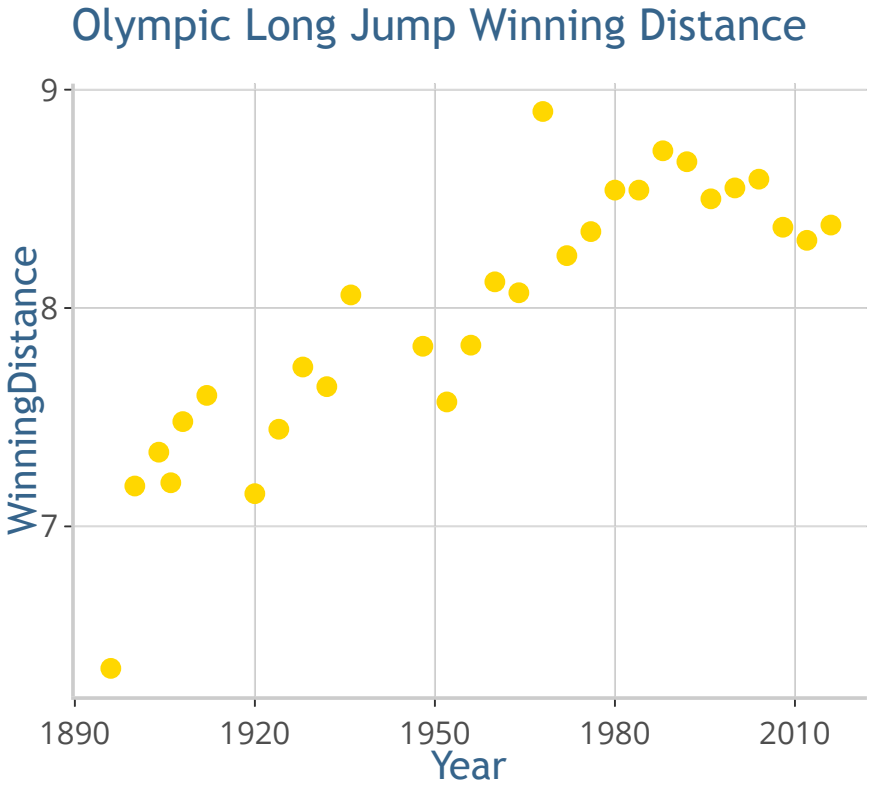
The **Cook's distance** of a data point in a simple linear regression is given by

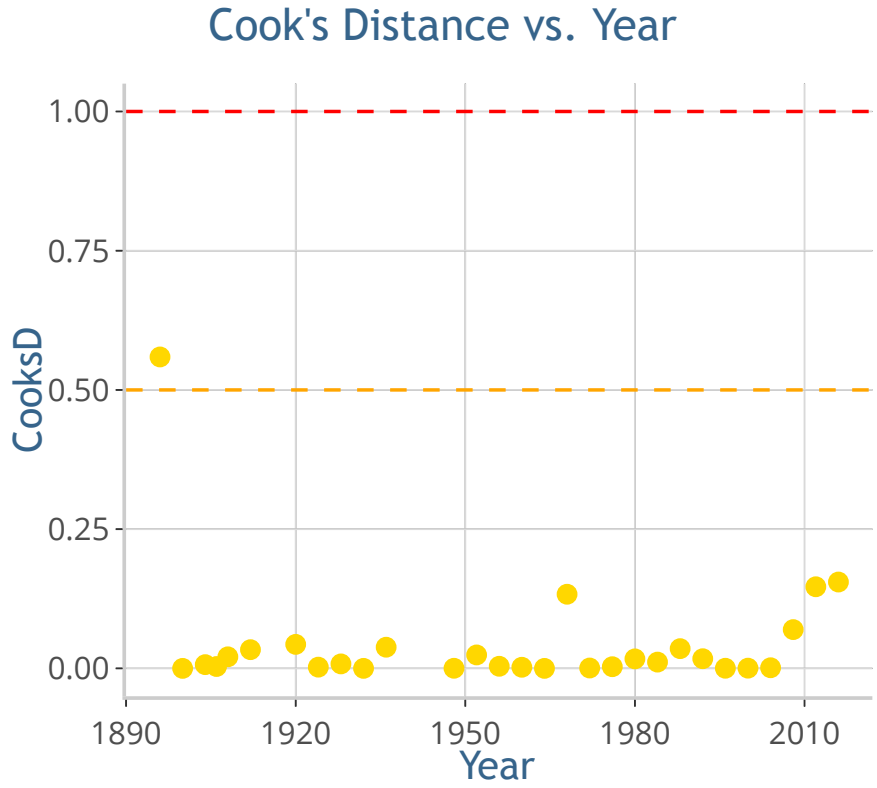
$$D_i = \frac{(\text{stdres}_i)^2}{2}\left(\frac{h_i}{1-h_i}\right)$$

where stdres_i and h_i are the standardized residual and leverage of the i^{th} data point.

- Cook's D measures the amount of influence that a particular data point has on the regression line. It depends on
  - how close the point lies to the trend of the rest of the data (as measured by its standardized or studentized residual) and
  - the leverage of the point (as measured by h_i).
- A large Cook's D occurs with a large standardized residual, a large leverage, or some combination of the two.
  - $D_i > 0.5$ indicates a moderately influential point
  - $D_i > 1$ indicates an especially influential point

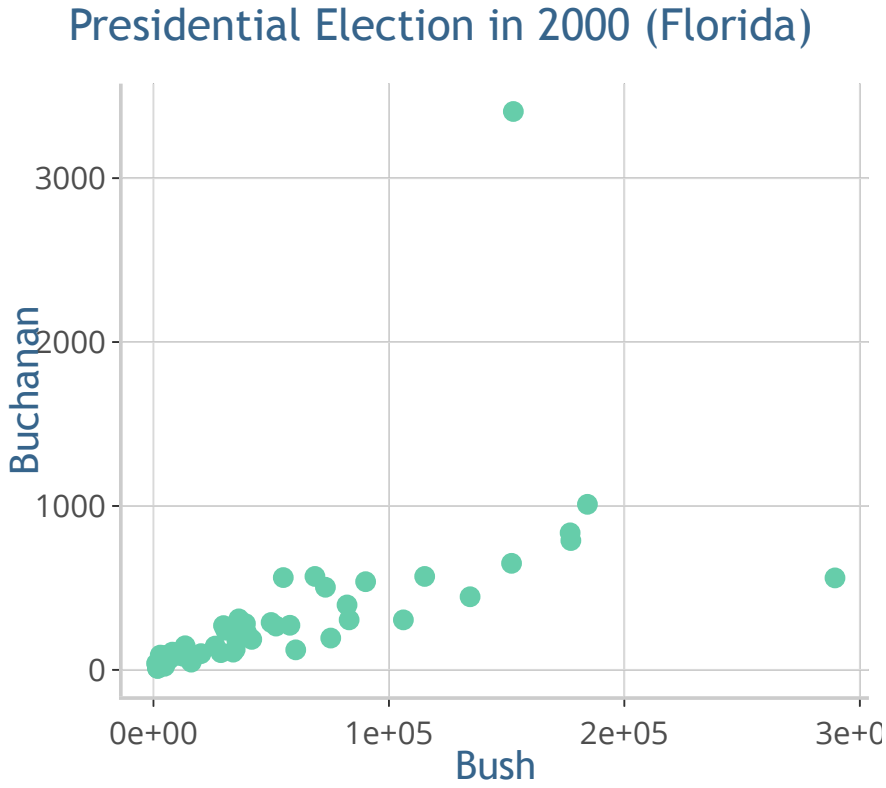
Cook's Distance - Example 1

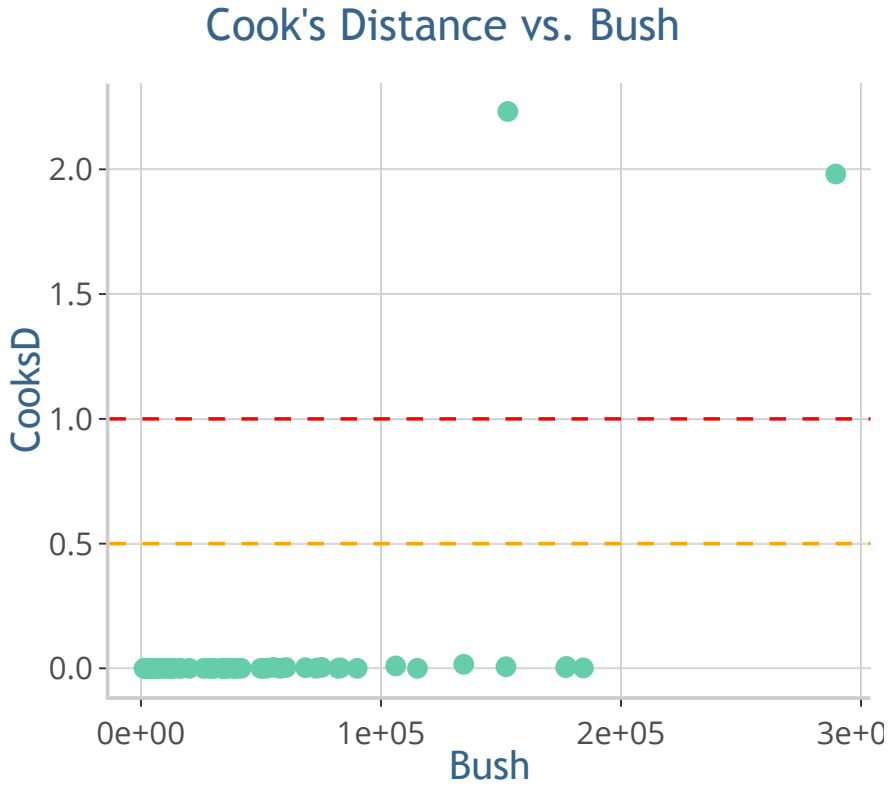




- The point for 1896 has the largest Cook's D value - moderately influential.
- The point for 1968 with a large positive residual has relatively small Cook's D value because it has small leverage.

Cook's Distance - Example 2





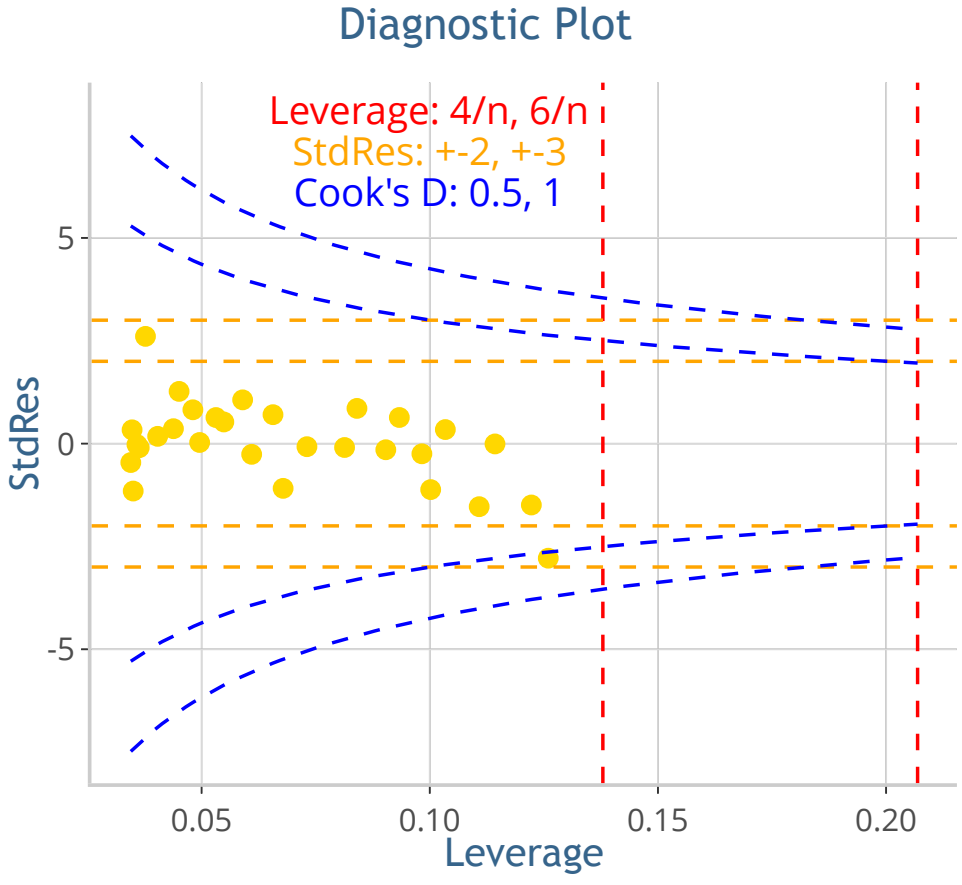
- The points for the Palm Beach county and the Dade county both have large leverages and large residuals and thus large Cook's D - especially influential.

Summarizing the three diagnostic statistics

For a simple linear regression model with n data points:

Statistic	Moderately unusual	Very unusual
Leverage, h_i	$> 4/n$	$> 6/n$
Standardized residual, stdres$_i$	beyond ± 2	beyond ± 3
Studentized residual, studres$_i$	beyond ± 2	beyond ± 3
Cook's distance, D_i	> 0.5	> 1

Three statistics in one diagnostic plot



- x-axis: leverage (cutoff $4/n$ and $6/n$)
- y-axis: standardized residuals (cutoff ± 2 and ± 3)
- Cook's D cutoff lines ($c = 0.5$ or 1) are found by

$$
\begin{aligned}
D_i &= \frac{(\text{stdres}_i)^2}{2} \frac{h_i}{1-h_i} > c \\
(\text{stdres}_i)^2 &> c \times 2 \frac{1-h_i}{h_i} \\
\text{stdres}_i &> \sqrt{2c \frac{1-h_i}{h_i}} \text{ or} \\
\text{stdres}_i &< -\sqrt{2c \frac{1-h_i}{h_i}}
\end{aligned}
$$

Three statistics in one diagnostic plot

```r
# Sample size
n <- nrow(election)

# SLR model
election.model <- lm(Buchanan ~ Bush, data=election)

# Pacakge for stdres() and studres()
library(MASS)

# Calculate the statistics
Leverage <- hatvalues(election.model) # leverage values
StdRes <- stdres(election.model) # standardized residuals
StudRes <- studres(election.model) # studentized residuals
CooksD <- cooks.distance(election.model) # Cook's distance

# Put all data into one dataframe
election <- data.frame(election, Leverage, StdRes, StudRes, CooksD)
```

Three statistics in one diagnostic plot

```
# Fina unusual points in the data
subset(election, Leverage > 4/n)
```

```
##    County Buchanan   Bush Leverage   StdRes   StudRes   CooksD
## 13   DADE      561 289456 0.297473 -3.05918 -3.280922 1.981366
```

```
subset(election, abs(StdRes) > 2)
```

```
##        County Buchanan   Bush   Leverage    StdRes   StudRes   CooksD
## 13       DADE      561 289456 0.29747301 -3.059180 -3.280922 1.981366
## 50 PALM BEACH     3407 152846 0.07085197  7.651072 24.080144 2.231935
```

```
subset(election, abs(StudRes) > 2)
```

```
##        County Buchanan   Bush   Leverage    StdRes   StudRes   CooksD
## 13       DADE      561 289456 0.29747301 -3.059180 -3.280922 1.981366
## 50 PALM BEACH     3407 152846 0.07085197  7.651072 24.080144 2.231935
```

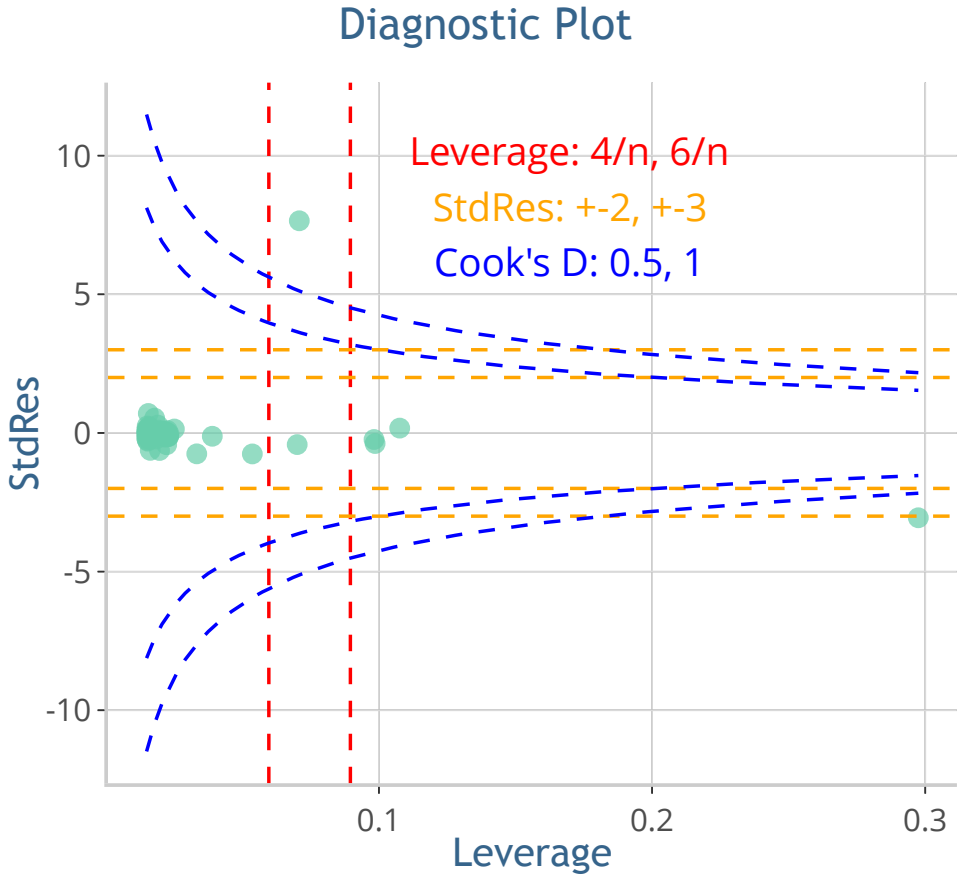
```
subset(election, CooksD > 0.5)
```

```
##        County Buchanan   Bush   Leverage    StdRes   StudRes   CooksD
## 13       DADE      561 289456 0.29747301 -3.059180 -3.280922 1.981366
## 50 PALM BEACH     3407 152846 0.07085197  7.651072 24.080144 2.231935
```

Three statistics in one diagnostic plot

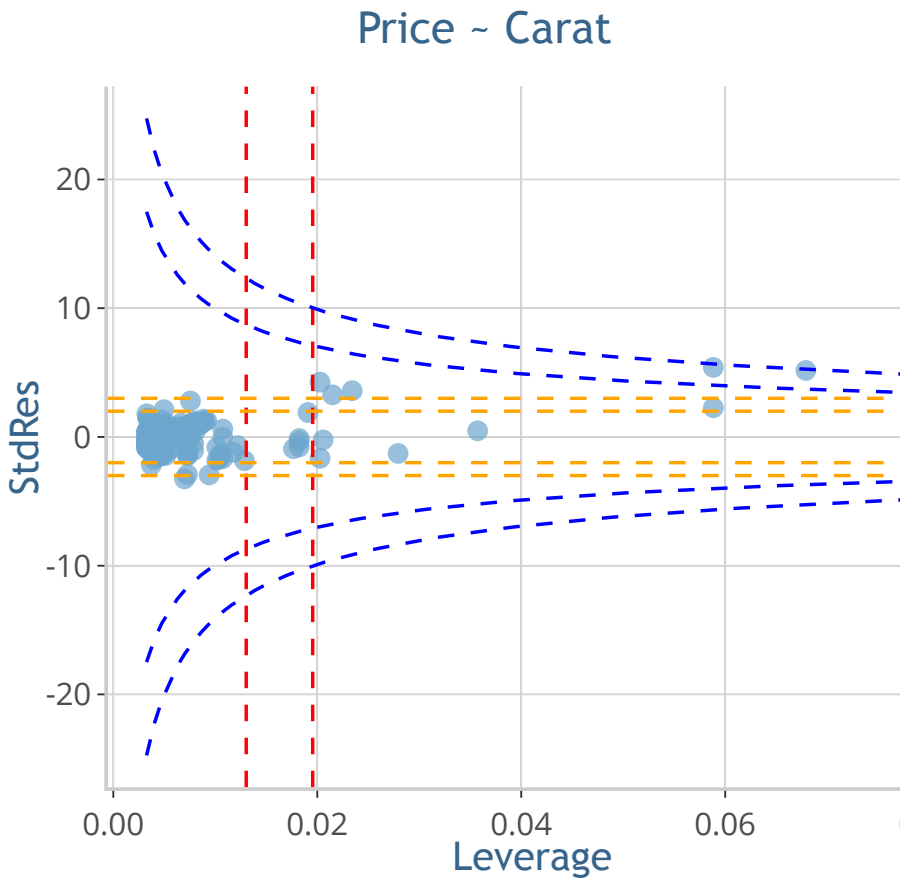
```
# Function for Cook's D cutoffs
cd <- function(h, type){
  sqrt((1-h)/h)*type
}
# Plot the ggplot
diag.plot <- ggplot(election, aes(x=Leverage, y=StdRes, label=County))+
  geom_point(color="aquamarine3", size=2.5)+
  geom_vline(xintercept = c(4/n, 6/n), color="red", linetype=2)+
  geom_hline(yintercept = c(-3,-2,2,3), color="orange", linetype=2)+
  stat_function(fun=cd, args=list(type=sqrt(2)), color="blue", linetype=2)+
  stat_function(fun=cd, args=list(type=-sqrt(2)), color="blue", linetype=2)+
  stat_function(fun=cd, args=list(type=1), color="blue", linetype=2)+
  stat_function(fun=cd, args=list(type=-1), color="blue", linetype=2)+
  ggtitle("Diagnostic Plot")
# Print the ggplot
diag.plot
# # Print the interaction ggplot using ggplotly()
library(plotly) # package for ggplotly()
ggplotly(diag.plot)
```

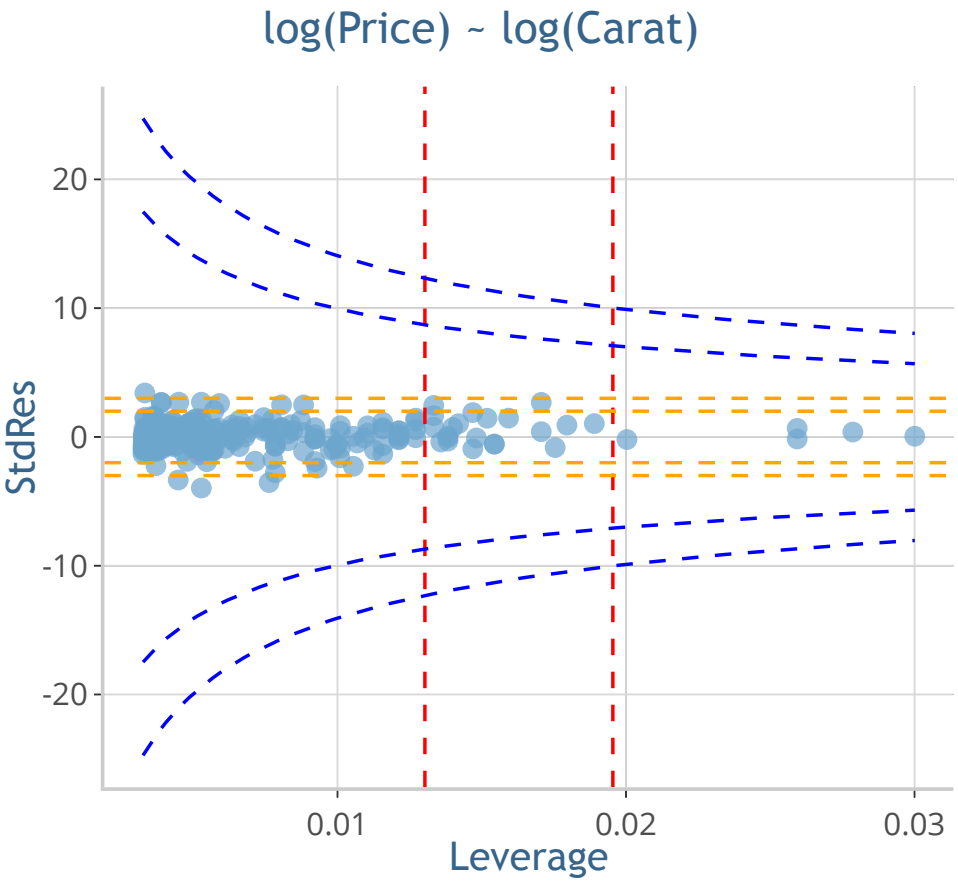
Three statistics in one diagnostic plot



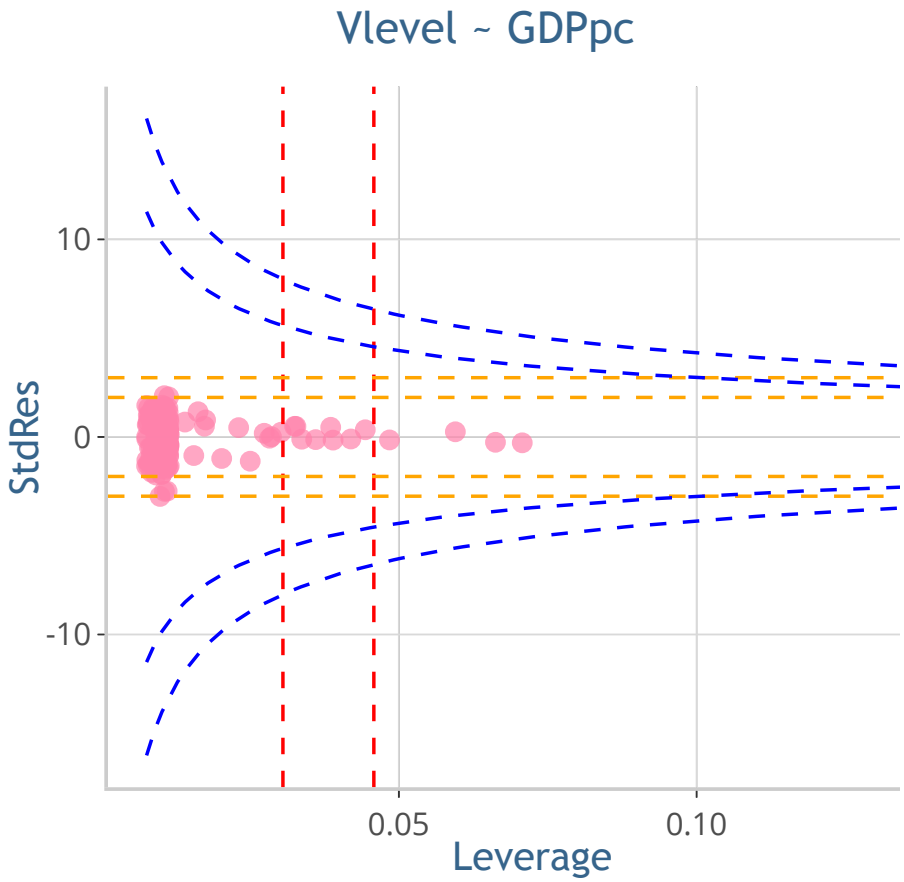
- The point *Palm Beach* has moderately high leverage, very high standardized residual and very unusal Cook's D value.
- The point *Dade* has extremely high leverage, very high standardized residual and very unusal Cook's D value.

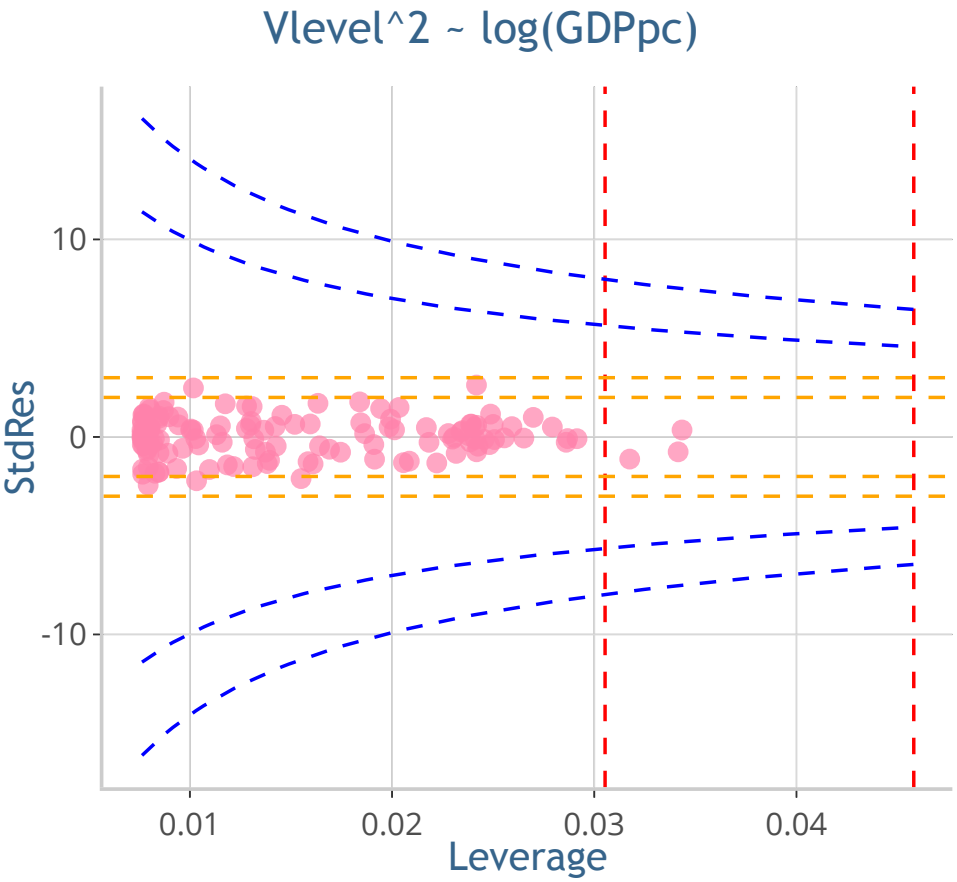
Application: Diamond Price vs. Carat





Application: Valentine's Day Vlevel vs. GDPpc





Some notes

- The goal of these diagnostic tools is to help us identify data points that might need further investigation.
  - Data errors?
  - Special cases? *Do an analysis with and without a suspicious point and see how the model is affected.* AVOID blindly deleting all unusual points until the data that remain are ``nice."
  - In many situations (like the butterfly ballot scenario), the most important features of the data would be lost if the unusual points were dropped from the analysis!

Summary

CHOOSE

- Exploratory data analysis; Model: $Y = \beta_0 + \beta_1 X + \epsilon$ where $\epsilon \overset{iid}{\sim} N(0, \sigma)$

FIT

- Maximum likelihood estimation (MLE)

ASSESS model

- Inference for the intercept and slope; ANOVA and R^2

ASSESS error

- Check conditions and transformations; Outliers and influential points

USE

- Predictions

Midterm Examination

- Time: 10/25/2018 Thursday in class
- Location: SC L26
- Lecture 1~13
- Mainly short answer questions similarly as Homework questions
  - No question about R programming/functions
  - But you should be able to understand R output.
- Closed-book; one two-sided letter size cheat sheet allowed.
- You'll need a calculator.
- Show your work and explain your reasoning.
- HW 6: given today and due on Monday 10/22 11:55 PM on Moodle
  - Solutions available on Tuesday 10/23