**Plan for Today**

- Another Linear Regression Example: Using 1996 to Predict 2000 Election Results (for the Reform Party Candidates)
  1. Load and explore data
  2. Identify and calculate (if necessary) the outcome (Y) and independent (X) variables
  3. Visualize the relationship between X and Y
  4. Fit a regression line using the least squares model
     - notice any anomalies? (e.g., outliers)
  5. Interpret the coefficients/regression estimates
  6. Calculate $R^2$: measure of prediction strength/model fitness

---

**Example: Using 1996 Results to Predict 2000**

- We are going to use the electoral outcome of Perot in 1996 to predict the electoral outcome of Buchanan in 2000
  - both were the reform party candidates
- Please download `gvt_201_11.R` and `florida.csv`, saved them in your GVT201 folder, and open `gvt_201_11.R`
1. Load and explore data

```r
data <- read.csv("florida.csv")
head(data)
```

- **counties**: county names
- **Clinton96, Dole96, Perot96**: number of votes received by each candidate in each county in 1996
- **Bush00, Gore00, Buchanan00**: number of votes received by each candidate in each county in 2000

What is the unit of analysis? In other words, what does each observation in the data represent?

- if you cannot figure it out based on the codebook and/or by simply looking at the data, find the one variable that its number of unique values equals the number of observations in the data

```r
dim(data) # dimension of data: rows, columns
## [1] 67 7
length(unique(data$counties))
## [1] 67
```

- in this case, counties has as many unique observations as the dataset itself so it is the identifier (the variable that identifies each observation)
- so each observation in this dataset represents? a county

2. Identify and calculate (if necessary) outcome and independent variables

- what is our outcome variable (Y)?
  - what do we want to predict?
  - electoral outcome of Buchanan in 2000 (Buchanan00)
- what is our independent variable (X)?
  - what are we going to use as the basis for our prediction?
  - electoral outcome of Perot in 1996 (Perot96)

3. Visualize the relationship between X and Y

What is the code? Complete the code in the R file

```r
plot(y=data$Buchanan00, x=data$Perot96)
```

- Does the relationship look positive or negative?
  - cor(X,Y)>0 or cor(X,Y)<0?
  - expected sign of \( \hat{\beta} \)?

4. Fit a regression line using the least squares model

What is the code? Complete the code in the R file

```r
Buchanan 2000_i = \hat{\alpha} + \hat{\beta} \text{ Perot 1996}_i \text{ (where } i=\text{counties)}
```

use `lm()` function in R to estimate the coefficients

```r
regression <- lm(data$Buchanan00 ~ data$Perot96)
regression
```

- Are there any anomalies?
5. Interpret Coefficients/Regression Estimates
   ▶ Let's focus on the model without the outlier

   `regression2` # call the `lm()` object

   ```r
   ## Call:
   ## lm(formula = Buchanan00 ~ Perot96, data = data2)
   ##
   ## Coefficients:
   ## (Intercept)     Perot96
   ##      45.84193      0.02435
   ##
   ## What is then the estimated model?
   Buchanan 2000, \( \hat{Y} = 46 + 0.02 \text{ Perot 1996} \), where \( i \) = counties
   ```

   ▶ Identify outlier and look at whether there is an explanation
   ▶ `resid()` or `residuals()` with `lm()` object inside will calculate the residuals of the regression

   ```r
   # [ADVANCED] we subset the name of counties to the one
   # for which the residual is the largest in the data
   data$counties[resid(regression) == max(resid(regression))]
   ```

   ▶ It turns out that Palm Beach county used a confusing ballot in 2000 that resulted in many voters mistakenly placing their vote for Buchanan when they, in fact, intended to vote for Al Gore

   ▶ What happens when we remove the outlier?

   ```r
   data2 <- data[data$counties != "PalmBeach",]
   regression2 <- lm(Buchanan00 ~ Perot96, data = data2)
   plot(y=data$Buchanan00, x=data$Perot96, col="gray")
   abline(regression, col="red", lty=2) # with outlier
   abline(regression2, col="red") # without outlier
   ```

   removing the outlier shifts the regression line considerably

   Buchanan 2000, \( \hat{Y} = 46 + 0.02 \text{ Perot 1996} \)

   ▶ Interpretation of \( \hat{\beta} \)?
   ▶ mathematically: \( \Delta \hat{Y} \) when \( \Delta X = 1 \)
   ▶ in this case: an additional vote in favor of Perot in the 1996 elections is associated with 0.02 more votes for Buchanan in 2000, on average
   ▶ then: an additional 100 votes in favor of Perot in the 1996 elections would be associated with how many more expected votes for Buchanan in 2000?
     ▶ 2 more votes (100*0.02=2)
   ▶ Interpretation of \( \hat{\alpha} \)?
   ▶ mathematically: \( \hat{Y} \) when \( X = 0 \)
   ▶ in this case: the average expected number of votes for Buchanan in 2000 in counties where Perot received no votes in 1996 is of 46 votes
6. Calculate $R^2$ for both models and compare them

```r
summary(regression)$r.squared # with Palm Beach
## [1] 0.5130333

summary(regression2)$r.squared # without Palm Beach
## [1] 0.8511675
```

- what does it mean for $R^2 = 0.85$?
  - the model explains 85% of the variation of Y
- which line provides a better fit for the data?
  - once we remove the outlier, the model fit improves significantly (i.e., becomes closer to 1)
  - this makes sense since we removed the observation with the largest error
- SSRT without outlier < SSRT with outlier
- $R^2$ without outlier > $R^2$ with outlier

---

**Today’s Class and Next**

**Today**
- Used linear regression for another prediction example
- New concept: outliers

**Next Class**
- An example of how to use linear regression to estimate the causal effect of $X$ on $Y$