version: uploaded 05162019

Correlation and Regression

Week 5

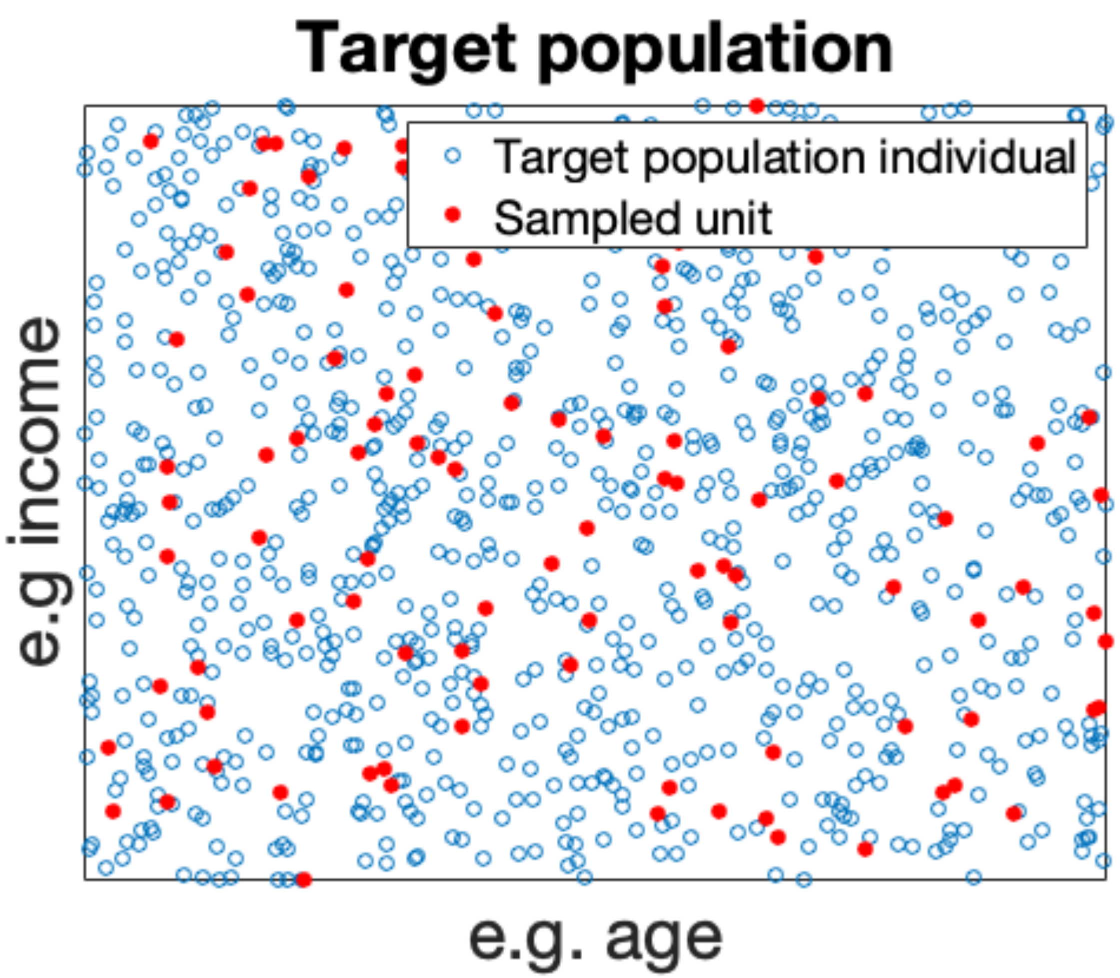
Yunkyu Sohn
School of Political Science and Economics
Waseda University

2019 Spring Statistics I

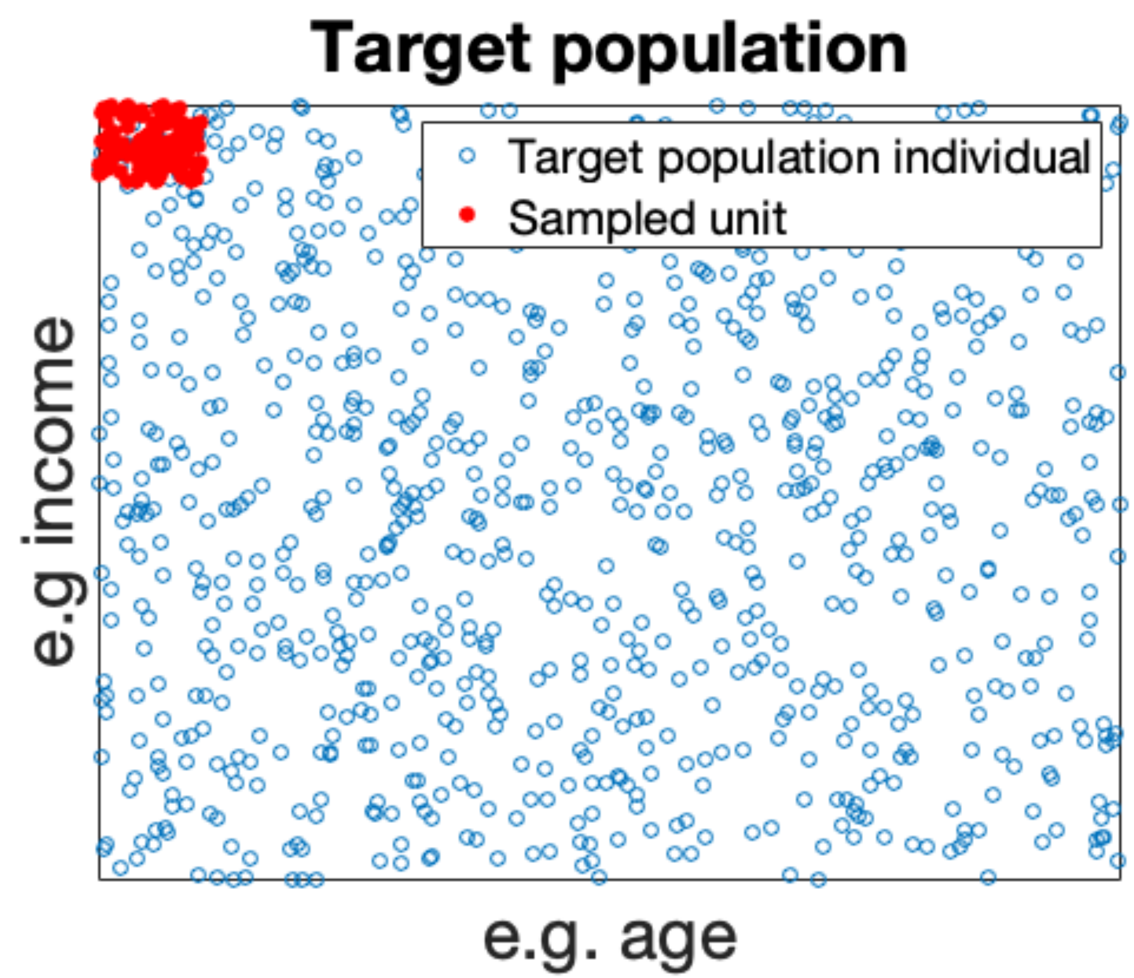
Contents (Book Chapter 3.6, 4.2)

Formalizing Survey Sampling: Basics

- Target population: Population of interest
  - e.g. entire eligible voters in a district; entire females in Waseda
- Sample selection bias: bias coming from sampling
  - not being **representative** of the target population



VS



Formalizing Survey Sampling: Basics

- Target population (TP): Population of interest
  - e.g. entire eligible voters in a district; entire females in Waseda
- Sample population (SP): Sub-population of TP being sampled
- Sampling frame:
  - Complete list of potential responders (may not contain all TP)
    - TP list impossible to obtain in most cases
- Sample selection bias: bias coming from the sample
  - not being representative of TP

Formalizing Survey Sampling: Basics

- Probability sampling:
  - every unit in TP has non-zero chance of being sampled
    - to ensure representativeness
- Simple random sampling:
  - Predetermined number sampled with each potential respondent having equal chance of being sampled
  - done without replacement (at most one interview per person)

Formalizing Survey Sampling: Sources of Bias

Population	Sample	Potential Bias
Target population		Frame bias
$\downarrow$		
Frame population $\rightarrow$	Sample	
	$\downarrow$	Unit non-response
	Respondents	
	$\downarrow$	Item non-response
	Completed items	
		Response bias

- Frame bias
  - e.g. cell phone contacts ➠ wealth, occupation
  - e.g. Internet surveys ➠ age, wealth
  - e.g. opt-in panels ➠ traits correlated with willingness

Formalizing Survey Sampling: Sources of Bias

Population	Sample	Potential Bias
Target population		Frame bias
$\downarrow$		
Frame population $\rightarrow$	Sample	
	$\downarrow$	Unit non-response
	Respondents	
	$\downarrow$	Item non-response
	Completed items	
		Response bias

- Unit non-response
  - e.g. cell phone contacts ➠ caller ID screening (traits)
  - e.g. offline canvassing ➠ occupation, age

Formalizing Survey Sampling: Sources of Bias

Population	Sample	Potential Bias
Target population		Frame bias
$\downarrow$		
Frame population $\rightarrow$	Sample	
	$\downarrow$	Unit non-response
	Respondents	
	$\downarrow$	Item non-response
	Completed items	
		Response bias

- Item non-response
  - e.g. sensitive questions (orientation, religion)
  - e.g. language problems (region of origin, age)

Formalizing Survey Sampling: Sources of Bias

Population	Sample	Potential Bias
Target population		Frame bias
$\downarrow$		
Frame population $\rightarrow$	Sample	
	$\downarrow$	Unit non-response
	Respondents	
	$\downarrow$	Item non-response
	Completed items	
		Response bias

- Response bias
  - e.g. sensitive questions (orientation, religion)
  - e.g. turnout surveys (social desirability bias)

Example: Public Opinion Survey in War Zones

- Asking sensitive questions in Afghanistan
  - Respondents at risk of providing truthful answers
  - Institutional Review Board does not allow direct questionaries
- Statistical solution:
  - List experiment or Item count technique
    - For a set of respondents ask a non-sensitive question
    - For another set of respondents ask a slightly different question
    - Use Difference-in-means estimator
      - All other responses equal except for the single difference

Example: Public Opinion Survey in War Zones

- Script for the control group:

  `I'm going to read you a list with the names of different groups and individuals on it. After I read the entire list, I'd like you to tell me how many of these groups and individuals you broadly support, meaning that you generally agree with the goals and policies of the group or individual. Please don't tell me which ones you generally agree with; only tell me how many groups or individuals you broadly support.`

  `Karzai Government; National Solidarity Program; Local Farmers`

Example: Public Opinion Survey in War Zones

- Script for the treatment group:

  `I’m going to read you a list with the names of different groups and individuals on it. After I read the entire list, I’d like you to tell me how many of these groups and individuals you broadly support, meaning that you generally agree with the goals and policies of the group or individual. Please don’t tell me which ones you generally agree with; only tell me how many groups or individuals you broadly support.`

  `Karzai Government; National Solidarity Program; Local Farmers; ISAF (Taliban)`

Example: Public Opinion Survey in War Zones

- Difference-in-means estimator
  - Vector for the numbers of items chosen by the treatment group

```
afghan$list.response[afghan$list.group == "ISAF"]
```

  - Vector for the numbers of items chosen by the control group

```
afghan$list.response[afghan$list.group == "control"]
```

`Karzai Government; National Solidarity Program; Local Farmers; ISAF (Taliban)`

  - Proportion of those who support ISAF:

```
mean(afghan$list.response[afghan$list.group == "ISAF"]) -
    mean(afghan$list.response[afghan$list.group == "control"])

## [1] 0.04901961
```

Review: Summarizing a Univariate Distribution

- Age distribution of sample B

Age	Frequency
16	5
18	10
20	11
22	10
24	5

- Spread
  - Range
  - upper (lower) quartile
  - IQR: inter-quartile range
  - Standard deviation

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Review: Summarizing a Univariate Distribution

- Age distribution of sample A

Age	Frequency
18	5
19	10
20	11
21	10
22	5

- Spread
  - Range: [18, 22]
  - upper (lower) quartile: 21 (19)
  - IQR: 2
  - Standard deviation: $\sqrt{1.5}$

- Age distribution of sample B

Age	Frequency
16	5
18	10
20	11
22	10
24	5

- Spread
  - Range: [16, 24]
  - upper (lower) quartile: 22 (18)
  - IQR: 4
  - Standard deviation: $\sqrt{6}$

Visualizing a Univariate Distribution: Histogram

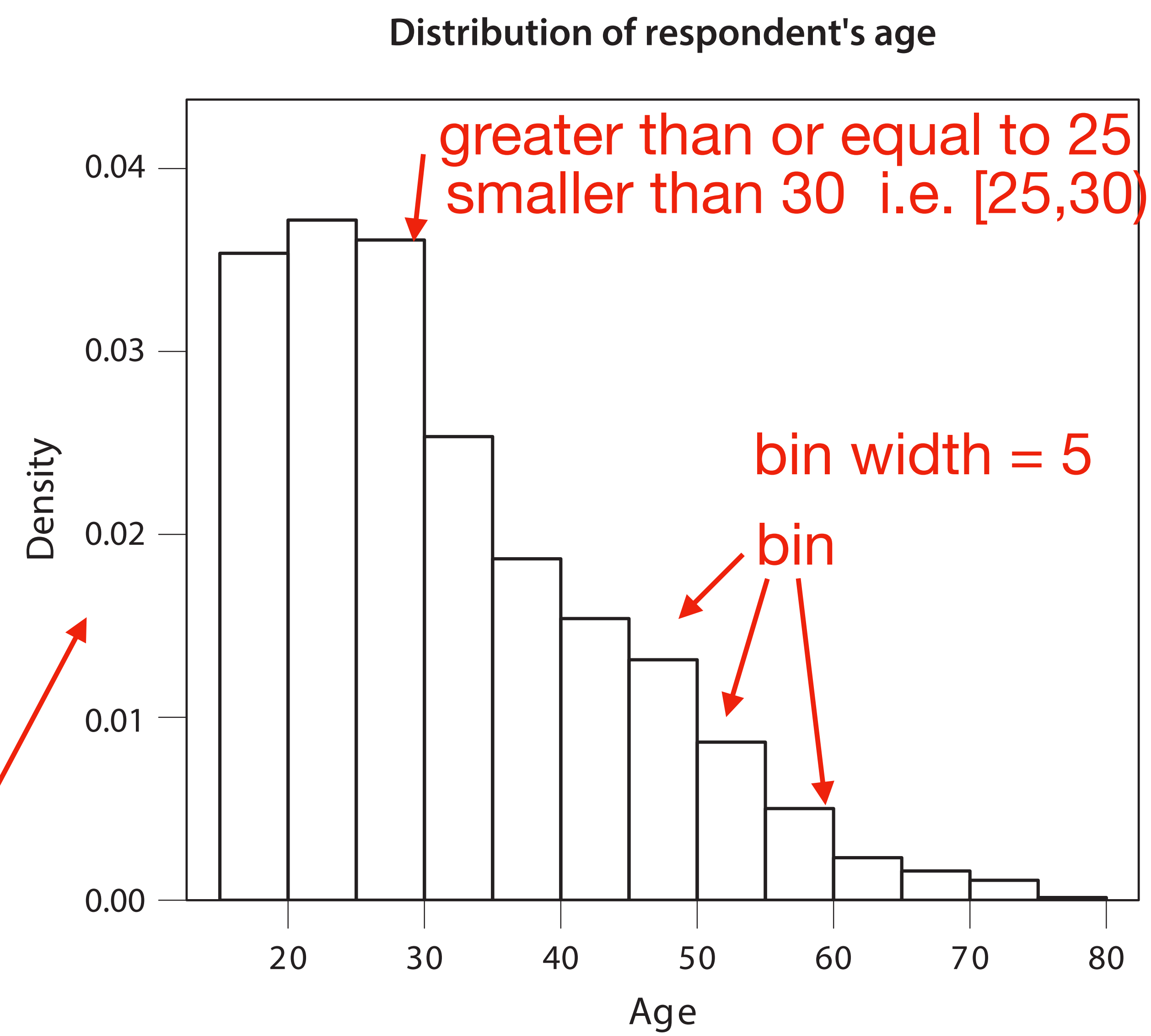
- Actual Afghanistan survey data in textbook

$$\text{density} = \frac{\text{proportion of observations in the bin}}{\text{width of the bin}}$$

$$= \frac{\textbf{Proportion of observations in the bin}}{5}$$

➠ sum of densities does not sum up to 1

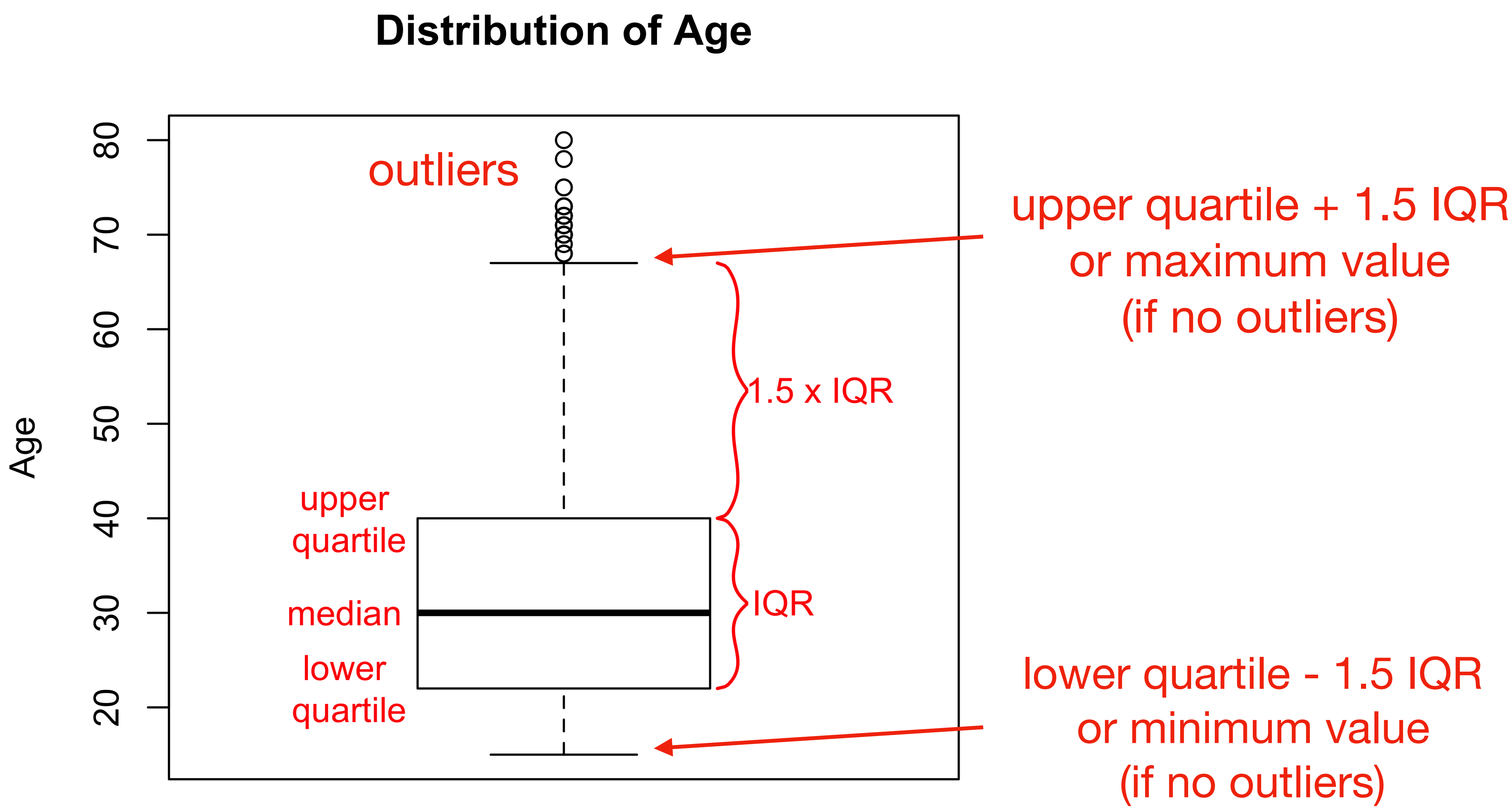
➠ sum of densities x bin width = 1




```r
hist(afghan$age, freq = FALSE, ylim = c(0, 0.04), xlab = "Age",
     main = "Distribution of Respondent's Age")
```

Visualizing a Univariate Distribution: Boxplot

- Actual Afghanistan survey data in textbook

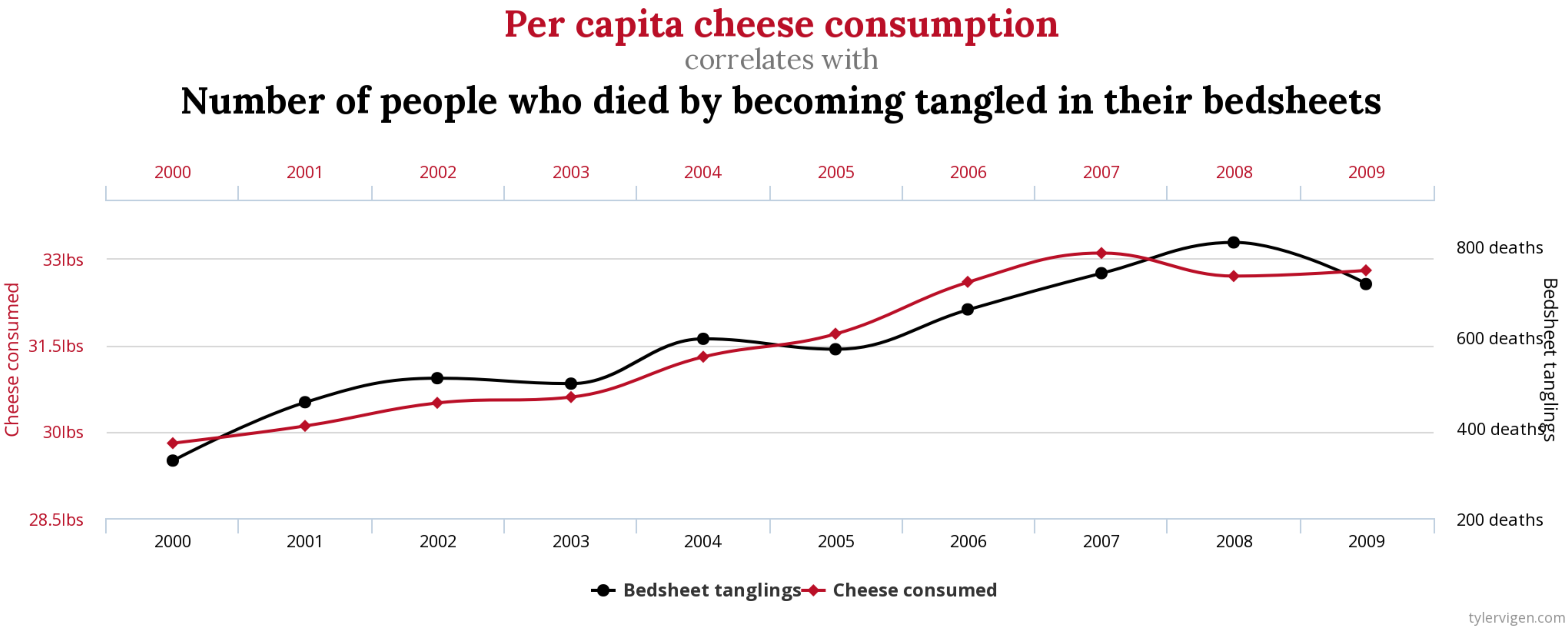


- Effective summary of a distribution (less informative than hist.)
  - compare multi distributions in compact manner

Contents (Book Chapter 3.6, 4.2)

- Review of survey sampling
- Correlation
  - z-score
  - Correlation coefficient
- Linear regression
  - Residuals
  - Lease squares
  - Example: Facial Competence and Vote Share

Correlation



- Correlation does not **always** guarantee causation
  - But causation yields correlation
    - We need measures for correlation

https://tylervigen.com/spurious-correlations

Correlation: Z-Score for Univariate Variable

- z-score (need to be defined before introducing a measure correlation)

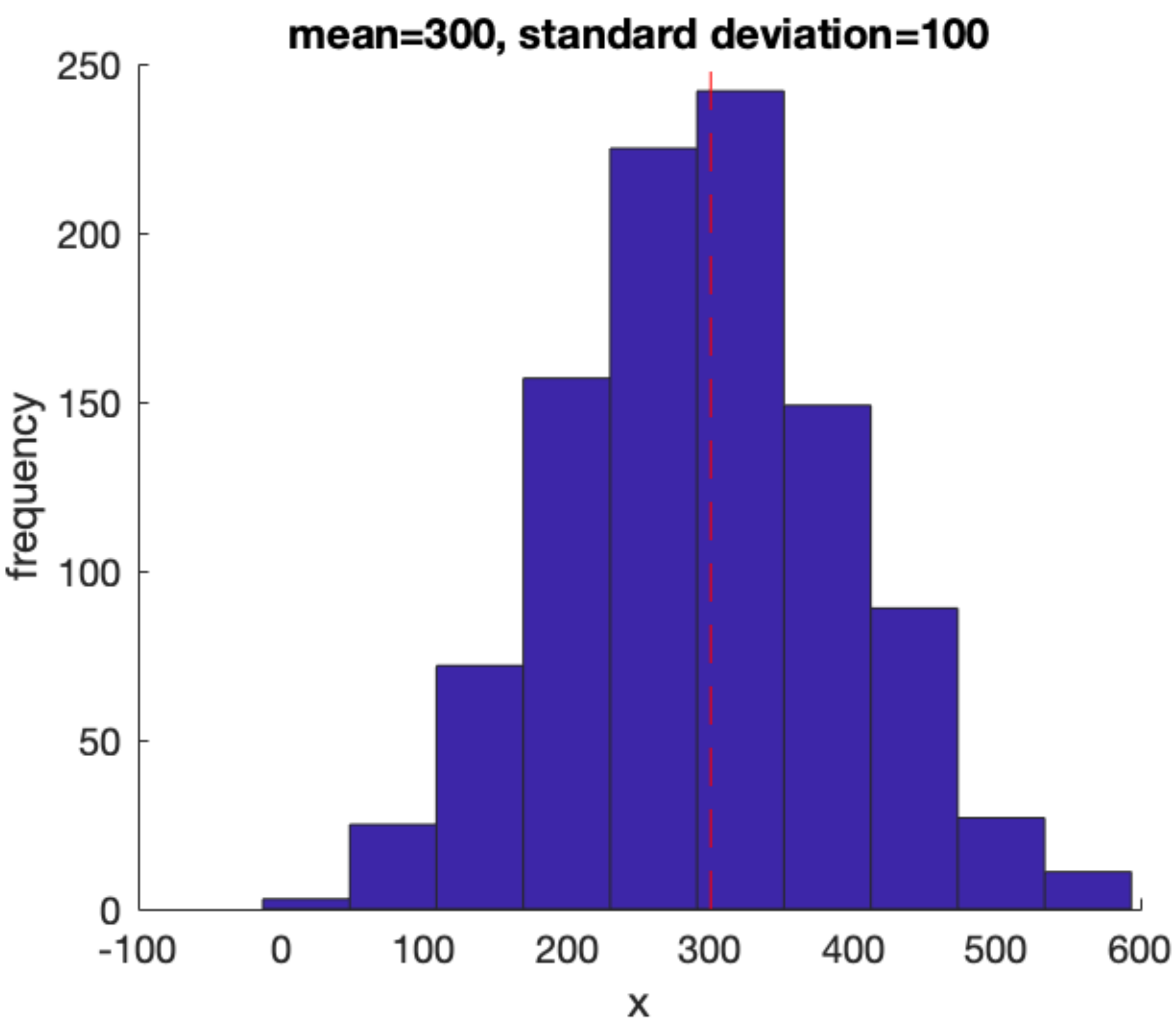
$$z\textbf{-score of } \mathbf{x}_i = \frac{x_i - \bar{x}}{S_x}$$

  - x_i: i-th observation of variable x
  - $\bar{x}$: mean of x
  - S_x: standard deviation of x
  - Meaning: the number of standard deviations an observation is above or below the mean
    - e.g. z-score: 2; z-score: -1.5

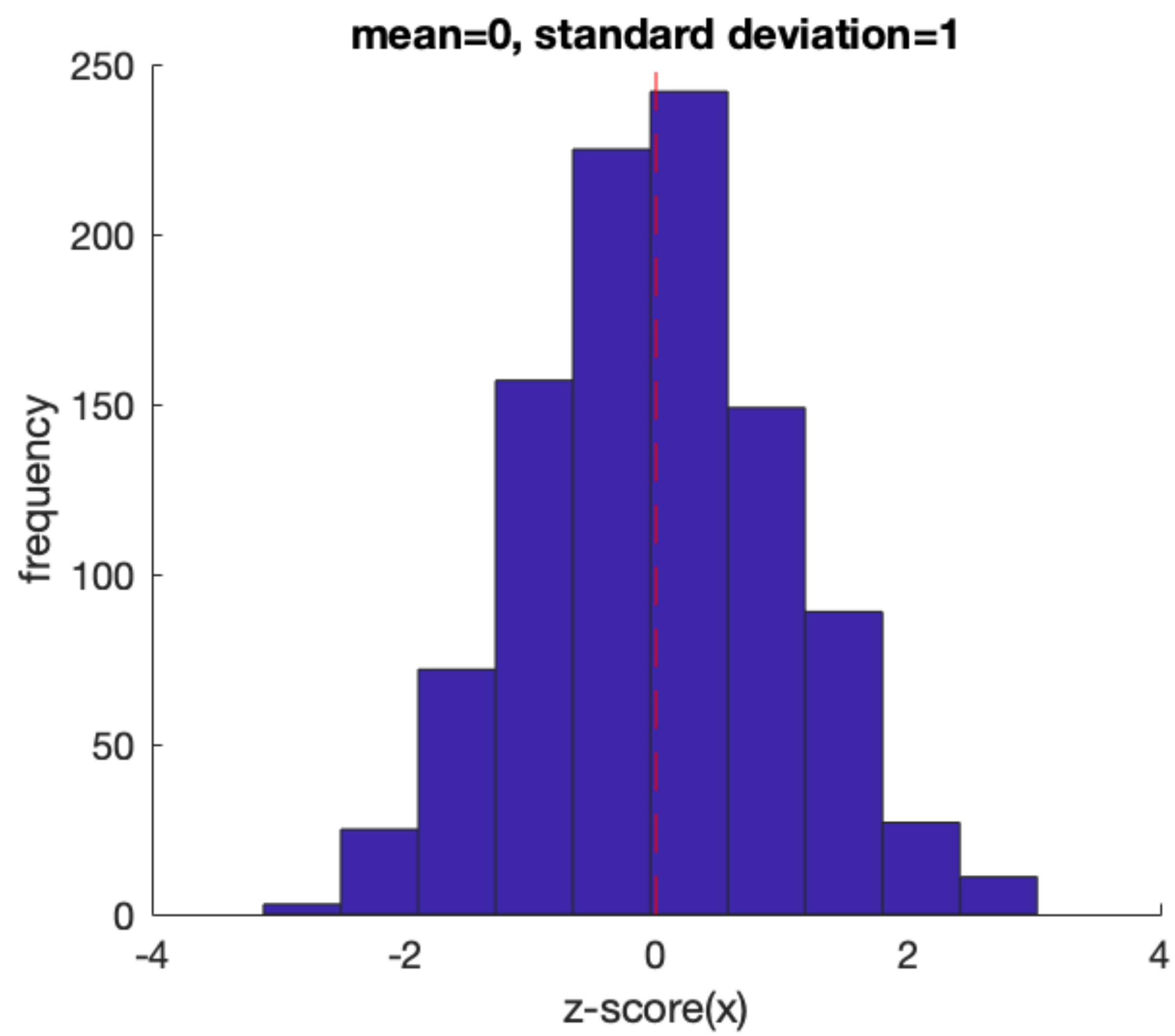
Correlation: Z-Score for Univariate Variable

- z-score

$$z\text{-score of } \mathbf{x}_i = \frac{x_i - \bar{x}}{S_x}$$



- Original



- Transformed

Correlation: Z-Score for Univariate Variable

- z-score (before introducing a measure correlation)

$$z\textbf{-score of } \mathbf{x}_i = \frac{x_i - \bar{x}}{S_x}$$

- Effect of scaling (a) and shift (b)

$$\begin{aligned} z\text{-score of } (ax_i + b) &= \frac{(ax_i + b) - \text{mean of } (ax + b)}{\text{standard deviation of } (ax + b)} \\ &= \frac{a \times (x_i - \text{mean of } x)}{a \times \text{standard deviation of } x} \\ &= z\text{-score of } x_i\,, \end{aligned}$$

Correlation: Bivariate Relationships

- Pearson correlation coefficient (use of n-1 will be discussed next week.)

$$\text{correlation}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(z\text{-score of } x_i \times z\text{-score of } y_i)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S_x} \times \frac{y_i - \bar{y}}{S_y}\right)$$

- Correlation is between -1 and 1
  - +/- represents positive/negative relationship
  - Absolute value represents magnitude of association
- Order does not matter
- Scale does not matter
- Correlation quantifies linear association

Numerical Exercise

Observation	1st sibling age	2nd sibling age
1	20	18
2	18	16
3	20	16
4	20	18

z**-score of $\mathbf{x}_i$** $= \dfrac{x_i - \bar{x}}{S_x}$

correlation(x, y) $=$

$$\frac{1}{n-1}\sum_{i=1}^{n}(z\text{-score of } x_i \times z\text{-score of } y_i)$$

(The use of n-1 will be discussed next week.)

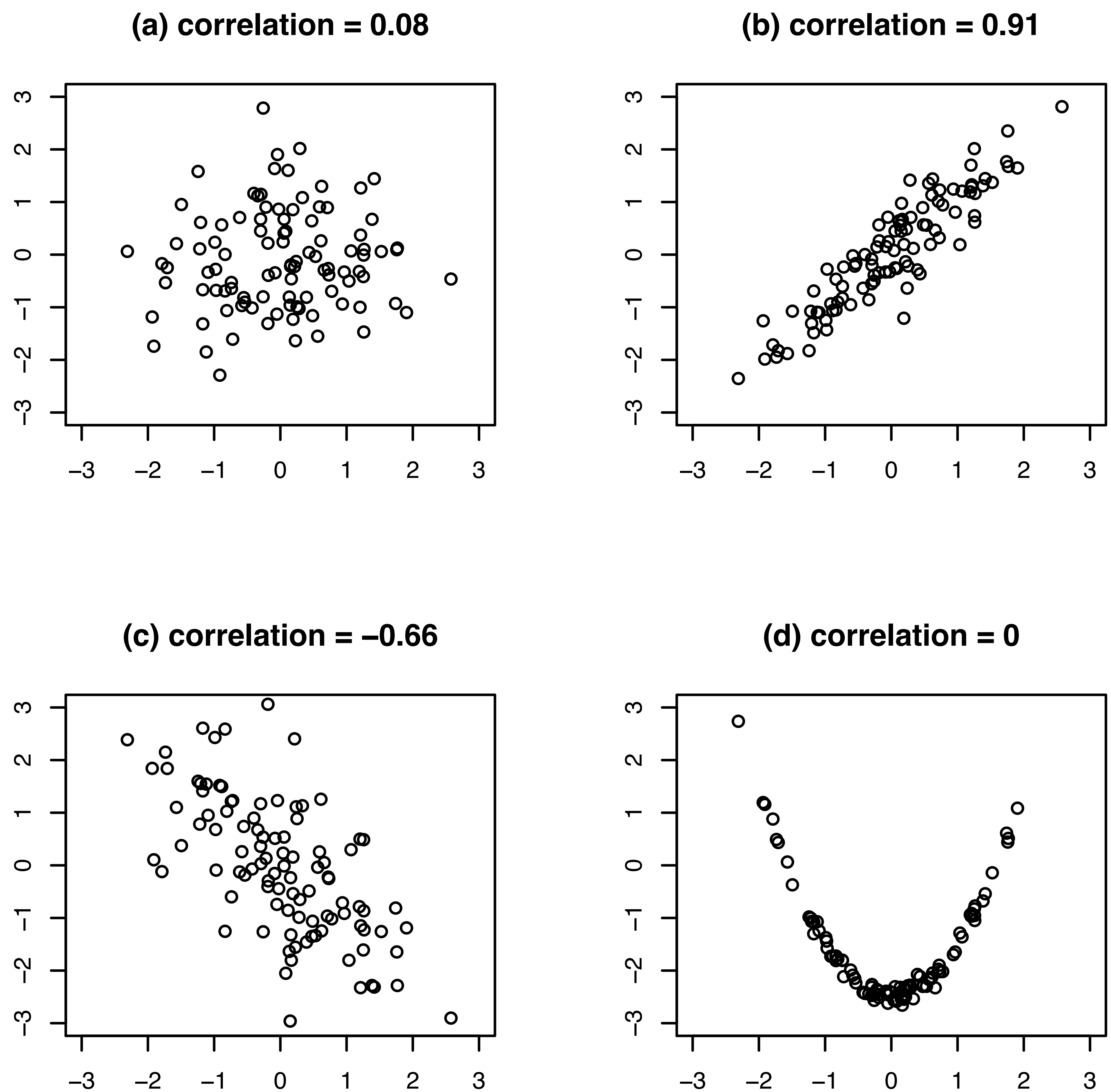
- z-score(1st sibiling age)
- z-score(2nd sibiling age)
- Correlation(1st sibiling age, 2nd sibiling age)

Contents (Book Chapter 3.6, 4.2)

- Review of survey sampling
- Correlation
  - z-score
  - Correlation coefficient
- Linear regression
  - Residuals
  - Least squares
  - Example: Facial Competence and Vote Share

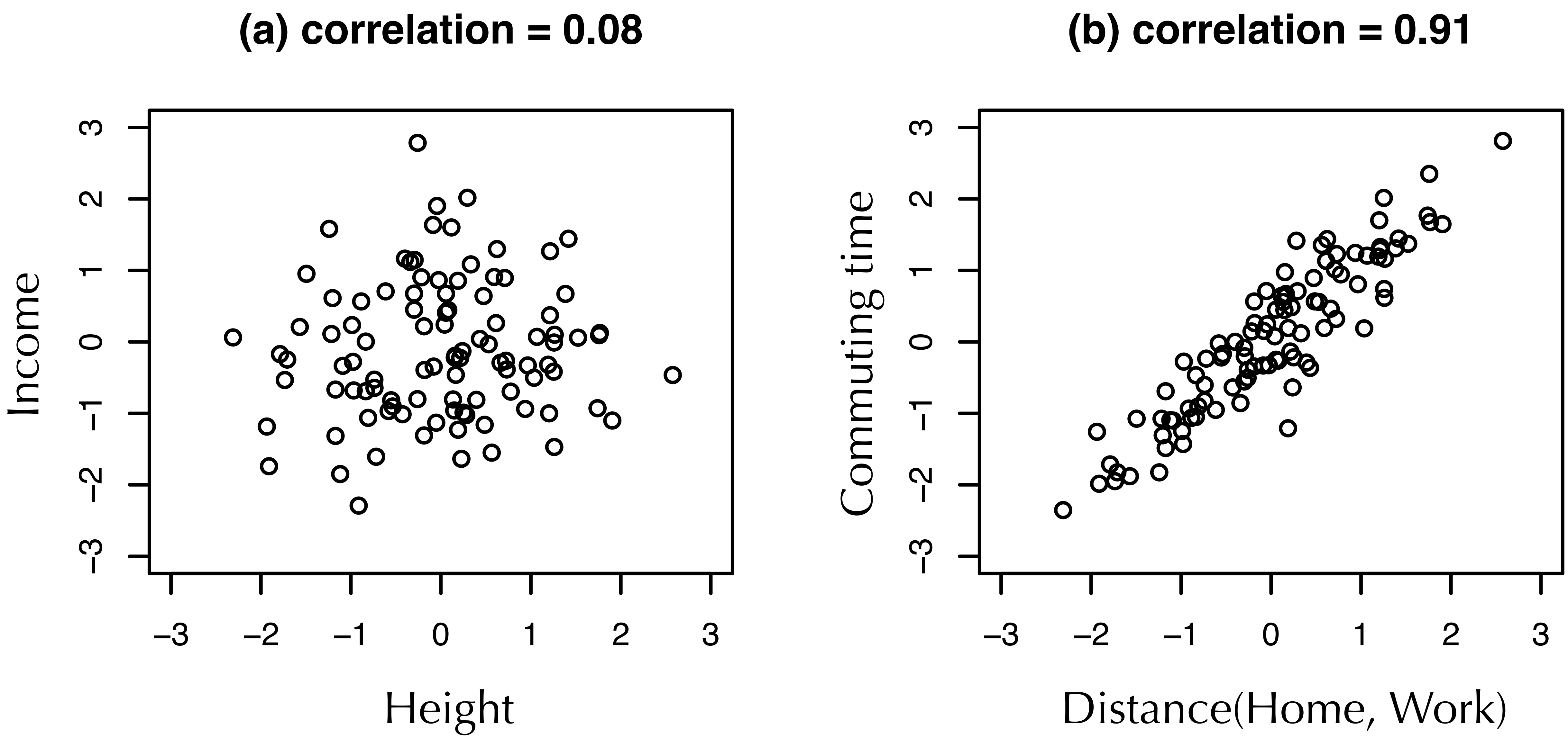
Correlation: Bivariate Relationships

- Few examples



Linear Regression Model

- What if x & y values are variables of interest?



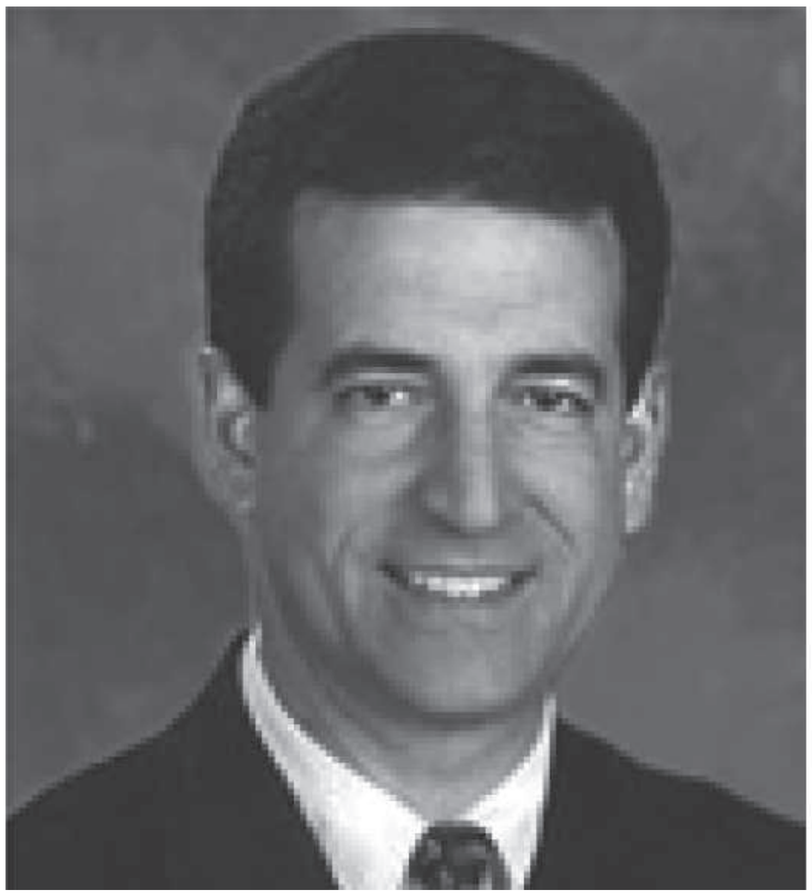
- Challenge: How would you quantify the level of association?

Linear Regression Model

- Model: $$Y = \underbrace{\alpha}_{\text{intercept}} + \underbrace{\beta}_{\text{slope}} X + \underbrace{\epsilon}_{\text{error term}}$$
  - Y: dependent/outcome/response variable
  - X: independent/explanatory variable, predictor
  - α, β: coefficients (parameters of the model)
  - ε: unobserved error/disturbance term (mean zero)
- Interpretation
  - $\alpha+\beta X$: mean of Y given the value of X
  - α: the value of Y when X is zero
  - β: increase in Y associated with one unit increase in X

Example: Facial Appearance and Election Outcomes

- Can you predict election outcome just from their faces?

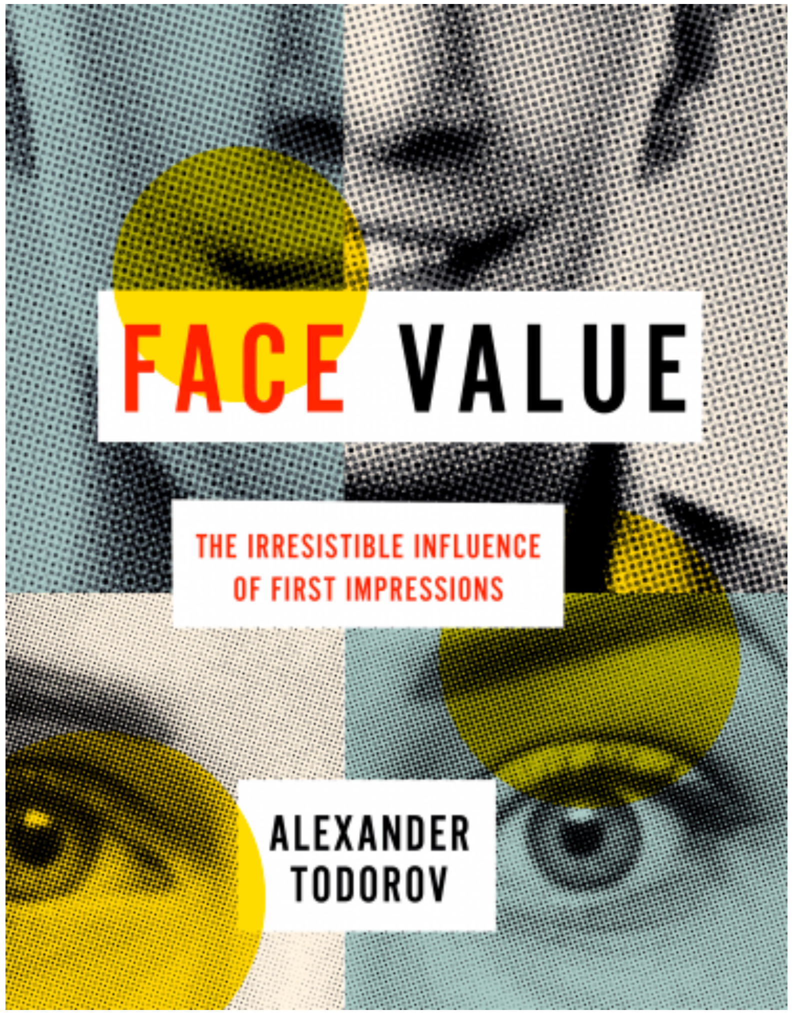


Which person is the most competent?

  - 2004 Wisconsin Senate election
  - Russ Feingold (D) 55% vs. Tim Micheles (R) 44%

Example: Facial Appearance and Election Outcomes

- Experimental design
  - Ask subjects choose one based on perceived level of competence
  - 1s exposure condition had similar results
  - Similar to the idea of crowd-sourcing



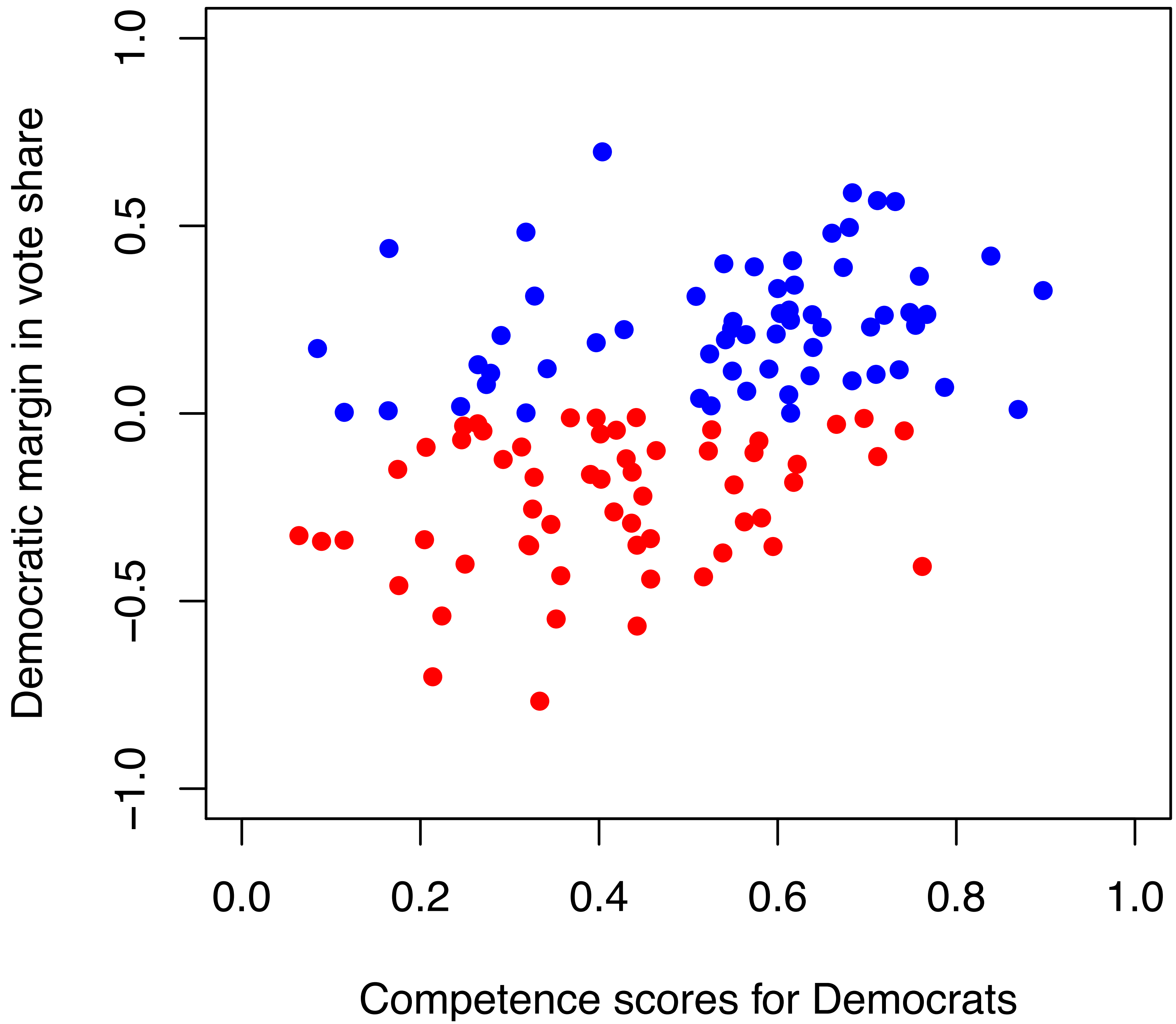
Example: Facial Appearance and Election Outcomes

- Variables in R dataset

Name	Description
`congress`	session of congress
`year`	year of election
`state`	state of election
`winner`	name of winner
`loser`	name of runner-up
`w.party`	party of winner
`l.party`	party of loser
`d.votes`	number of votes for Democratic candidate
`r.votes`	number of votes for Republican candidate
`d.comp`	competence measure for Democratic candidate
`r.comp`	competence measure for Republican candidate

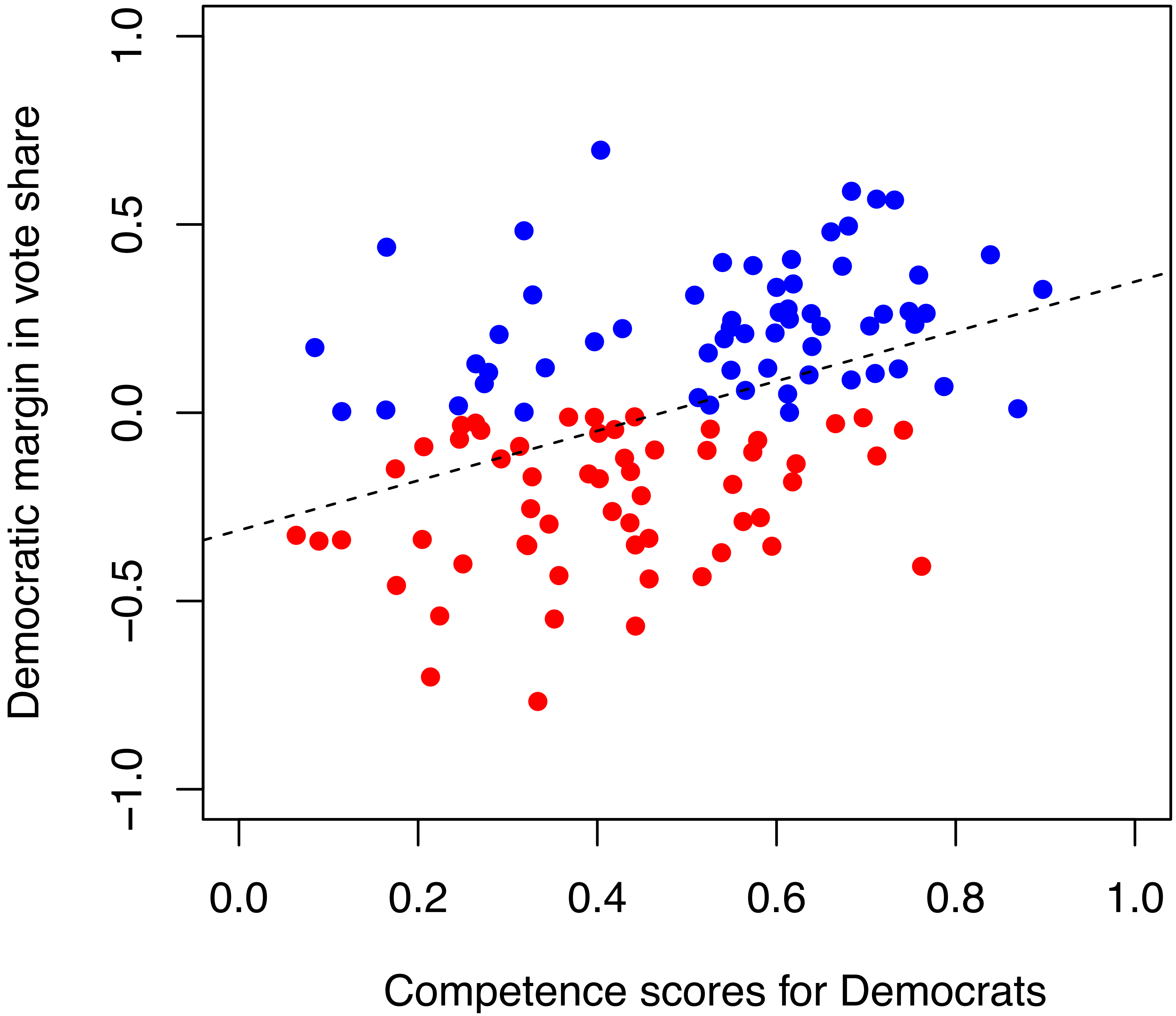
Example: Facial Appearance and Election Outcomes

- Experimental result



Example: Facial Appearance and Election Outcomes

- Linear regression draws a single line w. the highest level of explanation



Example: Facial Appearance and Election Outcomes

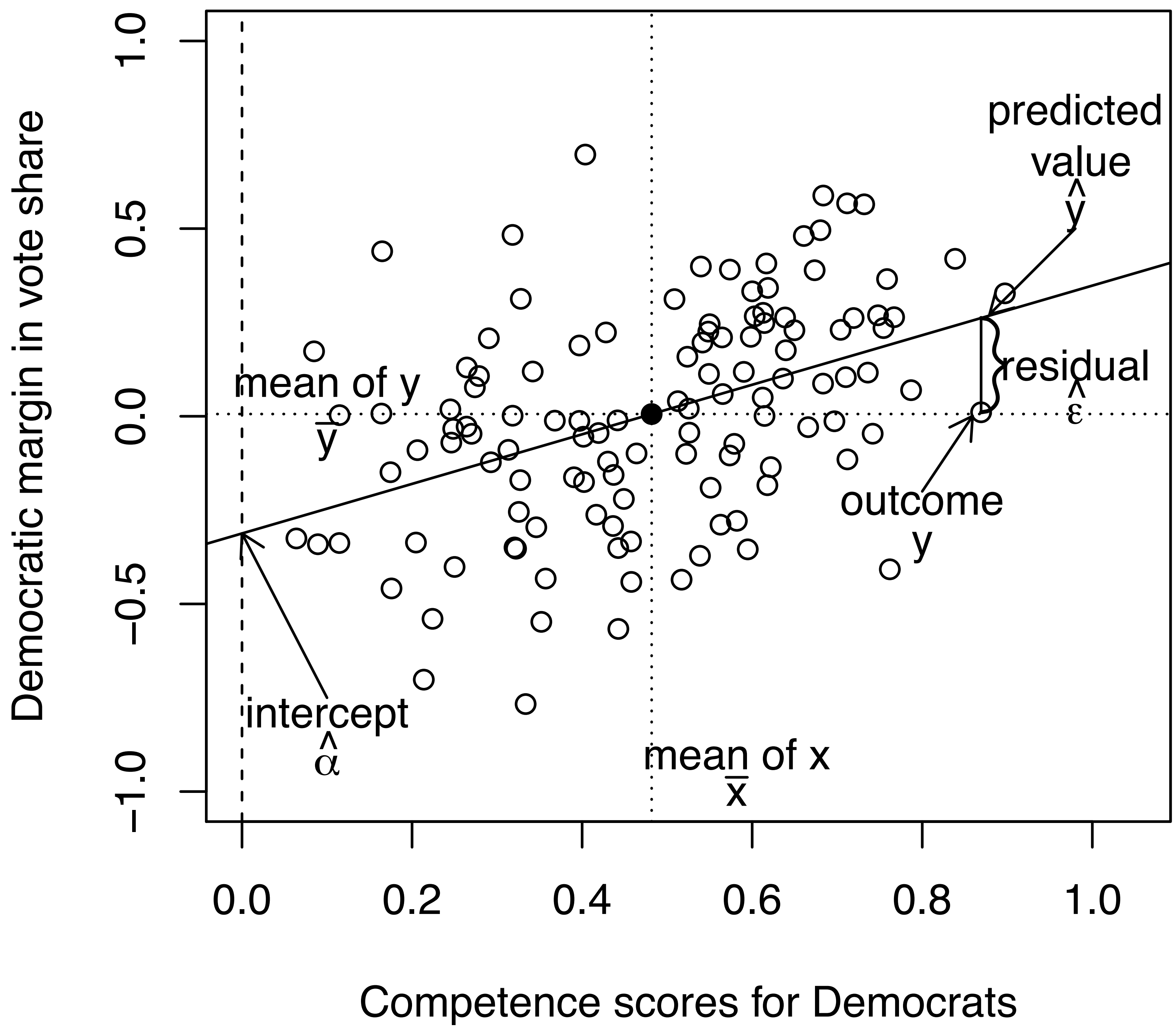
- Linear regression draws a single line w. the highest level of explanation

$$Y = \underbrace{\alpha}_{\text{intercept}} + \underbrace{\beta}_{\text{slope}} X + \underbrace{\epsilon}_{\text{error term}}$$

- Estimation (hat notation for estimated)
  - $\hat{\alpha}, \hat{\beta}$: estimated coefficients
  - $\hat{Y} = \hat{\alpha} + \hat{\beta}X$: predicted/fitted value
  - $\hat{\epsilon} = Y - \hat{Y}$: residuals

Example: Facial Appearance and Election Outcomes

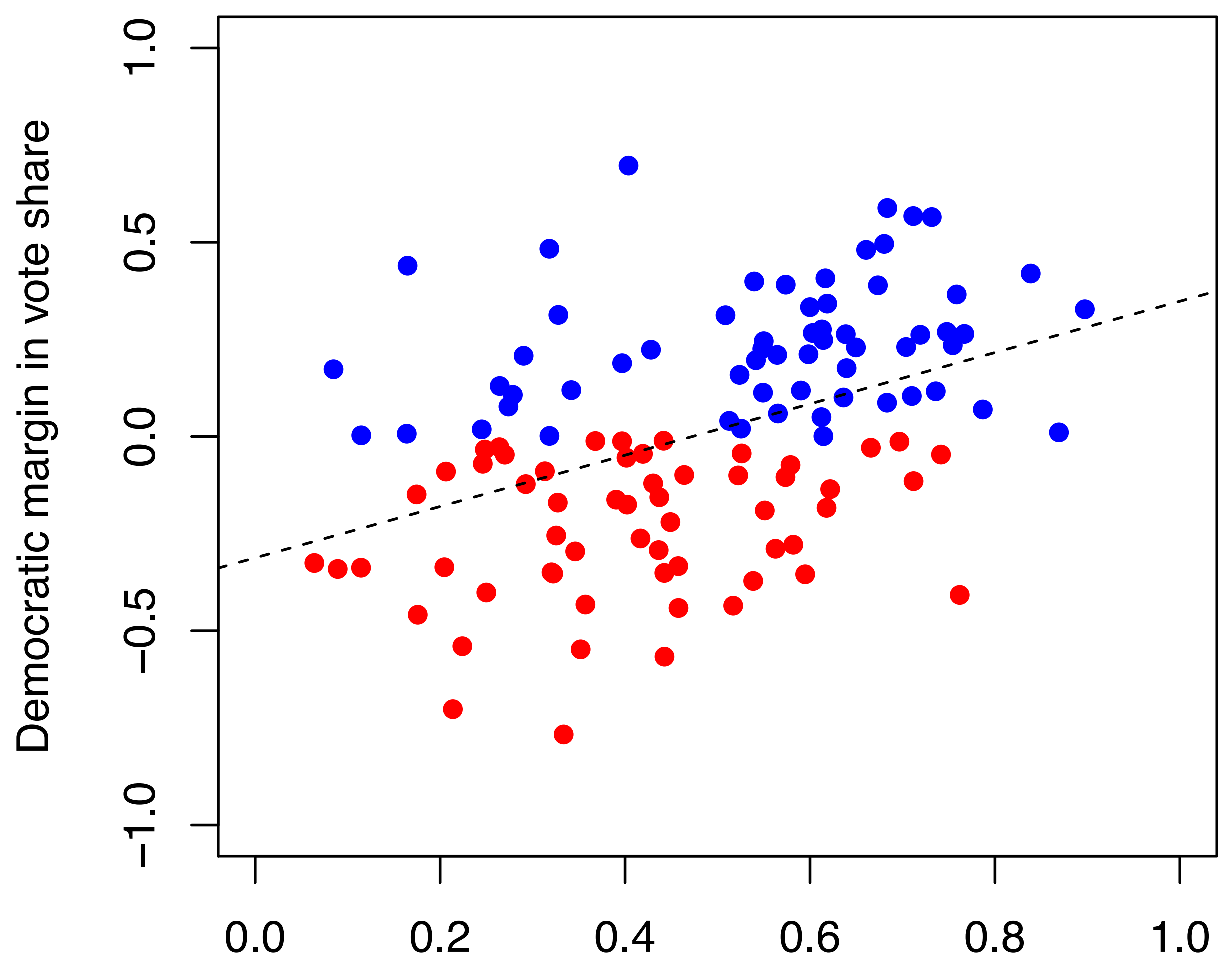
$$Y = \underbrace{\alpha}_{\text{intercept}} + \underbrace{\beta}_{\text{slope}} X + \underbrace{\epsilon}_{\text{error term}}$$



Example: Facial Appearance and Election Outcomes

- Linear regression draws a single line w. the highest level of explanation
  - = Estimating the coefficients (via least squares)
    - How?: Minimize the sum of squared residuals (SSR)

$$\text{SSR} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$



Example: Facial Appearance and Election Outcomes

- Estimated coefficients (will be covered next week)

$$
\begin{aligned}
\hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{X} \\
\hat{\beta} &= \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\
\hat{\beta} &= \text{correlation of } X \text{ and } Y \times \frac{\text{standard deviation of } Y}{\text{standard deviation of } X}
\end{aligned}
$$

- Least squares line goes through $(\bar{X}, \bar{Y})$

$$
\widehat{Y} = (\overline{Y} - \hat{\beta}\overline{X}) + \hat{\beta}\overline{X} = \overline{Y}
$$

- Mean of residuals is always zero:

$$
\text{mean of } \hat{\epsilon} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i) = \overline{Y} - \hat{\alpha} - \hat{\beta}\overline{X} = 0
$$

Example: Facial Appearance and Election Outcomes

- Fit the model

```
fit <- lm(diff.share ~ d.comp, data = face)
fit

##
## Call:
## lm(formula = diff.share ~ d.comp, data = face)
##
## Coefficients:
## (Intercept)       d.comp
##      -0.312        0.660
```

- Estimated coefficients

```
coef(fit)

## (Intercept)      d.comp
##      -0.312       0.660
```

Summary

- Measures
  - z-score
    - Standardized deviation from the mean
  - Correlation coefficient
    - How much z-scores of two variables are correlated
- Linear regression: predicting a linear trend
  - Residuals
  - Least Squares
  - Example: Facial Competence and Vote Share

See you next week.