Bush 631-600: Quantitative Methods

Lecture 5 (09.27.2022): Measurement II & Prediction Intro

Rotem Dvir

The Bush school of Government and Public Policy

Texas A&M University

Fall 2022

What is today's plan?

- In-class: *my first plot*..:))
- More on measurement.
- Latent concepts.
- Visuals: scatterplots.
- Correlation.
- Predictions: why? how?
- Predict with data: elections, defense spending
- R work: scatterplot, subset(), loops, if{}, if{}else{}

Working with R Markdown - Class Task

Data (BAAD v.2): 140 insurgent groups (1998-2012).

- Create **barplot**: religious groups
  - Base R: prop.table() vector and then plot
  - Tidyverse: only x var in aes()
- Create **histogram**: number of civilian casualties
  - Base R: define data and variable to plot ($)
  - Tidyverse: add geom_histogram()

Measurement

Why?

- Social science: develop and test causal theories.
- Leader background and conflict behavior.
- Minimum wage and levels of full-time employment?
- Concepts: level of unemployment, leader background, public approval.

How?

Measures - the context of theoretical concepts

Complex measurement

Latent concepts:

- Hard to measure.
- Variation in definitions.
- Democracy: the polity debate.
- Ideology: representative votes?

A new suspect:

- Terrorism: which violent events are terrorism?

What is terrorism?

Researchers $\rightarrow$ objective measures:

- Identity: perpetrators and victims.
- Population-wide psychological effects.
- Clear political objective.

The Public?

You tell me

Public views of terrorism?

Huff and Kertzer (2018):

- Objective: ‘facts on the ground’.
- Subjective: ‘who and why?’

The Method: Conjoint experiment

- No control group.
- Multiple treatments.
- Outcome: is it terrorism? (yes/no)
- How each factor contributes to viewing an incident as terrorism?

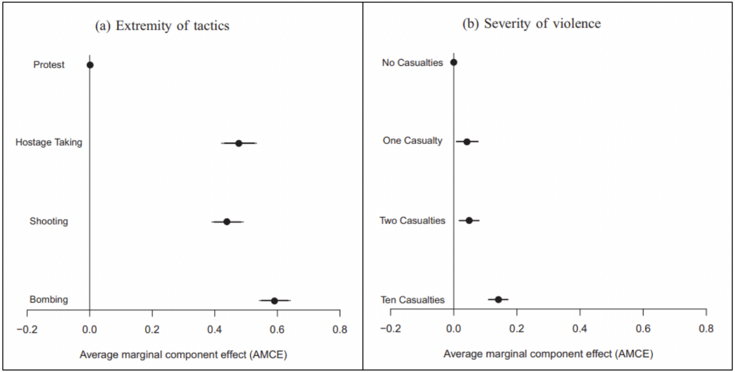
Conjoint experiment: Terrorism

Scenario 1
The incident: shooting
The incident occurred in a church in a foreign democracy with a history of human rights violation
Two individuals died.
The shooting was carried by a Muslim individual with history of mental illness.
News suggest the individual had ongoing personal dispute with one of the targets

Scenario 2
The incident: bombing
The incident occurred in a police station in a foreign dictatorship.
No fatalities reported.
The bombing was carried by a Muslim organization.

News suggest the group was motivated by the goal of overthrowing the government.

Objective path: results



Subjective path: results

FIGURE 5 Social Categorization Effects

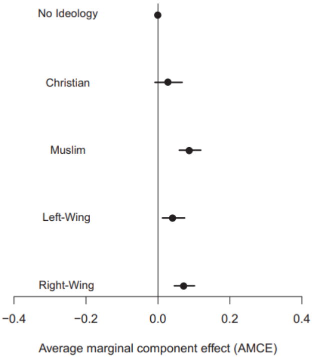
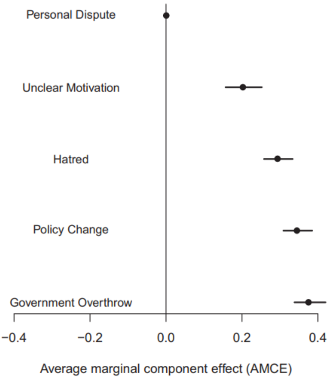


FIGURE 6 Motive Attribution Effects



Terrorism data

Type: event data

A lot of resources:

- GTD - START (Maryland).
- Individuals radicalization (PRIUS) - START (Maryland).
- Episodes of political violence (1946-2017) (Vienna, Austria).
- Suicide terrorism - CPOST (Chicago)
- List (Link)

Terrorism data

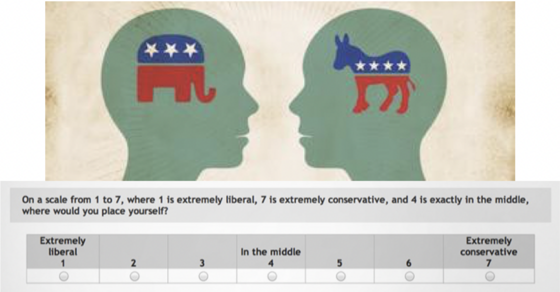
Global Terrorism Database (GTD):

- Time frame: 1970-2019.
- Events: International & domestic terrorism.
- Scope: over 100,000 cases.
- Sources: open source media.

Problem(s)?

- Events data $\rightarrow$ news sources.
- Temporal: less work prior to 1970.
- Biased and Selective reporting: strategic, sensational events.
- Errors in measurement.
- Measures matter - democracy and frequency of incidents (polity, strategic reporting).

Measuring ideology



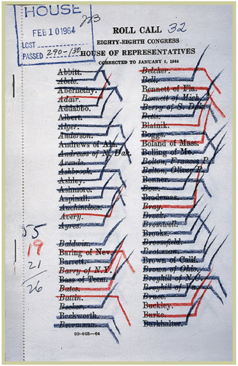
Measurement models:

- Summarize data.
- Learn about human behavior.

Measuring ideology

Legislators measurement model: congress roll-call votes

Voting $\rightarrow$ political orientation.



Complex concepts & measurement

What's the bottom-line?

- Latent concepts: democracy, ideology, terrorism.
- Tricky measurement: conjoint experiment, measurement models.

How to improve measures?

- Theoretical grounding.
- Replications.

Bivariate Relationships

Summarize relationship b-w 2 variables

Liberal-conservative ideology: Economy & Race

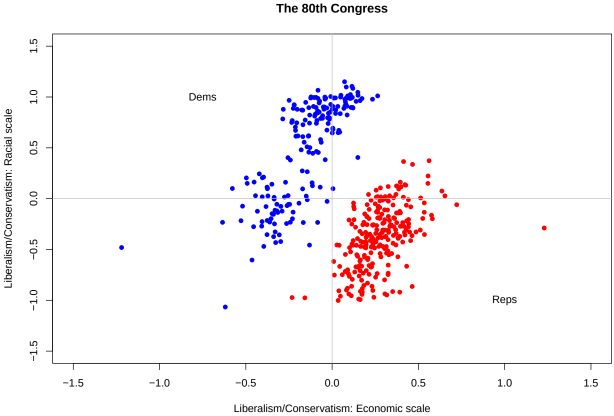
```
head(congress)
```

```
##   congress district   state    party       name dwnom1 dwnom2
## 1       80        0     USA Democrat     TRUMAN -0.276  0.016
## 2       80        1 ALABAMA Democrat  BOYKIN  F. -0.026  0.796
## 3       80        2 ALABAMA Democrat   GRANT  G. -0.042  0.999
## 4       80        3 ALABAMA Democrat ANDREWS  G. -0.008  1.005
## 5       80        4 ALABAMA Democrat   HOBBS  S. -0.082  1.066
## 6       80        5 ALABAMA Democrat   RAINS  A. -0.170  0.870
```

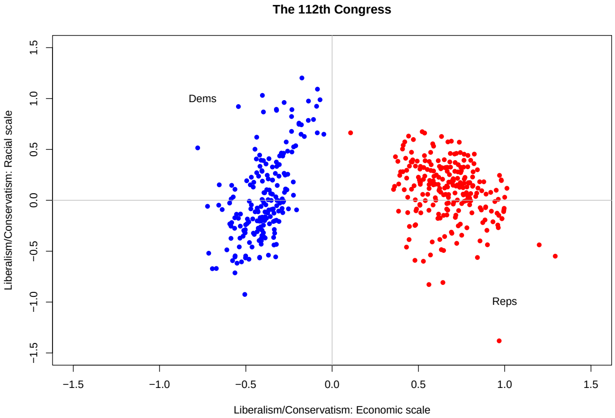
Back to visuals

Scatter plot

- Visualize relationship between 2 variables.
- Numeric/continuous values.



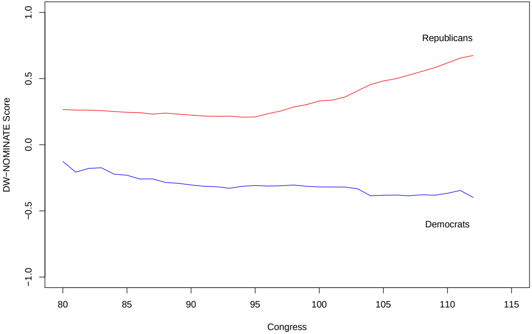
Congress ideology in the 21st century



Congress ideology: time trend

```
dem.med <- tapply(dem$dwnom1, dem$congress, median)
rep.med <- tapply(rep$dwnom1, rep$congress, median)

plot(names(dem.med), dem.med, col = "blue", type = "l",
     xlim = c(80,115), ylim = c(-1,1), xlab = "Congress",
     ylab = "DW-NOMINATE Score")
lines(names(rep.med), rep.med, col = "red")
text(110, -0.6, "Democrats")
text(110,0.8, "Republicans")
```



'International' Ideology

UN $\rightarrow$ International institution.

Voting patterns $\rightarrow$ countries orientation/ideology.

UN voting data (1946-2012)

```
dim(mydata)
```

```
## [1] 9120    6
```

```
summary(mydata)
```

```
##       Year       CountryAbb         CountryName          idealpoint     
##  Min.   :1946   Length:9120        Length:9120        Min.   :-2.6552  
##  1st Qu.:1972   Class :character   Class :character   1st Qu.:-0.6406  
##  Median :1987   Mode  :character   Mode  :character   Median :-0.1644  
##  Mean   :1985                                         Mean   : 0.0000  
##  3rd Qu.:2001                                         3rd Qu.: 0.7968  
##  Max.   :2012                                         Max.   : 3.0144  
##                                                                        
##   PctAgreeUS     PctAgreeRUSSIA  
##  Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.1395   1st Qu.:0.5053  
##  Median :0.2400   Median :0.6567  
##  Mean   :0.2960   Mean   :0.6219  
##  3rd Qu.:0.3902   3rd Qu.:0.7424  
##  Max.   :1.0000   Max.   :1.0000  
##  NA's   :1        NA's   :5       
```

Global ideologies

Voting with US → measure of foreign policy similarity.

Similar FP → similar global orientation.

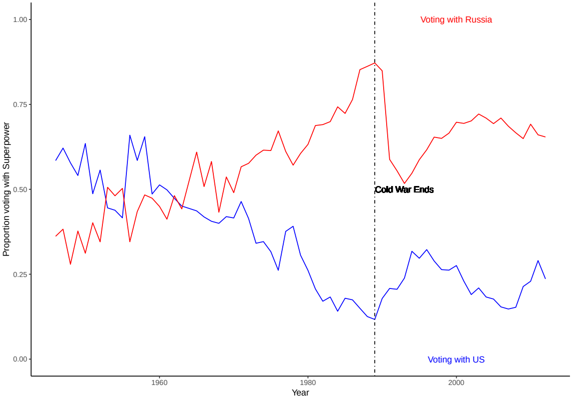
```
# Tidyverse approach to data management
# Arrange by year, calculate mean for US / Russia voting
annual.agree <- mydata %>%
  group_by(Year) %>%
  summarize(us.agree = mean(PctAgreeUS, na.rm = T),
            ru.agree = mean(PctAgreeRUSSIA, na.rm = T))

head(annual.agree)
```

```
## # A tibble: 6 x 3
##    Year us.agree ru.agree
##   <int>    <dbl>    <dbl>
## 1  1946    0.585    0.362
## 2  1947    0.621    0.383
## 3  1948    0.578    0.279
## 4  1949    0.541    0.377
## 5  1950    0.635    0.312
## 6  1951    0.487    0.402
```

Trends in global ideology

```
ggplot(data = annual.agree) +
  geom_line(mapping = aes(x = Year, y = us.agree), color = "blue") +
  geom_line(mapping = aes(x = Year, y = ru.agree), color = "red") +
  geom_text(aes(x = 2000, y = 0, label = "Voting with US"), color = "blue", data = data.frame()) +
  geom_text(aes(x = 2000, y = 1, label = "Voting with Russia"), color = "red", data = data.frame()) +
  geom_vline(aes(xintercept = 1989), linetype = "dotdash", color = "black") +
  geom_text(aes(x = 1993, y = 0.5, label = "Cold War Ends"), color = "black") +
  ylab("Proportion voting with Superpower") + theme_classic()
```



Grouping observations

Which side are you on?

Grouping countries: FP Similarity measures

```r
# Table for voting close to US
# USA
mydata %>%
  group_by(CountryName) %>%
  summarise(mean.pctUS = mean(PctAgreeUS)) %>%
  arrange(desc(mean.pctUS)) %>%
  head(n = 11) %>%
  filter(CountryName != "United States of America")
```

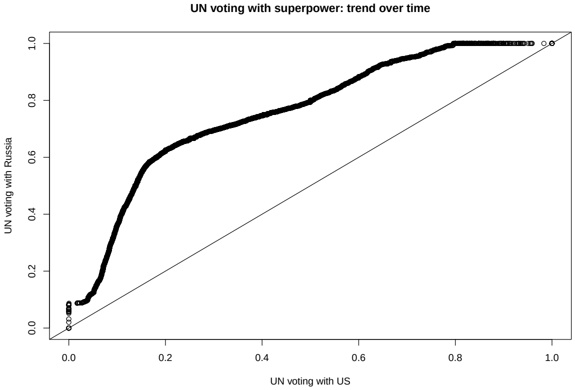
```
## # A tibble: 10 x 2
##    CountryName                    mean.pctUS
##    <chr>                               <dbl>
##  1 Palau                               0.736
##  2 United Kingdom                      0.652
##  3 Taiwan                              0.643
##  4 Israel                              0.640
##  5 Federated States of Micronesia      0.594
##  6 Canada                              0.586
##  7 Luxembourg                          0.571
##  8 Netherlands                         0.562
##  9 Belgium                             0.562
## 10 France                              0.549
```

Visualizing distributions

Qunatile Qunatile Plot

Scatter-plot of quantiles

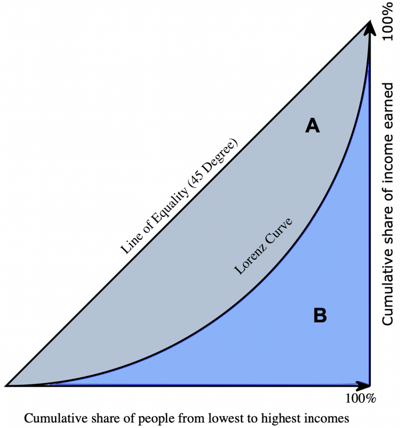
```
### Q-Q plot
qqplot(mydata$PctAgreeUS, mydata$PctAgreeRUSSIA, xlab = "UN voting with US",
       ylab = "UN voting with Russia",
       main = "UN voting with superpower: trend over time")
abline(0,1)
```



Political polarization: QSS textbook

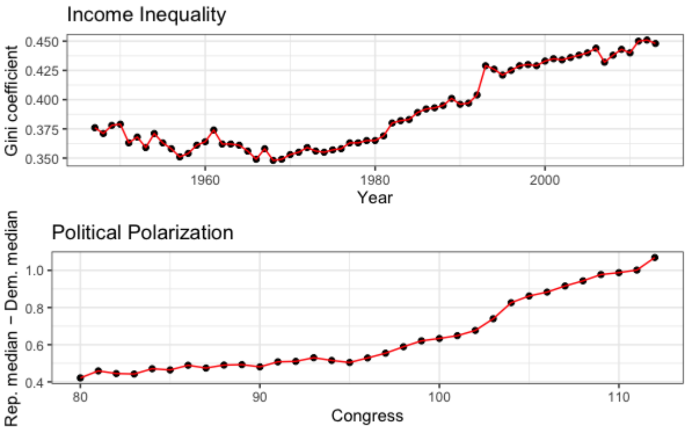
Income inequality → political polarization.

The *Gini coefficient*



US test case

Gini coefficient - Political Polarization

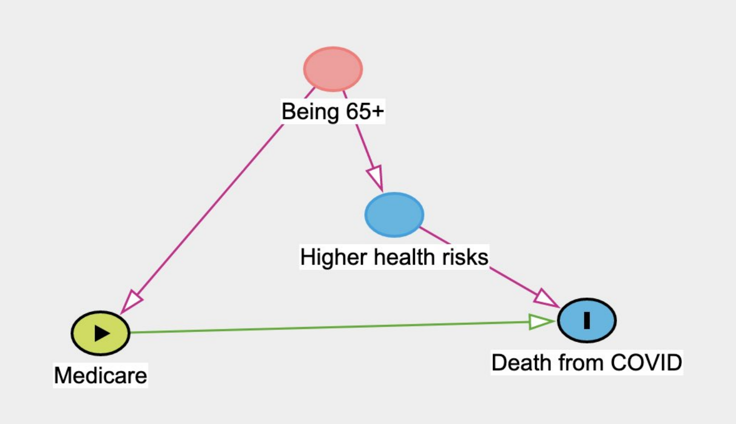


Association b-w variables: Correlation

Income inequality $\rightarrow$ Political polarization?

Correlation does not mean causation

Correlation & causality



Association b-w variables

Correlation:

- Summary of bivariate relationship.
- How two factors 'move together' on average.
- Always relative to mean value.

Product of z-scores:

$$cor(x, y) = \frac{1}{n} \sum_{i=1}^{n} (Z - x_i * Z - y_i)$$

Z-scores

- A measure for the deviation from the mean (in SD terms)
- Standardize variable
- Allows comparison with *common units*

$$Zscore(X_i) = \frac{x_i - \bar{x}}{SD(X_i)}$$

Z score $> 0 \rightarrow$ unit larger than mean
Z score $< 0 \rightarrow$ unit smaller than mean

z-score example: Test scores

Where do we stand versus our cohort?

- Total of 500 students
- Mean grade ($\bar{X} = 85$)
- SD ($\sigma = 6$)

```r
# Our grades = 81, 90, 65
z1 <- (81-85)/6
z1
```

```
## [1] -0.6666667
```

```r
z2 <- (90-85)/6
z2
```

```
## [1] 0.8333333
```

```r
z3 <- (65-85)/6
z3
```

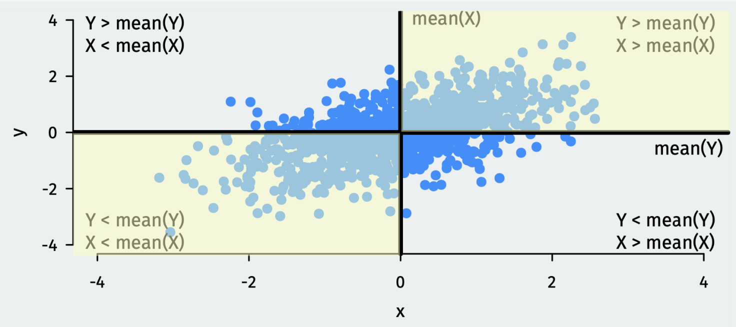
```
## [1] -3.333333
```

Correlation

- Average product of z-scores:
  - Positive correlation: when x is bigger than its mean, so is y
  - Negative correlation: when x is bigger than its mean, y is smaller
- z-score: not sensitive to unit used
- Correlation is identical even for different measuring units of variable

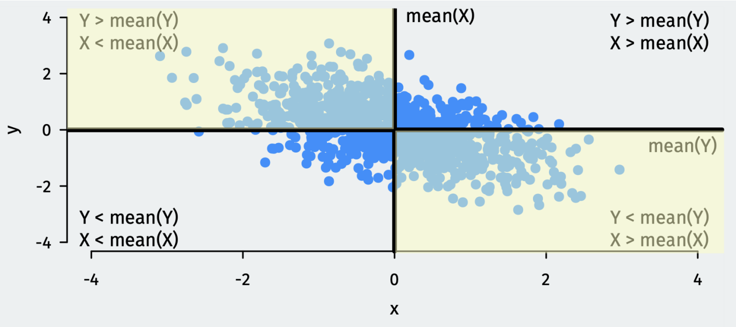
Correlation - how do the data look?

Positive correlation



Correlation - how do the data look?

Negative correlation



Correlation

- Measures **linear** association
- Order does not matter: cor(x,y) = cor(y,x)
- Interpretation:
  - Values range between (-1) to 1.
  - Close to ‘edges’ $\rightarrow$ stronger association.
  - Value of zero $\rightarrow$ no association.
  - Positive correlation $\rightarrow$ positive association.
  - Negative correlation $\rightarrow$ negative association.

Correlation in R

UN Voting: association b-w ideal point & liberal FP approach

```
# Voting with US
cor(mydata$idealpoint, mydata$PctAgreeUS, use = "pairwise")
```

```
## [1] 0.7498446
```

```
# Voting with Russia
cor(mydata$idealpoint, mydata$PctAgreeRUSSIA, use = "pairwise")
```

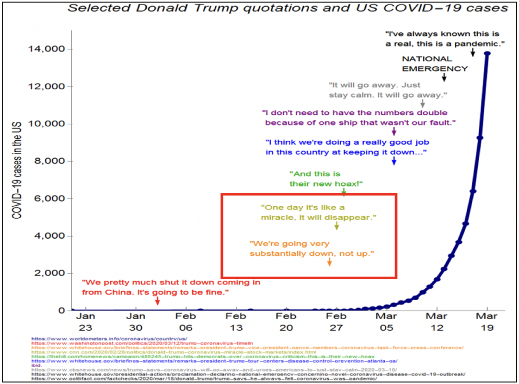
```
## [1] -0.7050107
```

Predicting with data

- Social science research:
  - Establish causality.
  - The role of measurement.
- Predictions:
  - Support for causal statements.
  - Generate accurate predictions about potential outcomes.

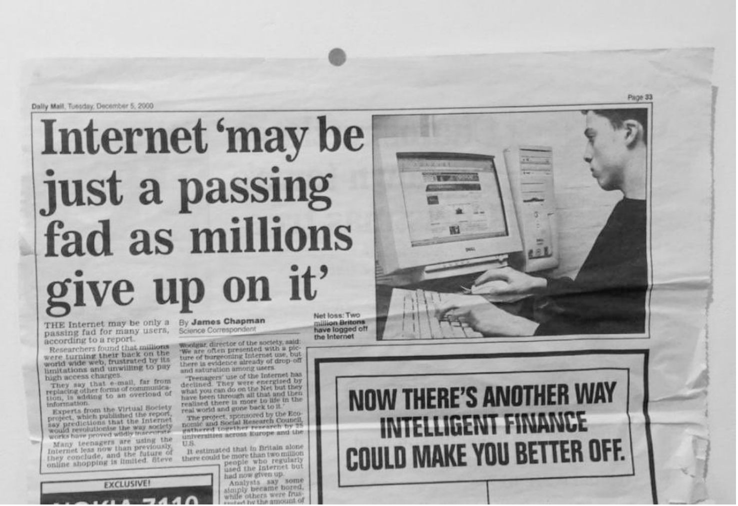
Not the best. . . predictions!

Oh no. . .



Some more gems

Daily Mail - December 5, 2000



Daily Mail, Tuesday, December 5, 2000

Page 33

Internet 'may be just a passing fad as millions give up on it'

By **James Chapman**
Science Correspondent

THE Internet may be only a passing fad for many users, according to a report.

Researchers found that millions were turning their back on the world wide web, frustrated by its limitations and unwilling to pay high access charges.

They say that e-mail, far from replacing other forms of communication, is adding to an overload of information.

Experts from the Virtual Society project, which published the report, say predictions that the Internet would revolutionise the way society works have proved wildly inaccurate.

Many teenagers are using the Internet less now than previously, they conclude, and the future of online shopping is limited. Steve Woolgar, director of the society, said: 'We are often presented with a picture of burgeoning Internet use, but there is evidence already of drop-off and saturation among users.

'Teenagers' use of the Internet has declined. They were energised by what you can do on the Net but they have been through all that and then realised there is more to life in the real world and gone back to it.'

The project, sponsored by the Economic and Social Research Council, gathered together research by 25 universities across Europe and the U.S.

It estimated that in Britain alone there could be more than two million people who regularly used the Internet but had now given up.

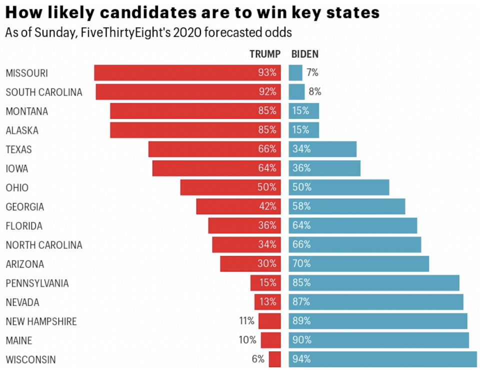
Analysts say some simply became bored, while others were frustrated by the amount of

Net loss: Two million Britons have logged off the Internet

Some groundwork

Loops

- Useful to repeat the same operation multiple times.
- Efficient analysis tool.



Loops in R

- Run similar code chunk repeatedly.

```r
for (i in X) {
  expression1
  expression2
  ...
  expression3
}
```

- Elements of loop:
  - i: counter (change as you like).
  - X: Vector of ordered values for the counter.
  - expression: set of expressions to run repeatedly.
  - {}: curly braces define the beginning and end of a loop.

Loops in R

```
weeks <- c(1,2,3,4,5)
n <- length(weeks)
t <- rep(NA,n)

# loop counter
for (i in 1:n){
  t[i] <- weeks[i] * 2
  cat("I completed Swirl HW number", weeks[i], "in",
      t[i], "minutes", "\n")
}
```

```
## I completed Swirl HW number 1 in 2 minutes
## I completed Swirl HW number 2 in 4 minutes
## I completed Swirl HW number 3 in 6 minutes
## I completed Swirl HW number 4 in 8 minutes
## I completed Swirl HW number 5 in 10 minutes
```

Conditional statements

Implement code chunks based on logical expressions.

If statements

Syntax: if(x = a condition){set of commands}

Run command(s) only if value if X is TRUE

```
weather <- "rain"
if (weather == "rain"){
  cat("I should take my umbrella")
}
```

```
## I should take my umbrella
```

Flexible if statements

Using if(){} else {}

```
weather <- "sunny"
if (weather == "rain"){
  cat("I should take my umbrella")
} else {
  cat("I should wear my Aggie hat")
}
```

```
## I should wear my Aggie hat
```

Complex conditional statements

Join conditional statements into a loop.

```r
days <- 1:7
n <- length(days)

for (i in 1:n){
  x <- days[i]
  r <- x %% 2

  if (r == 0){
    cat("Day", x, "is even and I need my umbrella \n")
  } else {
    cat("Day", x, "is odd and I need my Aggie cap \n")
  }
}
```

```
## Day 1 is odd and I need my Aggie cap
## Day 2 is even and I need my umbrella
## Day 3 is odd and I need my Aggie cap
## Day 4 is even and I need my umbrella
## Day 5 is odd and I need my Aggie cap
## Day 6 is even and I need my umbrella
## Day 7 is odd and I need my Aggie cap
```

Conditional statements

Nesting multiple conditional statements $\rightarrow$ MyApp Link

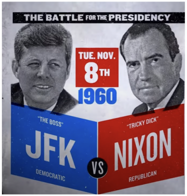
Caution:

- if(){} else{} are complex.
- Double check the curly braces for each statement.
- Use the automatic indentation.
- ‘Space-out’ your code.
- Add comments (using #) to clearly mark each step.

Predictions

- Awesome research tool...with the right design.
- Predict: elections, economic trends, behavior, Superbowl winners, etc.

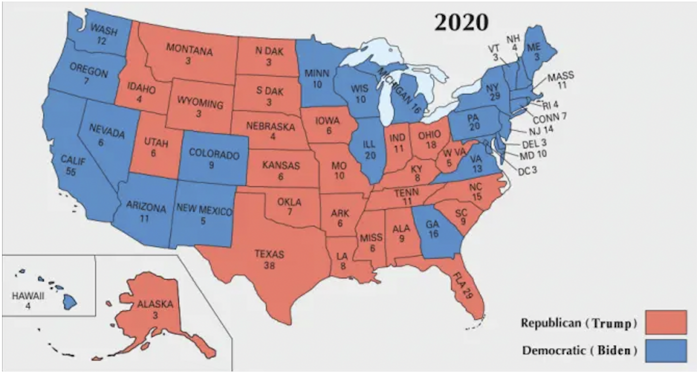
Elections winner



US electoral system

Electoral college

Plurality of votes in a state: “Winner-take-all”



Election predictions

Measurement problem:

- National vote vs. electoral votes.
- Bush - Gore (2000).
- Clinton - Trump (2016).

Electoral vote:

- Number of electors does not align with number of voters per state.
- Votes are “unaccounted”.

A Prediction problem:

- Accurate forecast of **each state** winner.

Polls and election predictions

Data: 2016 elections (polls)

```
head(polls16)
```

```
##   state  middate daysleft                  pollster
## 1    AK  8/11/16       89  Lake Research Partners
## 2    AK  8/20/16       80              SurveyMonkey
## 3    AK 10/20/16       19                    YouGov
## 4    AK 10/26/16       13 Google Consumer Surveys
## 5    AK  9/30/16       39 Google Consumer Surveys
## 6    AK 10/12/16       27 Google Consumer Surveys
##   clinton trump margin
## 1    30.0  38.0   8.00
## 2    31.0  38.0   7.00
## 3    37.4  37.7   0.30
## 4    38.0  39.0   1.00
## 5    47.5  36.7 -10.76
## 6    34.6  30.0  -4.62
```

Poll prediction by states (using R loop)

```
poll.pred <- rep(NA, 51) # place holder

# get list of unique state names to iterate over
st.names <- unique(polls16$state)

# add labels to holder
names(poll.pred) <- st.names

for (i in 1:51) {
  state.data <- subset(polls16, subset = (state == st.names[i]))

  latest <- state.data$daysleft == min(state.data$daysleft)

  poll.pred[i] <- mean(state.data$margin[latest])
}

head(poll.pred)
```

```
##     AK    AL    AR    AZ     CA    CO
##  14.73 29.72 20.02  2.50 -23.00 -7.05
```

Errors in polling

Prediction error = actual outcome - predicted outcome

```r
errors <- pres16$margin - poll.pred
names(errors) <- st.names
mean(errors)
```

```
## [1] 3.81
```

Root mean-square-error (RMSE): average magnitude of prediction error

```r
sqrt(mean(errors^2))
```

```
## [1] 9.6
```

Prediction challenges

Prediction of binary outcome variable $\rightarrow$ classification problem

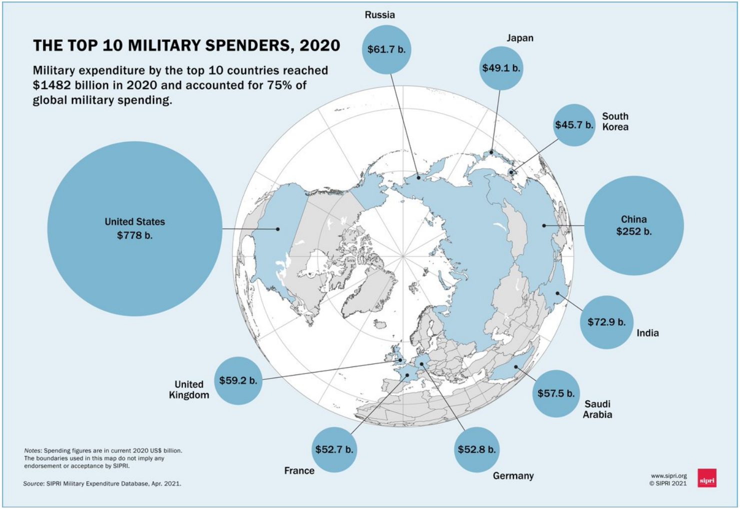
Wrong prediction $\rightarrow$ misclassification:

1. true positive: predict Trump wins when he actually wins.
2. **false positive**: predict Trump wins when he actually loses.
3. true negative: predict Trump loses when he actually loses.
4. **false negative**: predict Trump loses when he actually wins.

2016 elections: misclassification rate was high: 9.8% (5/51 states).

Predictions in INTA

Military spending across the globe



Predicting military spending

Our data:

- 157 Countries
- Time frame: 1999-2019
- Measure: military spending as proportion of total gov't spending.

Why this measure?

- Reflect state's preferences.
- Trade-off: *Guns vs. Butter*.

Our predictions:

- Using 1999-2019 data to predict 2020 levels.
- Test predictions with actual data.

Military spending data

```
dim(mil_exp)
```

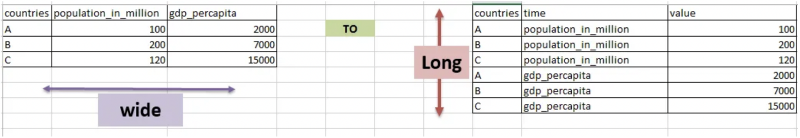
```
## [1] 157  25
```

```
head(mil_exp, n=8)
```

```
## # A tibble: 8 x 25
##   Country Group1 Subgr~1 `1999` `2000` `2001` `2002` `2003` `2004` `2005` `2
##   <chr>   <chr>  <chr>    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <
## 1 Algeria Africa North ~ 0.118  0.120  0.122  0.108  0.101  0.107  0.105  0.
## 2 Libya   Africa North ~ 0.115  0.103  0.0630 0.0524 0.0484 0.0490 0.0502 0.
## 3 Morocco Africa North ~ 0.145  0.0898 0.145  0.125  0.134  0.123  0.105  0.
## 4 Tunisia Africa North ~ 0.0618 0.0614 0.0605 0.0590 0.0603 0.0591 0.0601 0.
## 5 Angola  Africa Sub-Sa~ 0.274  0.129  0.108  0.0919 0.109  0.116  0.139  0.
## 6 Benin   Africa Sub-Sa~ 0.0452 0.0264 0.0232 0.0407 0.0473 0.0506 0.0482 0.
## 7 Botswa~ Africa Sub-Sa~ 0.0759 0.0817 0.0899 0.0900 0.0915 0.0848 0.0823 0.
## 8 Burkin~ Africa Sub-Sa~ 0.0576 0.0624 0.0588 0.0605 0.0610 0.0596 0.0594 0.
## # ... with 14 more variables: `2007` <dbl>, `2008` <dbl>, `2009` <dbl>,
## #   `2010` <dbl>, `2011` <dbl>, `2012` <dbl>, `2013` <dbl>, `2014` <dbl>,
## #   `2015` <dbl>, `2016` <dbl>, `2017` <dbl>, `2018` <dbl>, `2019` <dbl>,
## #   `2020` <dbl>, and abbreviated variable name 1: Subgroup1
## # i Use `colnames()` to see all variable names
```

Reshaping the data

- Use the gather() function
- Increase the data size.
- Each case (country for us) has multiple observations (rows).



countries	population_in_million	gdp_percapita
A	100	2000
B	200	7000
C	120	15000

countries	time	value
A	population_in_million	100
B	population_in_million	200
C	population_in_million	120
A	gdp_percapita	2000
B	gdp_percapita	7000
C	gdp_percapita	15000

Reshaping the data

gather() function: long-form data.

```
spend_long <- mil_exp2 %>%
  gather(year, exp, '1999':'2019',-Country, -Group1, -Subgroup1) %>%
  arrange(Country)

head(spend_long, n=9)
```

```
## # A tibble: 9 x 5
##   Country     Group1         Subgroup1  year    exp
##   <chr>       <chr>          <chr>      <chr> <dbl>
## 1 Afghanistan Asia & Oceania South Asia 1999  NA
## 2 Afghanistan Asia & Oceania South Asia 2000  NA
## 3 Afghanistan Asia & Oceania South Asia 2001  NA
## 4 Afghanistan Asia & Oceania South Asia 2002  NA
## 5 Afghanistan Asia & Oceania South Asia 2003  NA
## 6 Afghanistan Asia & Oceania South Asia 2004   0.161
## 7 Afghanistan Asia & Oceania South Asia 2005   0.127
## 8 Afghanistan Asia & Oceania South Asia 2006   0.104
## 9 Afghanistan Asia & Oceania South Asia 2007   0.119
```

Predicting spending

Predict 2020 $\rightarrow$ mean of spending (1999-2019)

Use loop to calculate means for all countries

```
## loop
pred.mean <- rep(NA,157)
c.names <- unique(spend_long$Country)
names(pred.mean) <- as.character(c.names)

for (i in 1:157){
  c.dat <- subset(spend_long, subset = (Country == c.names[i]))
  pred.mean[i] <- mean(c.dat$exp, na.rm = T)
}
```

Predicting spending for 2020

```
pred.mean
         Afghanistan              Albania              Algeria               Angola            Argentina              Armenia
        7.693784e-02         4.803755e-02         1.167886e-01         1.142081e-01         2.865062e-02         1.572688e-01
           Australia              Austria           Azerbaijan              Bahrain           Bangladesh              Belarus
        5.117444e-02         1.621721e-02         1.159260e-01         1.365441e-01         1.024893e-01         3.055717e-01
             Belgium               Belize                Benin              Bolivia   Bosnia-Herzegovina             Botswana
        2.104063e-02         3.481603e-02         4.312747e-02         5.311684e-02         3.023730e-02         7.708387e-02
              Brazil               Brunei             Bulgaria         Burkina Faso              Burundi             Cambodia
        3.954679e-02         8.537055e-02         5.727167e-02         6.086991e-02         1.238733e-01         9.068995e-02
            Cameroon               Canada           Cape Verde Central African Rep.                 Chad                Chile
        7.432152e-02         2.898024e-02         1.845547e-02         1.090412e-01         1.641743e-01         1.010081e-01
               China             Colombia    Congo, Dem. Rep.  Congo, Republic of           Costa Rica        Côte d'Ivoire
        8.147621e-02         1.133810e-01         9.082535e-02         8.326183e-02         0.000000e+00         7.179591e-02
             Croatia               Cyprus              Czechia              Denmark             Djibouti       Dominican Rep.
        4.203798e-02         4.971926e-02         3.230034e-02         2.517054e-02         1.513522e-01         4.516247e-02
             Ecuador                Egypt          El Salvador    Equatorial Guinea              Estonia             eSwatini
        7.900969e-02         6.539493e-02         4.407673e-02         5.624585e-02         4.613709e-02         6.040772e-02
            Ethiopia                 Fiji              Finland               France                Gabon               Gambia
        1.032980e-01         5.669500e-02         2.704904e-02         3.599000e-02         7.089440e-02         3.735918e-02
             Georgia              Germany                Ghana               Greece            Guatemala               Guinea
        1.093521e-01         2.686035e-02         2.040455e-02         5.686649e-02         3.739819e-02         1.172825e-01
       Guinea-Bissau               Guyana                Haiti             Honduras              Hungary              Iceland
        9.553127e-02         4.376836e-02         6.134272e-06         4.366182e-02         2.511546e-02         0.000000e+00
               India            Indonesia                 Iran                 Iraq              Ireland               Israel
        9.692641e-02         4.121770e-02         1.431855e-01         6.366464e-02         1.471538e-02         1.420280e-01
               Italy              Jamaica                Japan               Jordan           Kazakhstan                Kenya
        3.099443e-02         2.671973e-02         2.559871e-02         1.535606e-01         4.722987e-02         6.172174e-02
        Korea, South               Kuwait           Kyrgyzstan                 Laos               Latvia              Lebanon
        1.276501e-01         1.222232e-01         4.838694e-02         2.179216e-02         3.728258e-02         1.416378e-01
             Lesotho              Liberia                Libya            Lithuania           Luxembourg           Madagascar
        4.794950e-02         2.041134e-02         6.558880e-02         3.439832e-02         1.313624e-02         5.316299e-02
              Malawi             Malaysia                 Mali                Malta           Mauritania            Mauritius
        2.908423e-02         6.375313e-02         8.162525e-02         1.457119e-02         1.070985e-01         7.006463e-03
```

Good prediction?

Checking for errors:

```
# Calculate errors & assign country names
errors <- mil_exp$`2020` - pred.mean
names(errors) <- c.names

# Average error
mean(errors, na.rm = T)
```

```
## [1] -0.01210775
```

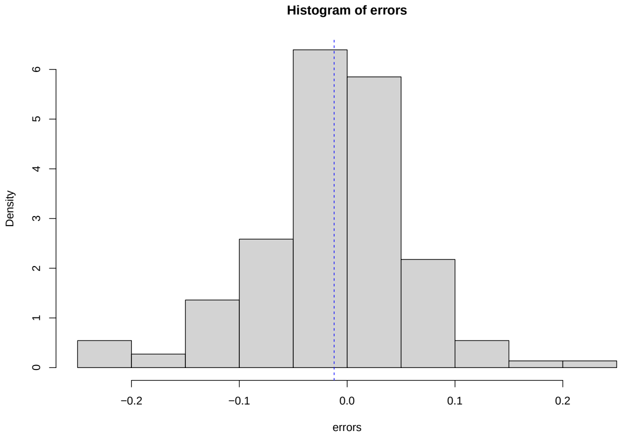
```
# RMSE
sqrt(mean(errors^2, na.rm = T))
```

```
## [1] 0.07380063
```

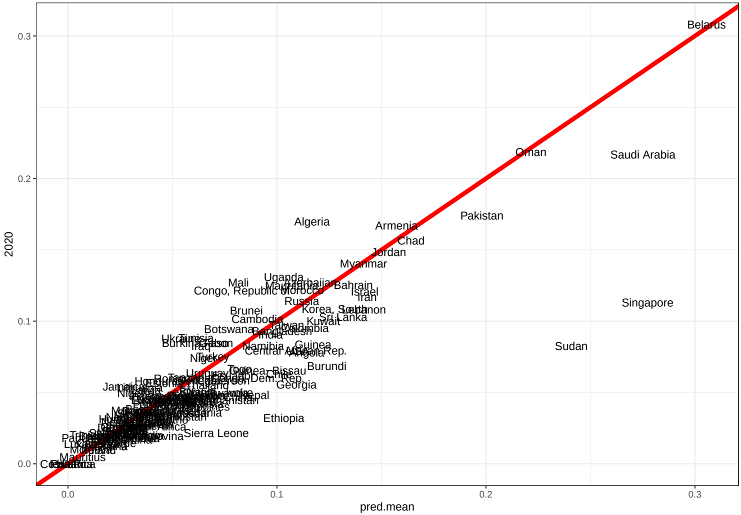
Prediction errors

How far off are we?

```
hist(errors, freq = FALSE)
abline(v = mean(errors, na.rm = T), lty = "dashed", col = "blue")
```



Accuracy of predictions



Find outlier predictions

Identify where we were off...

```
# Errors distribution
summary(n.dat$error)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## -0.164364 -0.017092 -0.004715 -0.008734  0.000374  0.053107        10
```

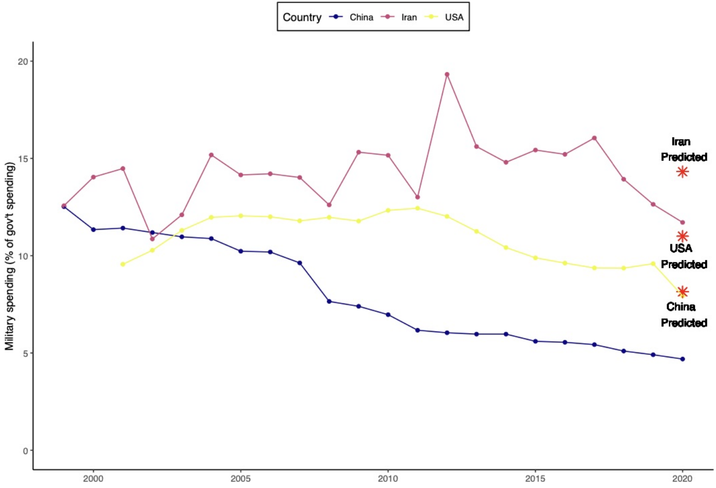
```
# Create variable for large outliers
n.dat$large.inc <- NA
n.dat$large.inc[n.dat$error > 0.01] <- "Much More"
n.dat$large.inc[n.dat$error < -0.01] <- "Much Less"

# Create subset of outliers: less than average
n.dat2 <- n.dat %>%
  filter(large.inc == "Much Less") %>%
  mutate(error = error * 100) %>%
  select(Group1, error) %>% arrange(desc(error))

tail(n.dat2, n=9)
```

```
##                      Group1      error
## Chile              Americas  -3.785553
## Nepal        Asia & Oceania  -4.102959
## Sierra Leone         Africa  -4.945523
## Georgia              Europe  -5.375066
## Burundi              Africa  -5.521676
## Saudi Arabia    Middle East  -5.806989
## Ethiopia             Africa  -7.119952
## Sudan                Africa -15.832405
## Singapore    Asia & Oceania -16.436356
```

Spending over time (and predicted 2020 - the ‘big 3’)



Wrapping up week 5

Summary:

- Measuring complex (latent) concepts: terrorism, ideology.
- Visualize bivariate relations: scatter plot, QQplot.
- z-scores and standardizing units.
- Correlation: how two factors 'move together'.
- Predictions: critical tool, how to? (loops, if/else).
- Predict elections or defense spedning with the average.
- R work: scatterplots, cor(), qqplot(), for loops, if{}else{}.

Task 1: Next Tuesday at midnight!!