Bush 631-603: Quantitative Methods

Lecture 6 (03.01.2022): Prediction vol. II

Rotem Dvir

The Bush school of Government and Public Policy

Texas A&M University

Spring 2022

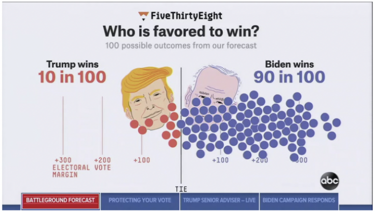
What is today's plan?

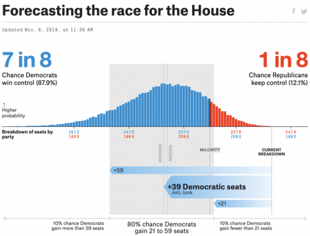
- Predictions: Improved (and more accurate) methods.
- Identify correlations in data with plots.
- The linear model: correlations, predictions, fit.
- R work: scatterplot(), lm(), cor().

Framing a messege with a plot

Predicting with data

Elections forecasting

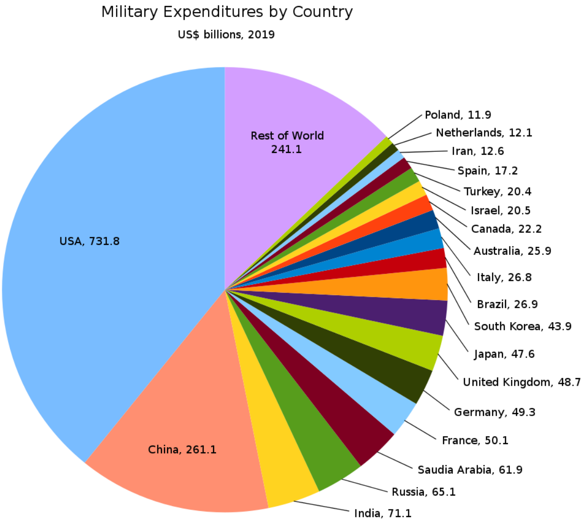




Predicting with Polls

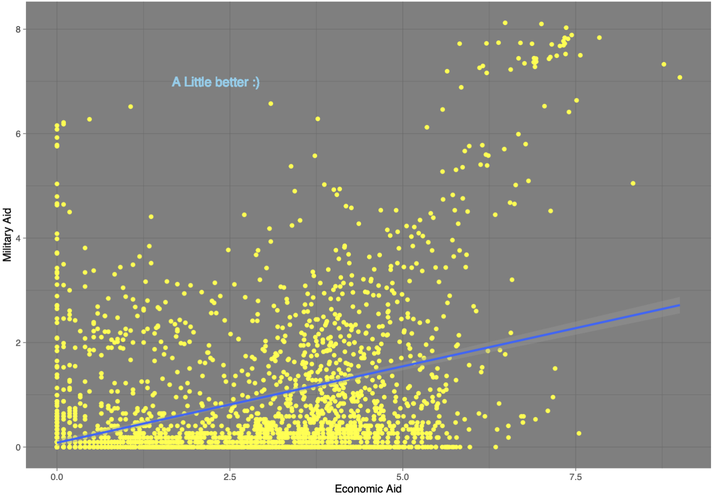
Predicting with data

Military spending → arms race



Predicting with data

Foreign aid (military and economic types)



Predicting with data

Method:

- Calculate values per group.
- Prediction = mean value.
- Elections: 51 US states (2016).
- Arms: 157 countries (1999-2019).
- Main benefit: simple and consistent.
- Foundation for customer outreach: Purchasing (Amazon); Content (Netflix).

However,

- Mean $\rightarrow$ sensitive to outliers/extreme values.
- Median?
- 'Ignore' context of special circumstances.

Better predicting with data

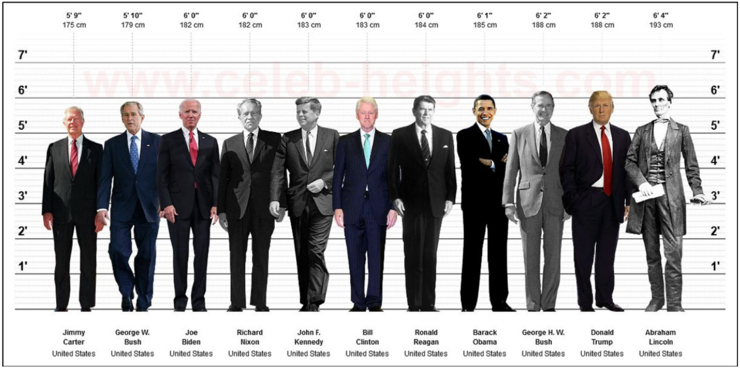
Explore linear relationship between factors

Advanced statistical methods to explore causality:

- Account for average and extreme values.
- Account for confounders.
- Integrate uncertainty in nature.

Data and linear relationship

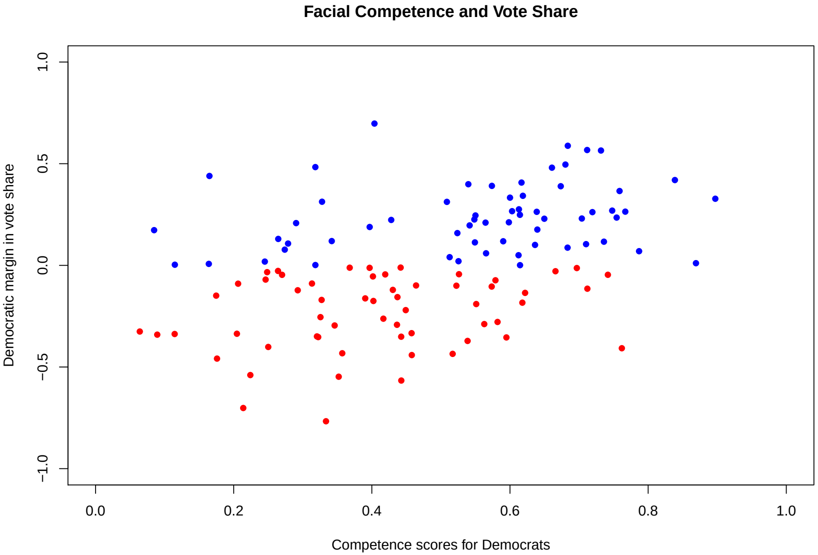
Physical appearance and electoral victory



Data and linear relationship

Facial appearance too?

Data and linear relationship



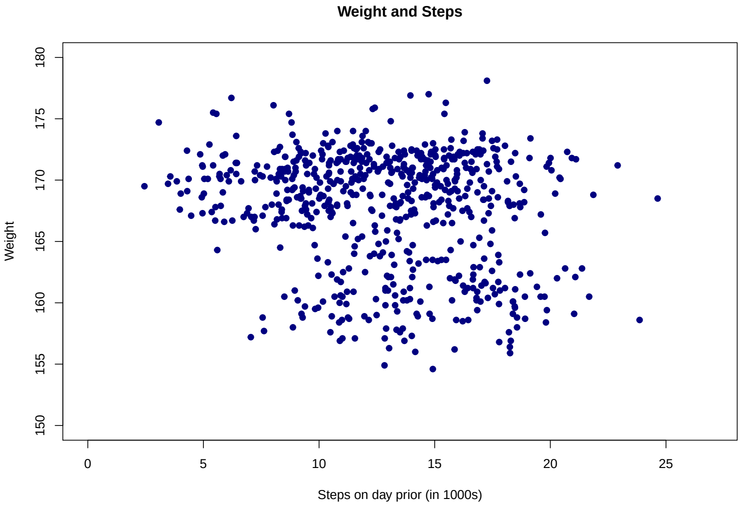
Checking correlation

- Upward trend linking competence score and winning.
- Facial appearance can help winning...
- Is it?

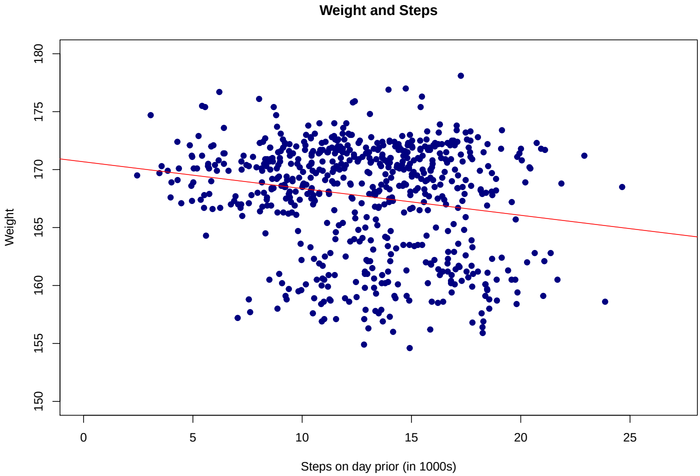
```r
# Correlation
cor(face$d.comp, face$diff.share)
```

```
## [1] 0.4327743
```

More examples



Should I walk to work??




```
cor(health$steps.lag, health$weight)
```

```
## [1] -0.1907032
```

Identify correlation in data

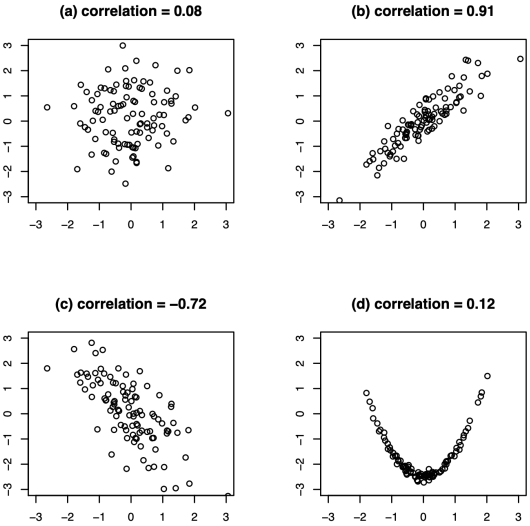
Correlation and scatter plots:

- Positive correlation $\rightarrow$ upward slope
- Negative correlation $\rightarrow$ downward slope
- High correlation $\rightarrow$ tighter, closer to a line
- Correlation cannot capture nonlinear relationship.

Can we see it?

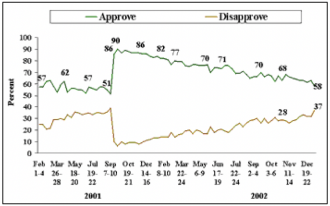
Identify correlation in data

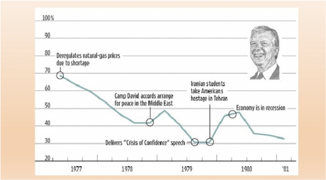
Scatter plots and correlations:



Correlations and predictions: INTA style

Crisis behavior and public approval





Crisis and public approval

Lin-Greenberg (2019):

- Conflict/crisis scenario.
- Actions mitigate public criticism.
- Method: experimental design
- Topic $\rightarrow$ *audience costs*

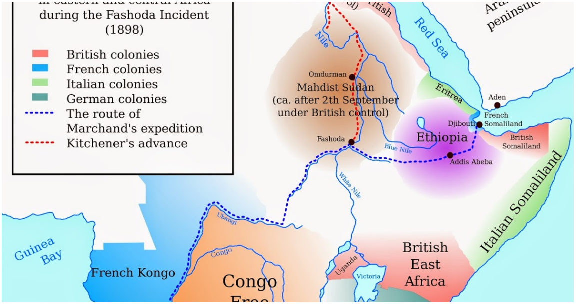
Audience costs

Fearon (1994)

- International crisis $\rightarrow$ "war of nerves"
- Public events, actions (threats, troop movements)
- The role of honor, credibility, and reputation
- Leaders' actions shaped by domestic audience
- The cost of *backing down*
- The strategic implications of audience costs

Audience costs

- Main problem? Observability.
- Can we ‘see’ audience costs?



Measuring audience costs

The solution: experimental research designs

- Conflict scenario
- Leader issues a public threat
- Main treatment: follow-through or back-down
- Compare public approval $\rightarrow$ measure for AC

Are there audience costs?

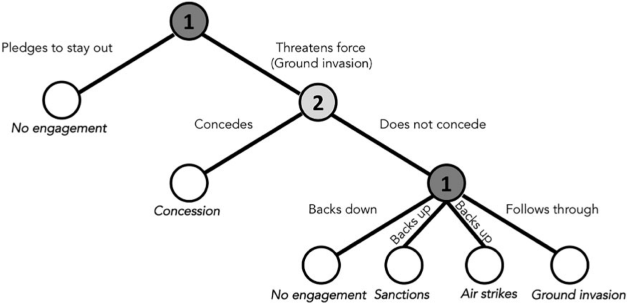
Tomz (2007): experimental design

	Public reaction to empty threat (%)	−	Public reaction to staying out (%)	=	Difference in opinion (%)	Summary of differences (%)
Disapprove						
Disapprove very strongly	31 (27 to 35)		20 (17 to 23)		11 (6 to 17)	16 (10 to 22)
Disapprove somewhat	18 (14 to 21)		13 (10 to 16)		5 (0 to 9)	
Neither						
Lean toward disapproving	8 (6 to 11)		9 (7 to 11)		0 (−3 to 3)	−4 (−9 to 2)
Don't lean either way	21 (17 to 24)		21 (18 to 24)		0 (−5 to 4)	
Lean toward approving	8 (6 to 11)		11 (9 to 14)		−3 (−6 to 0)	
Approve						
Approve somewhat	8 (5 to 10)		13 (11 to 16)		−6 (−9 to −2)	−12 (−17 to −8)
Approve very strongly	6 (4 to 9)		13 (10 to 16)		−7 (−10 to −3)	

Backing-up, not down. . .

Lin-Greenberg (2019):

- Employ less risky action → reduce audience costs



Backing-up, not down...

Backing-up?

Obama's "Red line" (2012-2013)

India-Pakistan standoff (2001-2002)

Measuring audience costs

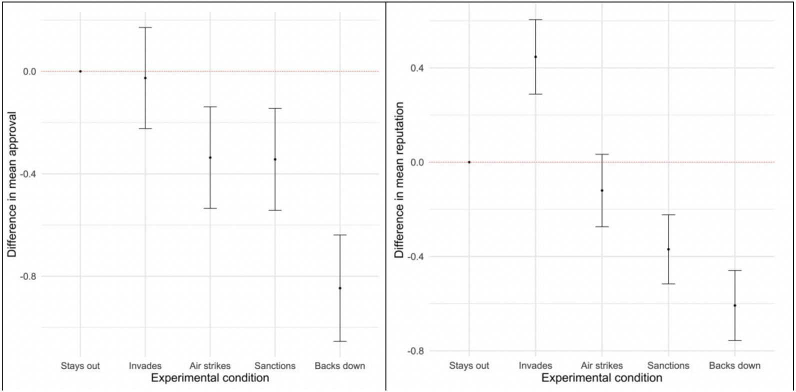
Compare:

- Does policy action matter?
- Approval
- Reputation

Our goal?

- Explore approval & reputation ratings.

Some results



The data

```r
dim(mydata)
```

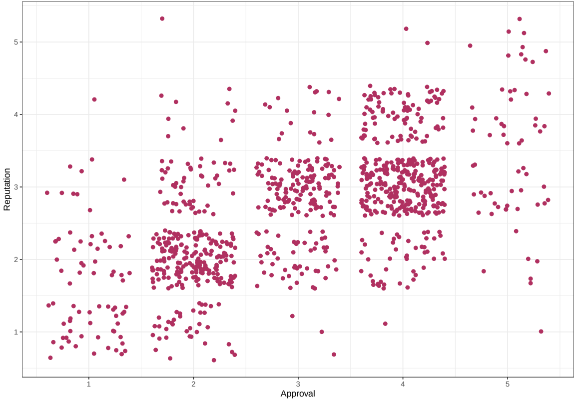
```
## [1] 1006   23
```

```r
head(mydata, n=5)
```

```
##   None Invades Airstrikes Sanctions Backs.Down Intro.Q Approval Justification_2
## 1    0       0          1         0          0       1        4               1
## 2    0       0          1         0          0       1        1               2
## 3    0       0          1         0          0       1        2               9
## 4    0       0          1         0          0       1        5               5
## 5    0       0          1         0          0       1        2               2
##   Justification Criticize.Sitting.Out Consistence Reputation Future.Threats
## 1             1                    NA           3          3              3
## 2             2                    NA           4          2              4
## 3             2                    NA           2          2              3
## 4             1                    NA           4          5              4
## 5             2                    NA           2          2              3
##   Competence FPView Gender Age Education Ideology PolActive Mil Income
## 1          4      3      2  27         6        3         1   1      3
## 2          3      1      2  29         5        4         2   1      1
## 3          3      5      1  36         3        3         2   1      3
## 4          5      5      1  31         5        4         1   1      1
## 5          3      2      1  58         7        2         2   1      2
##   treatment
## 1         3
## 2         3
## 3         3
## 4         3
## 5         3
```

Detecting correlations

```
# Scatter plot: tidyverse approach
ggplot(mydata, aes(Approval,Reputation)) +
  geom_jitter(color = "maroon", cex = 1.9) + theme_bw()
```

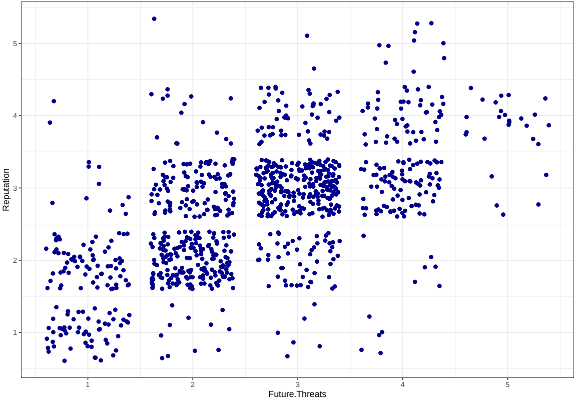



```
cor(mydata$Approval,mydata$Reputation)
```

```
## [1] 0.6221307
```

Detecting correlations

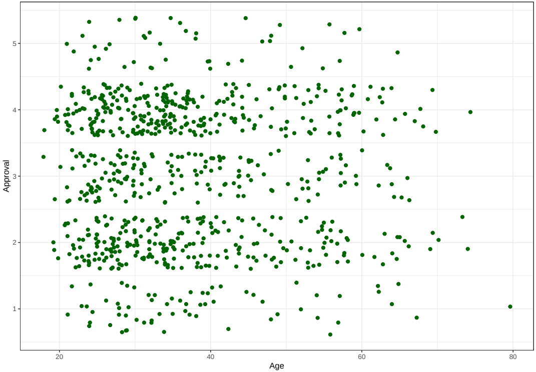
```
# Scatter plot: tidyverse approach
ggplot(mydata, aes(Future.Threats,Reputation)) +
  geom_jitter(color = "darkblue", cex = 1.9) + theme_bw()
```




```
cor(mydata$Future.Threats,mydata$Reputation)
```

```
## [1] 0.6230729
```

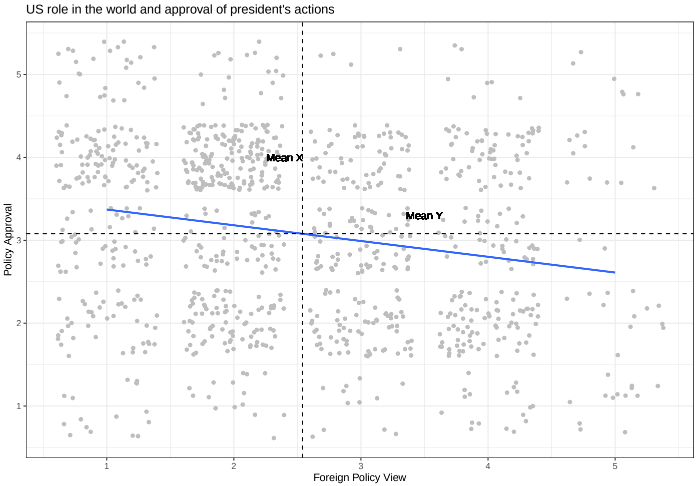
Detecting correlations




```
cor(mydata$Approval,mydata$Age)
```

```
## [1] -0.1106591
```

What about negative correlations?



Negative association

Increase in global involvement & decrease in approval

```
cor(mydata$FPView, mydata$Approval)
```

```
## [1] -0.2001058
```

```
cor(mydata$FPView, mydata$Ideology)
```

```
## [1] 0.1514648
```

Least squared

A Linear model

$$Y = \alpha + \beta * X_i + \epsilon$$

Elements of model:

- *Intercept* (α): the average value of Y when X is zero.
- *Slope* (β): the average increase in Y when X increases by 1 unit.
- *Error/disturbance term* (ϵ): the deviation of an observation from a perfect linear relationship.

Our model:

- **Y** $\rightarrow$ approval for leader's actions.
- **X** $\rightarrow$ leader's actions (back-down or back-up).

Least squared

- Assumption: model $\rightsquigarrow$ Data generation process (DGS)
- **Parameters/coefficients** (α, β): true values unknown.
- Use data to estimate $\alpha, \beta \Longrightarrow \hat{\alpha}, \hat{\beta}$
- Predicting (finally!):
  - Use the *regression line*.
  - Calculate *fitted value* ($\neq$ observed value)

$$\hat{Y} = \hat{\alpha} + \hat{\beta} * x$$

Linear model elements

- *Residual/prediction error*: the difference b-w fitted and observed values.
- Real error is unknown $\Rightarrow \hat{\epsilon}$

$$\hat{\epsilon} = Y - \hat{Y}$$

Linear model estimation

Least squared:

- A method to estimate the regression line.
- Use data (values of Y & X_i).
- 'select' $\hat{\alpha}, \hat{\beta}$ to minimize SSR.
- Calculate RMSE: average magnitude of prediction error (magnitude of least squared).

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

Few more points:

- Mean of residuals ($\hat{\epsilon}$) $== 0$.
- Regression line goes through center of data ($\bar{X}$, $\bar{Y}$).
- $\bar{X}, \bar{Y}$: Sample means of X & Y.

Linear regression in R

Fit the model

- Syntax: lm(Y ∼ x, data = mydata)
- Y = dependent variable; x = independent variable(s).

How does it look like?

Leaders' Audience costs: fitting the model

```
# Fit the model
fit <- lm(Approval ~ FPView, data = mydata)
fit
```

```
##
## Call:
## lm(formula = Approval ~ FPView, data = mydata)
##
## Coefficients:
## (Intercept)       FPView
##      3.5605      -0.1901
```

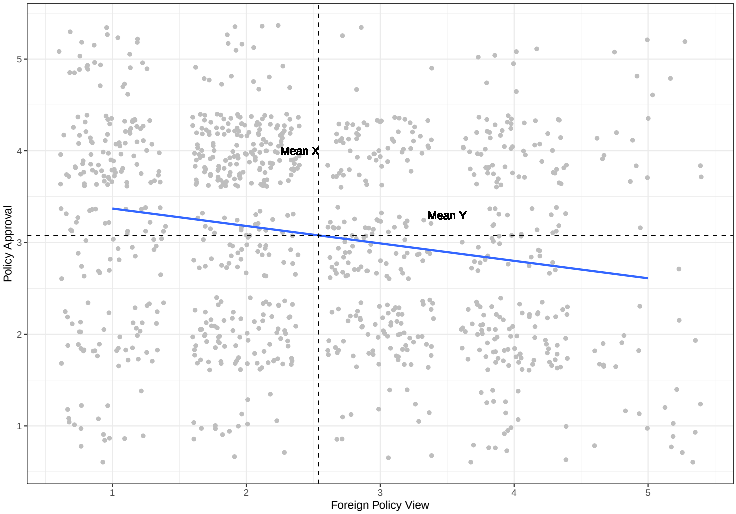
```
# Directly obtain coefficients
coef(fit)
```

```
## (Intercept)      FPView
##   3.5605290  -0.1900987
```

```
# Directly pull fitted values
head(fitted(fit))
```

```
##        1        2        3        4        5        6
## 2.990233 3.370430 2.610036 2.610036 3.180332 3.180332
```

Fitted model on plot



Approval & Reputation: regression models

Back-up (Airstrikes) or Back-down

```
# Fit model
fit2 <- lm(Approval ~ Reputation, data = mydata2)
fit2
```

```
##
## Call:
## lm(formula = Approval ~ Reputation, data = mydata2)
##
## Coefficients:
## (Intercept)   Reputation
##      0.7181       0.8382
```

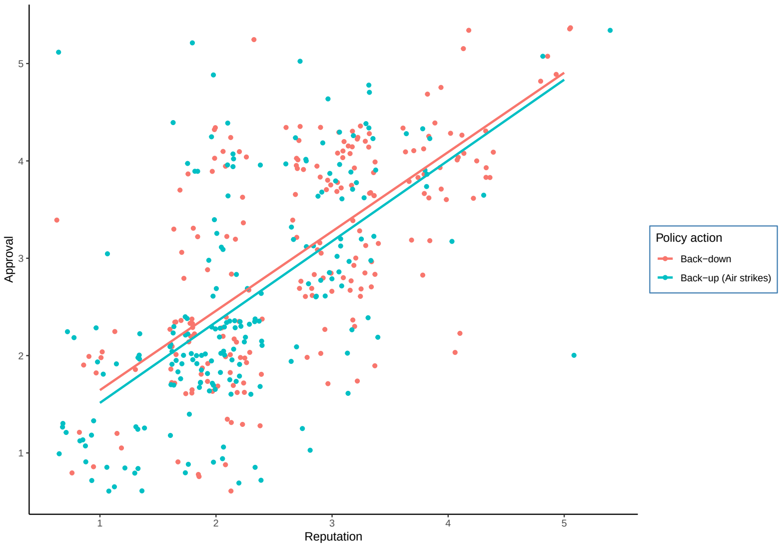
```
# Fitted (predicted) values
head(fitted(fit2))
```

```
##        1        2        3        4        5        6
## 3.232531 2.394373 2.394373 4.908849 2.394373 3.232531
```

```
# Errors
head(resid(fit2))
```

```
##          1          2          3          4          5          6
##  0.7674687 -1.3943726 -0.3943726  0.0911513 -0.3943726  0.7674687
```

Plotting both conditions



Approval & Reputation: different actions

How do leaders' FP actions matter for Approval - Reputation link?

```r
# Subset of Air strike action
mydata3 <- subset(mydata, subset = (treatment == 5))
cor(mydata3$Approval,mydata3$Reputation)
```

```
## [1] 0.6116879
```

```r
# subset of Backing down
mydata4 <- subset(mydata, subset = (treatment == 3))
cor(mydata4$Approval,mydata4$Reputation)
```

```
## [1] 0.688027
```

Least square

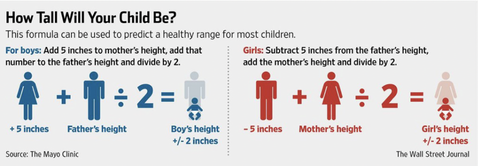
- Regression line $\rightarrow$ "line of best fit"
- Minimize prediction error
- Predictions of fitted line are accurate. How come?
- $\bar{\hat{\epsilon}} = 0$.
- Linear model: not necessarily represent DGS (assumption).

Errors/Curses/Anomalies

Cursed??

Errors/Curses/Anomalies

Fighter pilots performance?



My kids height?

Actually

Regression to the mean

- Empirical - data driven.
- Explained by (random) chance.
- High (low) observations are followed by low (high) observations.
- Observations ‘regress’ towards the average value of the data.

Merging data sets

- Combine data with shared variables.
- Expand data available: more years, same information.
- Technical: use columns / rows.
- Multiple approaches.

Merging

(1) merge function:

- Join two datasets.
- Merge based on common variable (*by* argument).
- 2008-2012 voting data: state Abb. name (QSS pp. 150-151).
- Common variable: matching of rows and columns.
- Other common columns? Appended with .x or .y after name.

(2) cbind function:

- Column binding of multiple datasets.
- Main drawback: assumes similar sorting.
- Keeps duplicates.
- rbind(): join data by rows (add observations to data).

Merging

(3) Join (tidyverse):

- More flexible: multiple options.
- Keep one data, join by common variable.
- Keep all data, join by common variable.

ID	X1
1	a1
2	a2

ID	X2
2	b1
3	b2

inner_join

ID	X1	X2
2	a2	b1

left_join

ID	X1	X2
1	a1	NA
2	a2	b1

right_join

ID	X1	X2
2	a2	b1
3	NA	b2

full_join

ID	X1	X2
1	a1	NA
2	a2	b1
3	NA	b2

semi_join

ID	X1
2	a2

anti_join

ID	X1
1	a1

Apply prediction with regression

- Linear model $\rightarrow$ predict Y using X_i
- Using linear predictions - policy:
  - Predict crime waves - deploy police resources.
  - Predict students performance - target interventions.
- Using linear predictions - business:
  - Predict preferred products based on previous purchases.
  - Predict Netflix/Spotify content based on what I saw/heard?

Model fit

Our well does a linear model predict the data (outcome)?

Model fit:

- Measures to assess model predictive accuracy.

Coefficient of determination (R^2):

- The proportion of total variation in outcome explained by model.
- How much variation in Y explained by our model.
- Values from 0 (no correlation) to 1 (perfect correlation).

Model fit: R-squared

$$R^2 = \frac{TSS - SSR}{TSS}$$

TSS (Total sum of squares): prediction error with mean Y only

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

SSR (Sum of squared residuals): prediction error with model

$$SSR = \sum_{i=1}^{n} \hat{\epsilon}^2$$

Model fit with data: Florida (1996-2000)

Independent candidates ‘inertia’?

```r
# Use summary function
summary(fit3 <- lm(Buchanan00 ~ Perot96, data = florida))
```

```
##
## Call:
## lm(formula = Buchanan00 ~ Perot96, data = florida)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -612.74  -65.96    1.94   32.88 2301.66
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.34575   49.75931   0.027    0.979
## Perot96      0.03592    0.00434   8.275 9.47e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 316.4 on 65 degrees of freedom
## Multiple R-squared:  0.513,  Adjusted R-squared:  0.5055
## F-statistic: 68.48 on 1 and 65 DF,  p-value: 9.474e-12
```

- 51% of Buchanan (2000) explained by Perot (1996) voters.

Model fit with data: Florida (1996-2000)

'Conventional' candidates: Clinton - Gore

```
summary(lm(Gore00 ~ Clinton96, data = florida))
```

```
##
## Call:
## lm(formula = Gore00 ~ Clinton96, data = florida)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30689.3  -1161.5   -622.4   1040.3  23309.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 434.49448  921.26520   0.472    0.639
## Clinton96     1.13120    0.01216  92.997   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6523 on 65 degrees of freedom
## Multiple R-squared:  0.9925, Adjusted R-squared:  0.9924
## F-statistic:  8648 on 1 and 65 DF,  p-value: < 2.2e-16
```

Model fit with data: Florida (1996-2000)

'Conventional' candidates: Dole - Bush

```r
summary(lm(Bush00 ~ Dole96, data = florida))
```

```
##
## Call:
## lm(formula = Bush00 ~ Dole96, data = florida)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18276.9   -781.9   -105.3   1599.5  21759.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 799.82813  701.76481    1.14    0.259
## Dole96        1.27333    0.01262  100.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4587 on 65 degrees of freedom
## Multiple R-squared:  0.9937, Adjusted R-squared:  0.9936
## F-statistic: 1.018e+04 on 1 and 65 DF,  p-value: < 2.2e-16
```

Model fit with data: Florida (1996-2000)

Where did the independents go for the millennium?

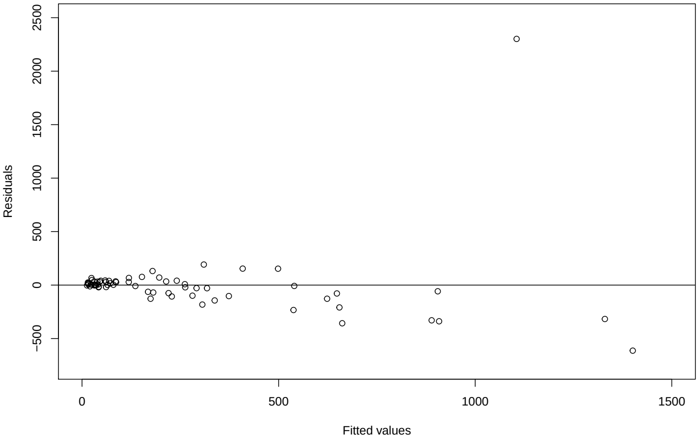
```
summary(lm(Bush00 ~ Perot96, data = florida))
```

```
##
## Call:
## lm(formula = Bush00 ~ Perot96, data = florida)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -49100  -5003  -2951   -582 145169
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1810.4147  3853.0142    0.47     0.64
## Perot96        5.7646     0.3361   17.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24500 on 65 degrees of freedom
## Multiple R-squared:  0.8191, Adjusted R-squared:  0.8163
## F-statistic: 294.2 on 1 and 65 DF,  p-value: < 2.2e-16
```

Model fit with data: Florida (1996-2000)

Maybe not all of them? *Palm beach county*

```
plot(fitted(fit3), resid(fit3), xlim = c(0,1500), ylim = c(-750,2500),
     xlab = "Fitted values", ylab = "Residuals")
abline(h=0)
```



Model fit with data: Florida (1996-2000)

Remove outlier - better prediction

```
summary(lm(Buchanan00 ~ Perot96, data = florida_cut))
```

```
##
## Call:
## lm(formula = Buchanan00 ~ Perot96, data = florida_cut)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -206.70  -43.51  -16.02   26.92  269.03
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 45.841933  13.892746    3.30  0.00158 **
## Perot96      0.024352   0.001273   19.13  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 87.75 on 64 degrees of freedom
## Multiple R-squared:  0.8512, Adjusted R-squared:  0.8488
## F-statistic:   366 on 1 and 64 DF,  p-value: < 2.2e-16
```

Model fit

- R^2: measure of *in-sample* fit.
- *Out-of-sample-fit*: how model predicts outcomes 'outside' the sample.

OVERFITTING:

- OLS $\rightarrow$ good for in-sample.
- Poor performance for out-of-sample.
- Example: use gender to predict 2016 democratic primaries winner.

Avoid overfitting

- Multiple mitigating procedures.
- **Cross validation**:
  - Test set: select randomly.
  - Training set: estimate coefficients.
  - Asses model fit with test set.
  - Repeat test with training set.
  - Average results.

You know machine learning 101!

Wrapping up week 7

Summary:

- Prediction: beyond sample means.
- Using plots to find correlations/trends in data.
- Least squared method.
- Linear model and estimating coefficients.
- Predictions based on linear model.
- Merging data.
- Model fit.

Looking ahead

- **Final Project**:
  - Objective.
  - Technical aspects.
- Next task - research proposal:
  - What is the topic / area?
  - Why important?
  - How will you study it?
  - Sources: substance and data.
  - Final visual product outline.

Proposal due March 22, 2022