Bush 631-603: Quantitative Methods

Lecture 7 (02.28.2023): Prediction vol. II

Rotem Dvir

The Bush school of Government and Public Policy

Texas A&M University

Spring 2023

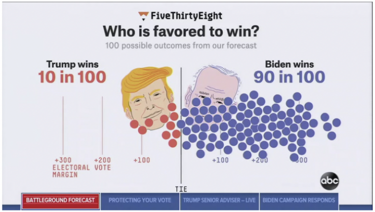
What is today's plan?

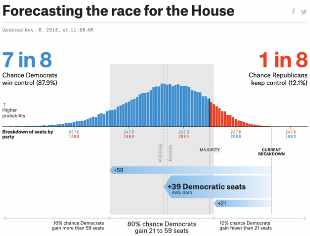
- Predictions: Improved (and more accurate) methods.
- Identify correlations in data with plots.
- The linear model: correlations, predictions, fit.
- R Tech: Plotting in Markdown
- R work: scatterplot(), lm(), cor().

Framing a messege with a plot

Predicting with data

Elections forecasting

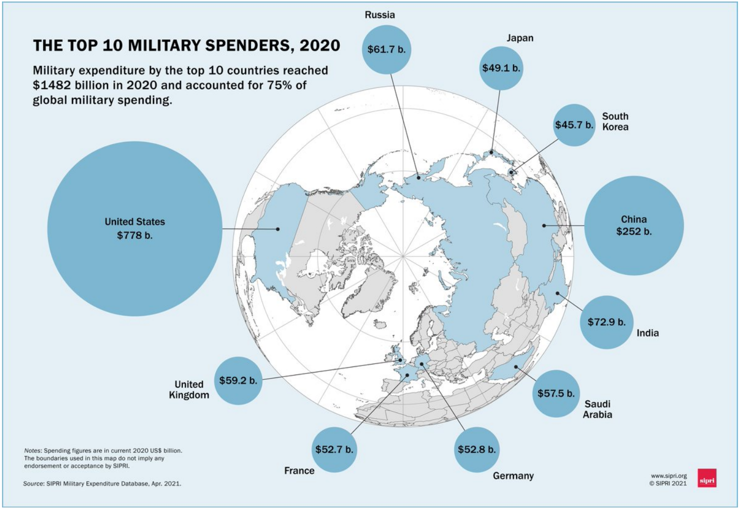




Predicting with Polls

Predicting with data

Military spending $\rightarrow$ arms race



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

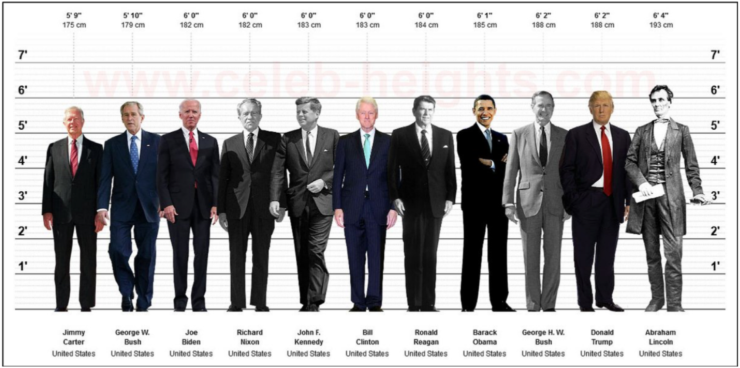
Explore linear relationship between factors

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

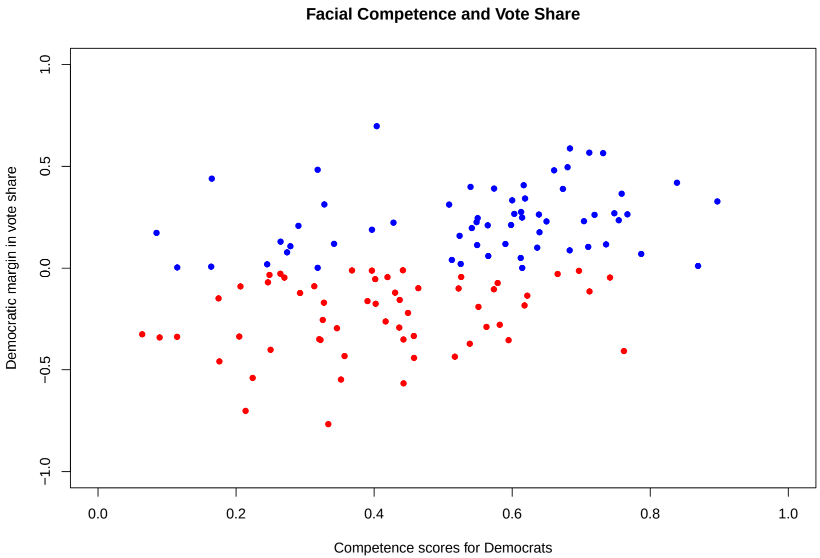


Data and linear relationship

Facial appearance too?

Which person is the more competent?

Data and linear relationship



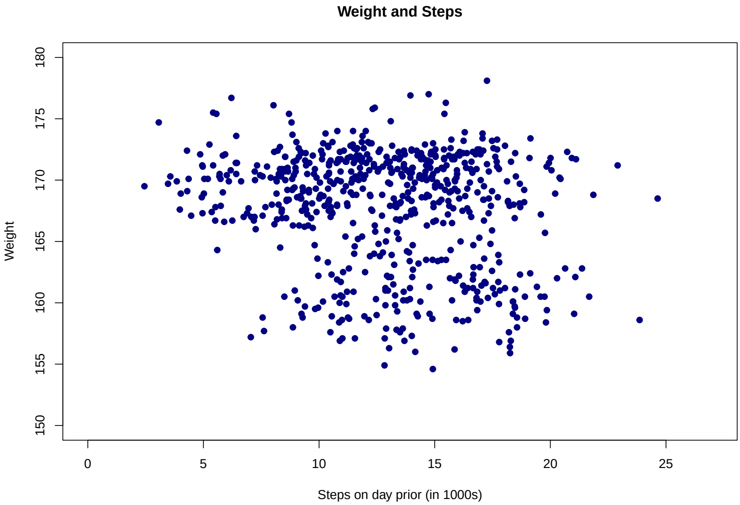
Checking correlation

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

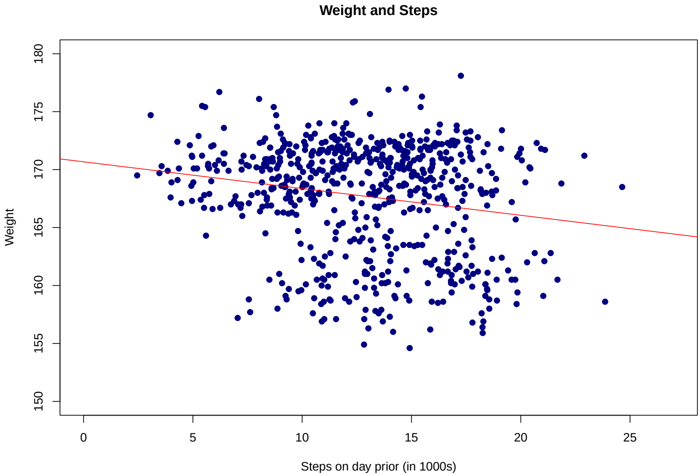
```r
# Correlation
cor(face$d.comp, face$diff.share)
```

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

More examples



Should I walk to work??




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

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

Identify correlation in data

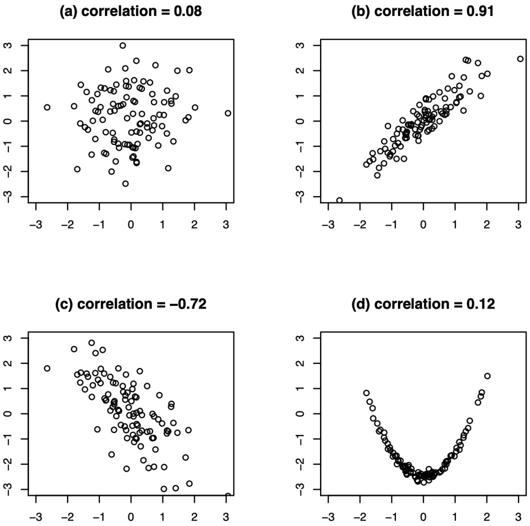
Correlation and scatter plots:

- 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?

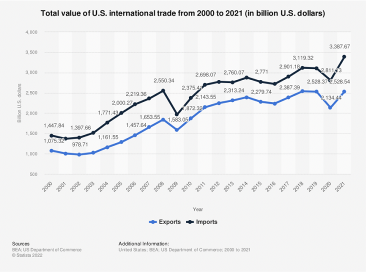
Identify correlation in data

Scatter plots and correlations:



Correlations and predictions: INTA style

Global Trade Flows



- Volume (Q1 - 2022): $7.7 trillion.
- Increases in goods and services (20-25% higher than Q1 2021)

Explaining international trade

The Gravity Model

- "Workhorse of int'l trade"
- Trade volume b-w countries:

1. Size of economies.
2. Distance.

Measuring Gravity and Trade

- Distance, land area, population size, borders, etc.

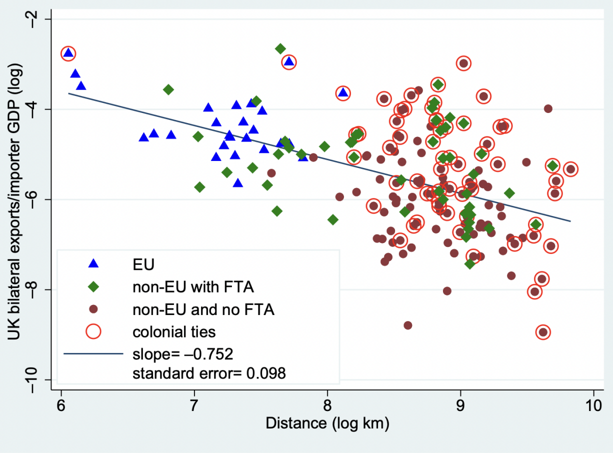


Fig. 2. *UK Bilateral Exports/Importer GDP and Distance, 2017.*

The Gravity Model

Trade and global processes

- International conflict / global alliances:
  - Trade persist b-w strong economies.
  - Weak and strong economy: trade increases with defense pact.
  - Weak and strong economy: trade decays with military conflict.
- Move towards Democratization:
  - Increased trade $\rightarrow$ consolidate democracy.
  - Openness (free trade) increase democratization.

International Trade and democracy promotion

Doces and Magee (2015)

- Benefits of globalization:
  - Abundant labor $\rightarrow$ trade helps workers (and harms capital).
  - Abundant capital $\rightarrow$ trade helps capital (and harms workers).

- Trade $\rightarrow$ strengthen democracy (labor abundant).
- Trade $\rightarrow$ weaken democracy (capital abundant).

Trade and Democracy

- Data: democracy and econ (1960-2007)

```
dim(trade)
```

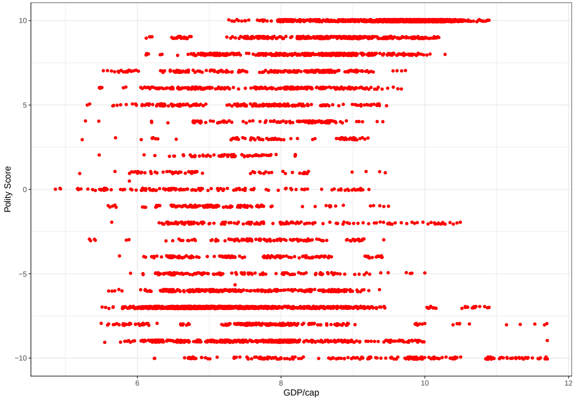
```
## [1] 10421    33
```

```
head(trade, n=5)
```

```
## # A tibble: 5 x 33
##    year my_code open_hat1 wb_code country  pwt_c~1 polity2 America Europe Af
##   <dbl>   <dbl>     <dbl> <chr>   <chr>      <dbl>   <dbl>   <dbl>  <dbl>  <
## 1  1960       1      15.8 AFG     Afghani~      NA     -10       0      0
## 2  1961       1      15.7 AFG     Afghani~      NA     -10       0      0
## 3  1962       1      15.5 AFG     Afghani~      NA     -10       0      0
## 4  1963       1      16.1 AFG     Afghani~      NA     -10       0      0
## 5  1964       1      17.5 AFG     Afghani~      NA      -7       0      0
## # ... with 23 more variables: Pacific <dbl>, oil <dbl>,
## #   female_percent_pop <dbl>, pop_15_64 <dbl>, pop_15_under <dbl>, urban <db
## #   region_polity_20 <dbl>, region_polity_10 <dbl>, region_open_20 <dbl>,
## #   region_open_10 <dbl>, lang_num2 <dbl>, ethnic_num2 <dbl>,
## #   religion_num2 <dbl>, colony_1945 <dbl>, yrs_indep <dbl>, time <dbl>,
## #   open3 <dbl>, ln_gdppc8 <dbl>, ln_pop8 <dbl>, kl8 <dbl>, median_kl8 <dbl>
## #   above_median_kl8 <dbl>, above_avg_kl8 <dbl>, and abbreviated variable ..
```

Gravity model - Trade data

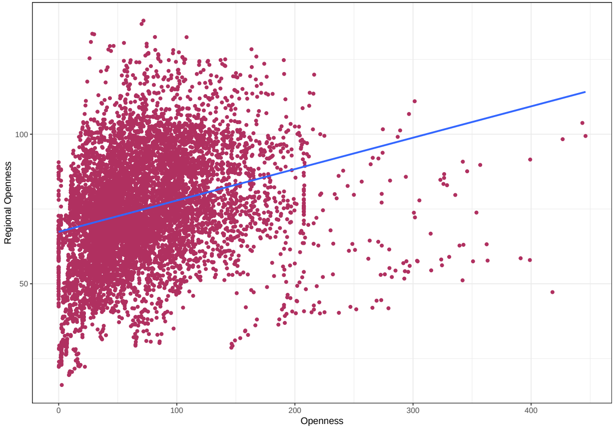
```
ggplot(trade, aes(ln_gdppc8,polity2)) +
  geom_jitter(color = "red", size = 1.3) +
  theme_bw() +  xlab("GDP/cap") + ylab("Polity Score")
```




```
cor(trade$polity2,trade$ln_gdppc8, use = "complete")
```

```
## [1] 0.4396297
```

Gravity model - Trade data




```
## [1] 0.2756198
```

Trade and democracy - with a caveat

Labor abundant → more workers: Trade boost democracy

```r
# Only labor abundant countries
labor.trade <- trade %>%
  filter(above_median_kl8 == 0)

# Trade and religious diversity
cor(labor.trade$open3, labor.trade$religion_num2, use = "complete")
```

```
## [1] -0.210249
```

```r
# Trade and working population
cor(labor.trade$open3, labor.trade$pop_15_64, use = "complete")
```

```
## [1] 0.1137331
```

```r
# Trade and Democracy
cor(labor.trade$open_hat1, labor.trade$polity2, use = "complete")
```

```
## [1] 0.1928551
```

Trade and democracy - with a caveat

Capital abundant $\rightarrow$ less workers: Trade harms democracy

```
# Only capital abundant countries
cap.trade <- trade %>%
  filter(above_median_kl8 == 1)

# Trade and religious diversity
cor(cap.trade$open3, cap.trade$religion_num2, use = "complete")
```

```
## [1] 0.3193981
```

```
# Trade and Democracy
cor(cap.trade$open_hat1, cap.trade$polity2, use = "complete")
```

```
## [1] -0.09662274
```

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 change 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** $\rightarrow$ Democracy score (polity).
- **X** $\rightarrow$ Extent of int'l trade (openness).

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.
- Capture the gap b-w actual values (data) and predictions.
- 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?

Trade and democracy - fitting the model

```r
# Fit the model
fit <- lm(polity2 ~ open3, data = trade)
fit
```

```
##
## Call:
## lm(formula = polity2 ~ open3, data = trade)
##
## Coefficients:
## (Intercept)        open3
##    0.711664    -0.006503
```

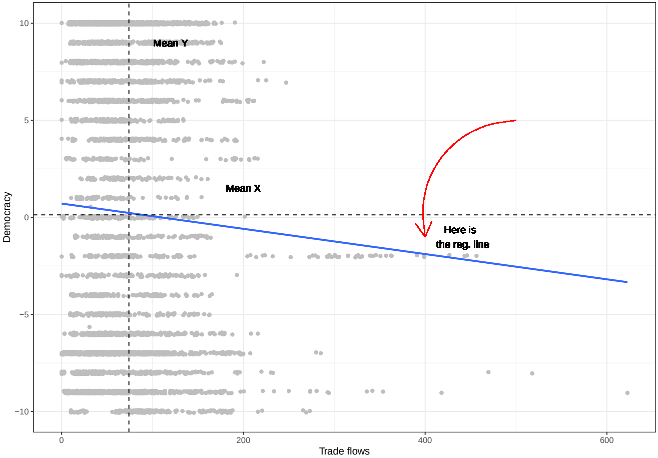
```r
# Directly obtain coefficients
coef(fit)
```

```
##  (Intercept)        open3
##  0.711663968 -0.006502699
```

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

```
##         1         2         3         4         5         6
## 0.6452021 0.6453733 0.6452360 0.6022816 0.5264752 0.5652951
```

Trade and democracy - visualize the model



Trade and democracy - labor vs. capital

```r
# Labor abundant
fit2 <- lm(polity2 ~ open3, data = labor.trade)
fit2
```

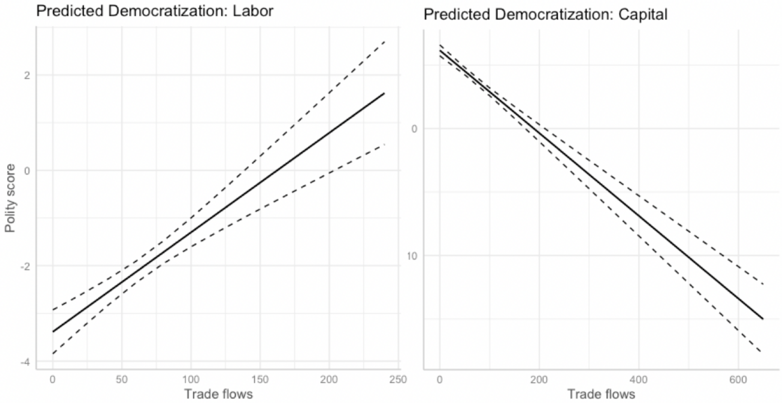
```
##
## Call:
## lm(formula = polity2 ~ open3, data = labor.trade)
##
## Coefficients:
## (Intercept)        open3
##    -3.38618      0.02086
```

```r
# Capital abundant
fit3 <- lm(polity2 ~ open3, data = cap.trade)
fit3
```

```
##
## Call:
## lm(formula = polity2 ~ open3, data = cap.trade)
##
## Coefficients:
## (Intercept)        open3
##     6.14921     -0.03257
```

Trade and democracy - labor vs. capital

Predicted Polity score $\leftarrow$ Trade volume



Least square

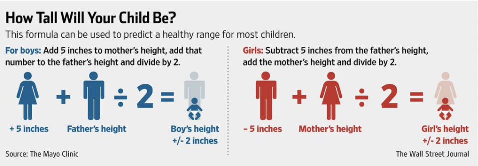
- Regression line $\rightarrow$ "line of best fit"
- Minimize prediction error
- Predictions of fitted line are accurate. How come?
- $\bar{\hat{\epsilon}} = 0$.
- Linear model: not necessarily represent DGS (assumption).

Errors/Curses/Anomalies

$\boxed{\text{Cursed??}}$

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.

R Tech

- Plotting in Markdown plots
- Code chunk definitions: echo, include, message.
- Display plot: out.width, fig.align.
- How to add/remove code for plots.

Working with R - Class Task

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

- Upload data - STATA file!!
- Use base R or tidyverse.
- Plot variable: real GDP per capita
- Plot variables: Civilian fatalities vs. police and military fatalities.

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

```
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?

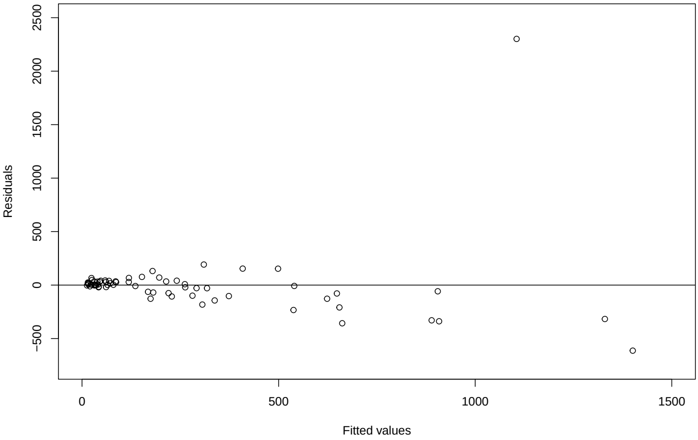
```
summary(lm(Bush00 ~ Perot96, data = florida))
```

```
##
## 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*

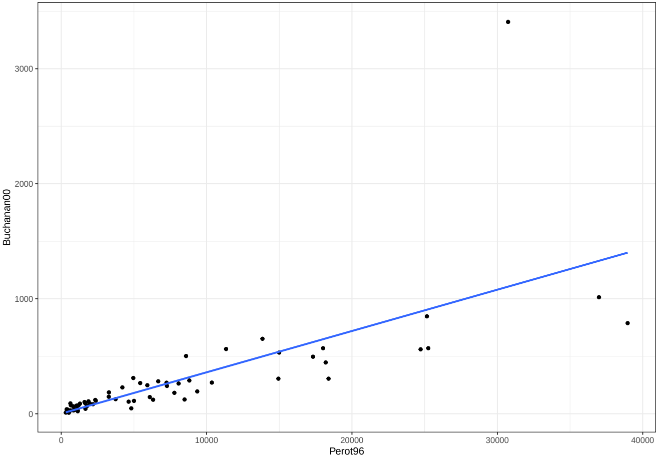
```
plot(fitted(fit3), resid(fit3), xlim = c(0,1500), ylim = c(-750,2500),
     xlab = "Fitted values", ylab = "Residuals")
abline(h=0)
```



Model fit with data: Florida (1996-2000)

How's the correlation?

```
# Plotting Dole/Buchanan correlation
ggplot(florida, aes(x=Perot96, y=Buchanan00)) +
  geom_point() + geom_smooth(method = "lm", se = F) +
  theme_bw()
```



Model fit with data: Florida (1996-2000)

Remove outlier - better prediction

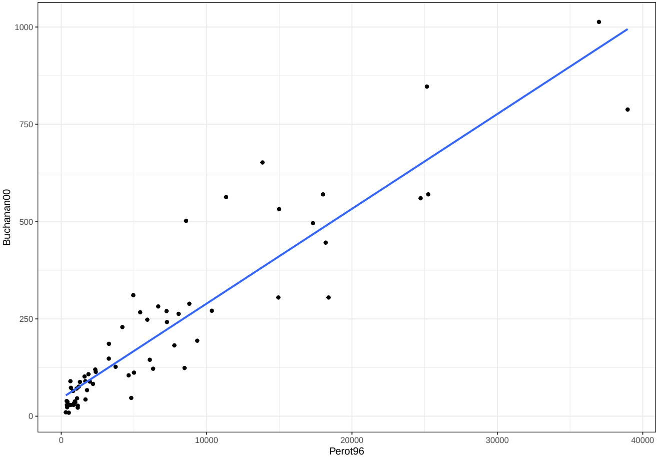
```
summary(lm(Buchanan00 ~ Perot96, data = florida_cut))
```

```
##
## 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 with data: Florida (1996-2000)

And now? How's the correlation?

```
# Plotting Dole/Buchanan correlation
ggplot(florida_cut, aes(x=Perot96, y=Buchanan00)) +
  geom_point() + geom_smooth(method = "lm", se = F) +
  theme_bw()
```



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.

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.

R Task Friday at midnight!!