Introduction to Social Data Analytics: Class 23

Today: Regression in R

By the end of today's lecture, you should be able to:

- Perform regression analysis to determine linear relationships between variables
- Interpret coefficient estimates and add best fit lines to scatter plots of the data

We'll work with three datasets: nba.csv, florida.csv, and face.csv.

Review of regression basics

- Regression: finds the best fit line between k independent and one dependent variables.
- One independent variable case: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- Produces estimates: $\hat{\beta}_0$ and $\hat{\beta}_1$
- Interpretation:
  - $\hat{\beta}_1$: a one-unit increase in x-units is associated with a $\hat{\beta}_1$ average increase in y-units.
  - $\hat{\beta}_0$: expected value of y when $x = 0$.

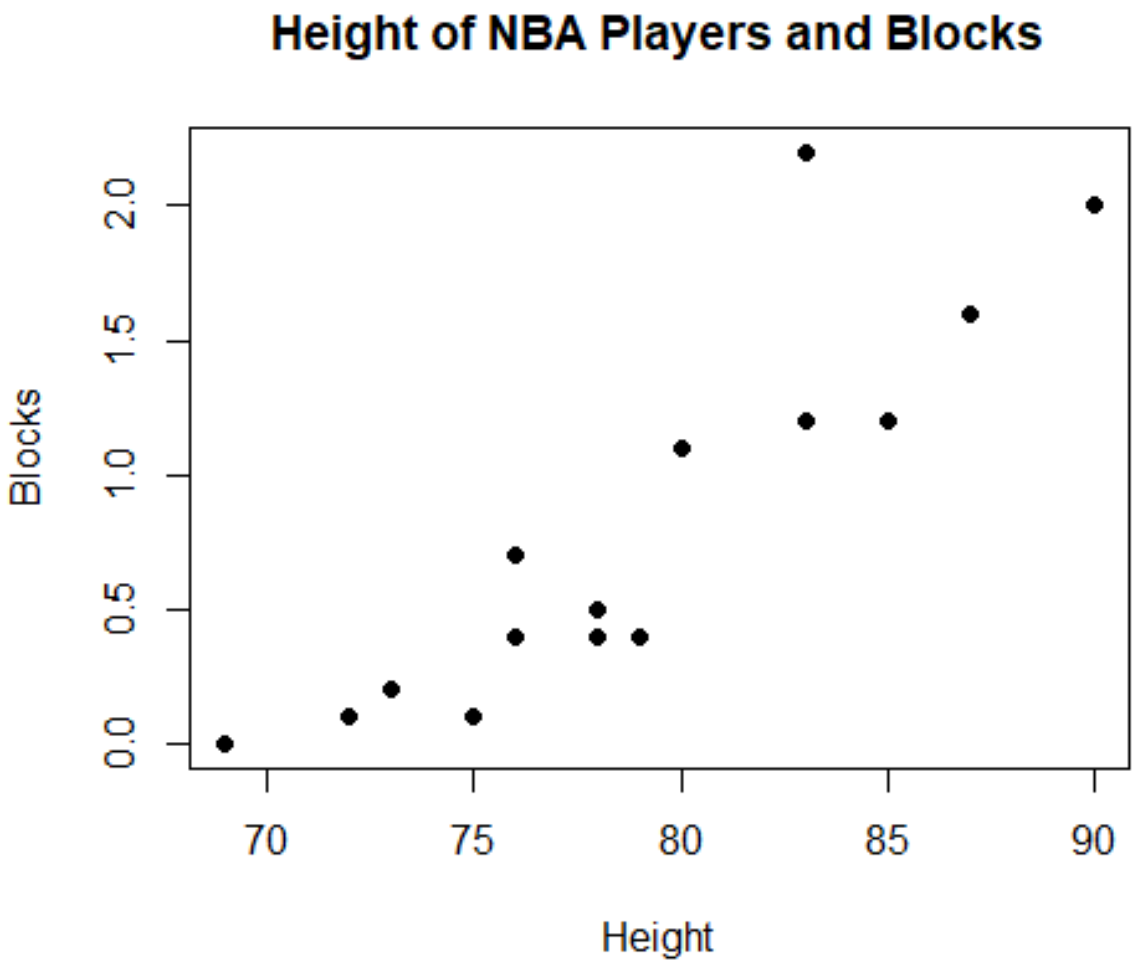
Some useful terms

- Fitted values, predicted values, $\hat{y}$:
  - predicted y when x is set to a certain value.
  - predicted y for x's observed in the dataset
- Residuals
  - Difference between predicted and actual value of y.
- Let's return to the NBA dataset.

Height and Blocks

```
# Make scatter plot
plot(nba$height, nba$blocks,
  xlab = "Height",
  ylab = "Blocks",
  pch = 16,
  main = "Height of NBA Players and Blocks")
```

Height and Blocks



Estimating Height and Blocks Relationship

```
# What is the correlation between height and blocks?
cor(nba$height, nba$blocks)
[1] 0.8797669

# Fit a regression line of the form:
#   blocks = b0 + b1 * height
fit <- lm(blocks ~ height, data=nba)

# Check the results
summary(fit)
```

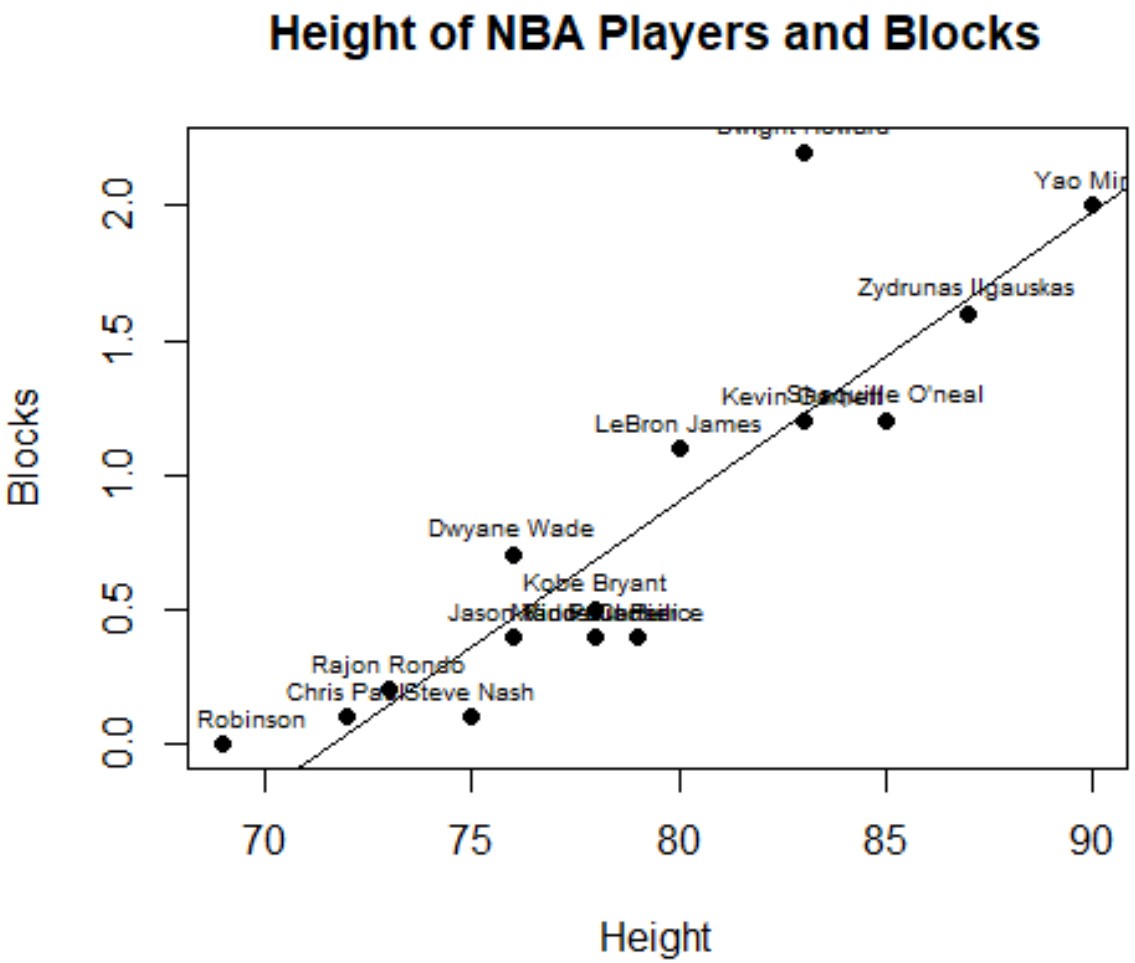
Estimating Height and Blocks Relationship

```
summary(fit)

Call:
lm(formula = blocks ~ height, data = nba)

Residuals:
     Min       1Q   Median       3Q      Max
-0.39475 -0.24761 -0.04251  0.09506  0.97341

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.73418    1.23275  -6.274 2.04e-05 ***
height       0.10796    0.01559   6.924 7.05e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1
```

Add best fit line and players' names to plot

```
# Now add the best fit line to the plot
abline(fit)

# Now add the players' names
text(nba$height, nba$blocks+.1, nba$name, cex=0.6)
```

Height and Blocks



Election 2000, Butterfly Ballot

- Suppose we wanted to predict third-party votes in Florida
- Use 1996 election (Perot) to predict 2000 election (Buchanan)

```
head(florida)
    county Clinton96 Dole96 Perot96 Bush00 Gore00 Buchanan00
1  Alachua     40144  25303    8072  34124  47365        263
2    Baker      2273   3684     667   5610   2392         73
3      Bay     17020  28290    5922  38637  18850        248
4 Bradford      3356   4038     819   5414   3075         65
5  Brevard     80416  87980   25249 115185  97318        570
6  Broward    320736 142834   38964 177323 386561        788
```

Do Votes for Perot Predict Votes for Buchanan?

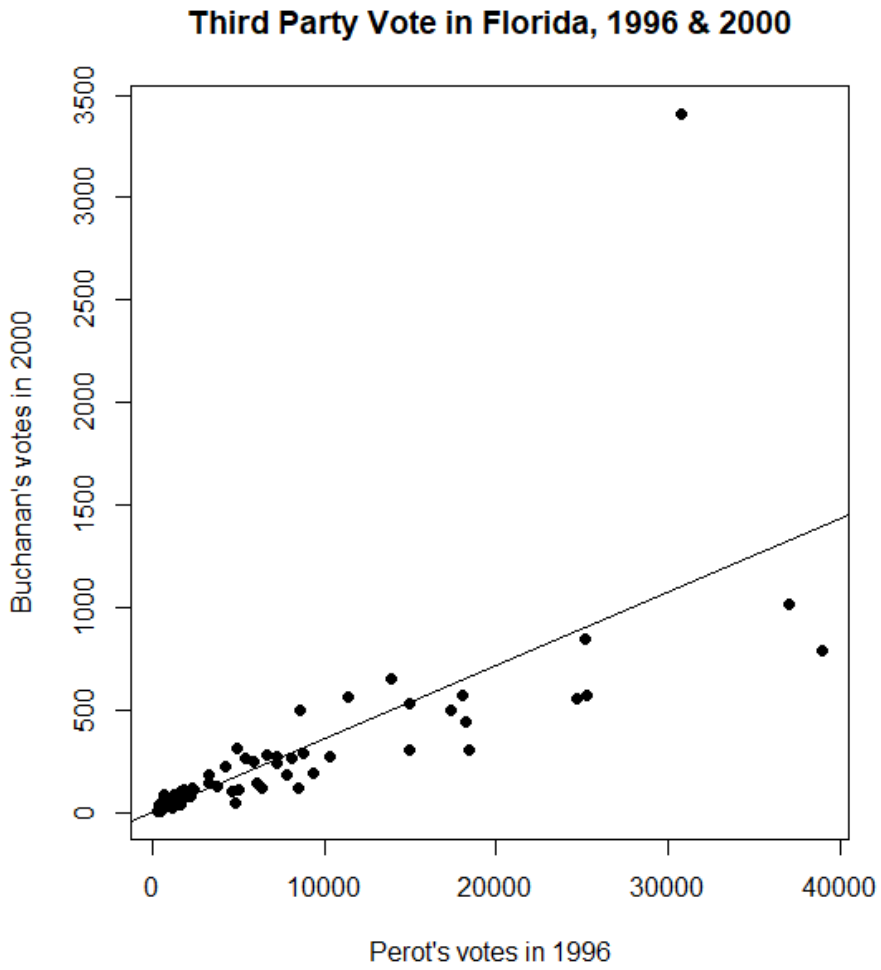
Your task is to do the following:

- Plot votes for Perot96 vs votes for Buchanan00 (both third party presidential candidates)
- Fit a best fit line through your scatterplot
- What seems off?

Do Votes for Perot Predict Votes for Buchanan?

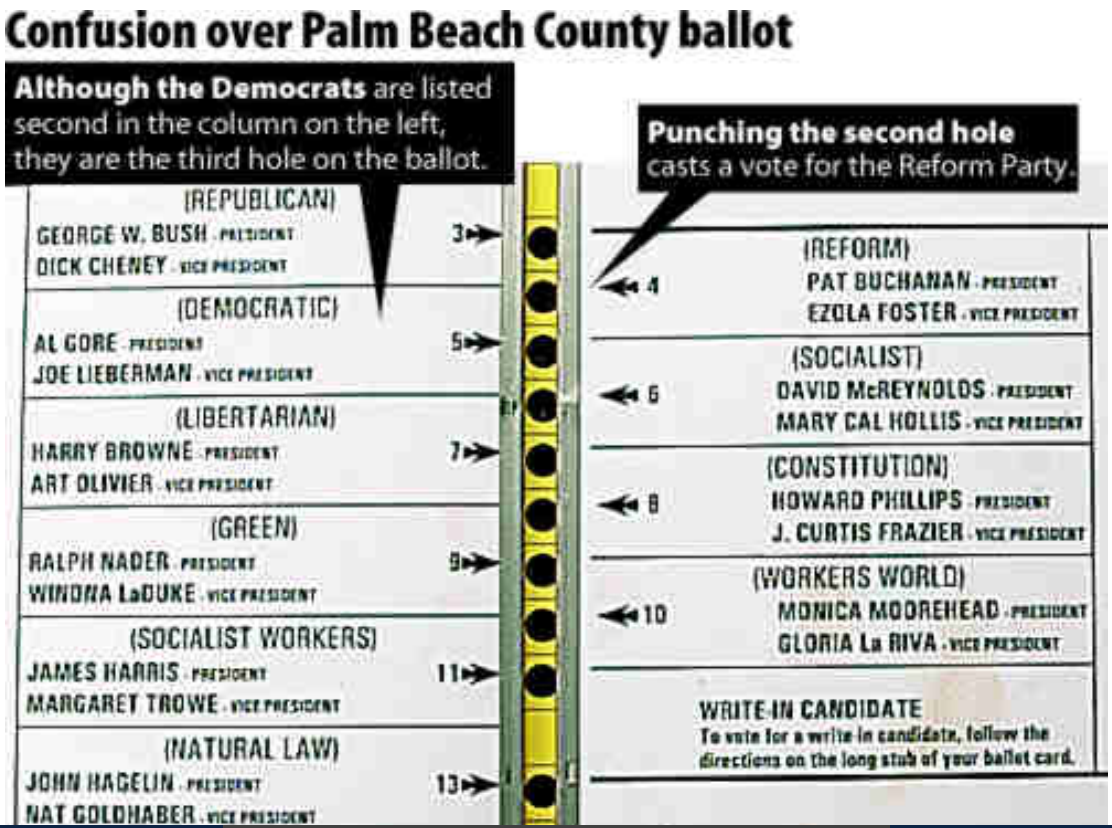
```
fit2 <- lm(Buchanan00 ~ Perot96, data = florida)
summary(fit2)

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 ***
```

What seems off?



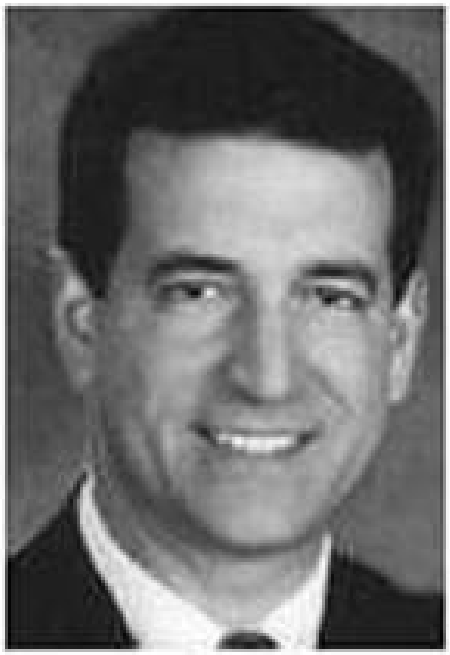
Which county in Florida is the outlier?

```
florida$county[resid(fit2) == max(resid(fit2))]
[1] PalmBeach
```



Regression with experiments

- Experiment: Does facial appearance predict vote outcomes?
- Todorov et al (2005):
  - Quickly show picture of candidates to subjects
  - Ask them to rate on "competence"



Which person is more competent?

The data

```
head(face)
```

```
  year state    winner     loser w.party l.party    d.comp    r.comp d.votes r.vote
1 2000    CA Feinstein  Campbell       D       R 0.5645676 0.4354324 5790154 377932
2 2000    DE    Carper      Roth       D       R 0.3419122 0.6580878  181387  14268
3 2000    FL    Nelson  McCollum       D       R 0.6123680 0.3876320 2987644 270360
4 2000    GA    Miller Mattingly       D       R 0.5415328 0.4584672 1390428  93369
5 2000    HI     Akaka   Carroll       D       R 0.6802323 0.3197677  251130   8465
6 2000    IN     Lugar   Johnson       R       D 0.3205024 0.6794976  684242 141962
```

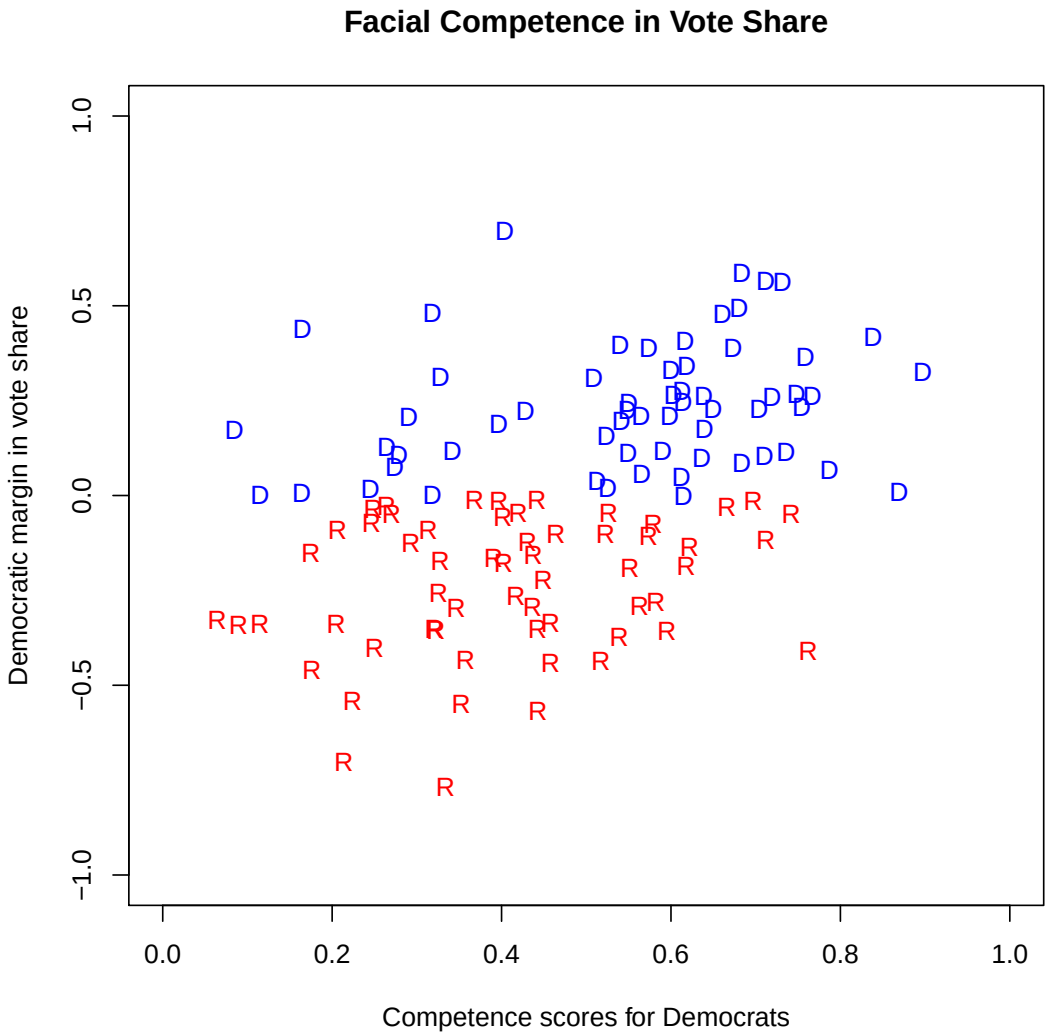
`d.comp` is the fraction of subjects who think the democratic candidate appears more competent.

Does Facial Appearance Predict Vote Share?

```
#define vote share for Dems and Reps
face$d.share <- face$d.votes/(face$d.votes + face$r.votes)
face$r.share <- face$r.votes/(face$d.votes + face$r.votes)
face$diff.share <- face$d.share - face$r.share
```

- Add a plot of Democratic candidate's competence score on the x-axis and Democratic margin in vote share on the y-axis, and label as appropriate.
- Regress the model diff.share $= \beta_0 + \beta_1 *$ d.comp
- Add a trendline to your plot
- Is there a causal effect of appearance on voting outcome?

Does Facial Appearance Predict Vote Share?



How do we interpret these coefficients?

```
fit3 <- lm(diff.share ~ d.comp, data = face)
summary(fit3)

Call:
fit3 <- lm(formula = diff.share ~ d.comp, data = face)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67487 -0.16600  0.01399  0.17741  0.74297

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.31223    0.06596  -4.733 6.24e-06 ***
d.comp       0.66038    0.12718   5.193 8.85e-07 ***
---
```

Does Facial Appearance Predict Vote Share?