Week 2 - Finding and Describing Associations

Ryan Miller

Week #2 Outline

- Video #1
  - Comparing Groups
- Video #2
  - Correlation
- Video #3
  - Regression

Introduction

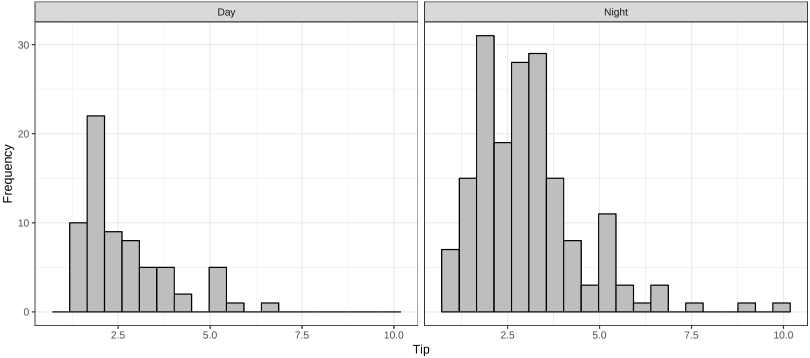
- Last week, we introduced *contingency tables* as a method for summarizing relationships between *two categorical variables*

Introduction

- Last week, we introduced *contingency tables* as a method for summarizing relationships between *two categorical variables*
- This week, we'll cover methods for summarizing relationships for combinations of variables of other types
  - Side-by-side graphs and differences (categorical and quantitative)
  - Correlation and regression (quantitative and quantitative)

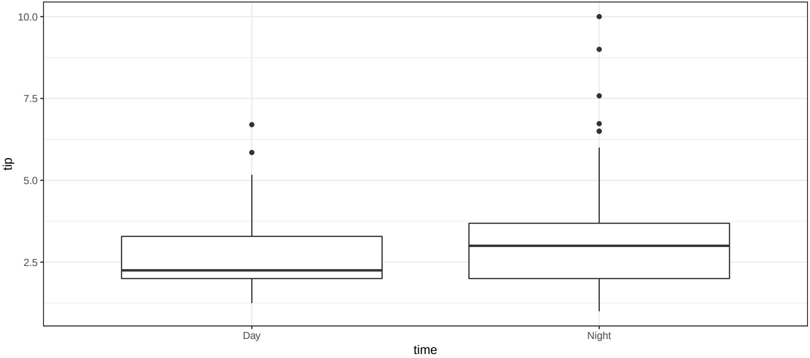
Side-by-side Graphs

- A simple way of comparing two or more groups (as defined by a categorical variable) is split up the cases by group and graph them side-by-side



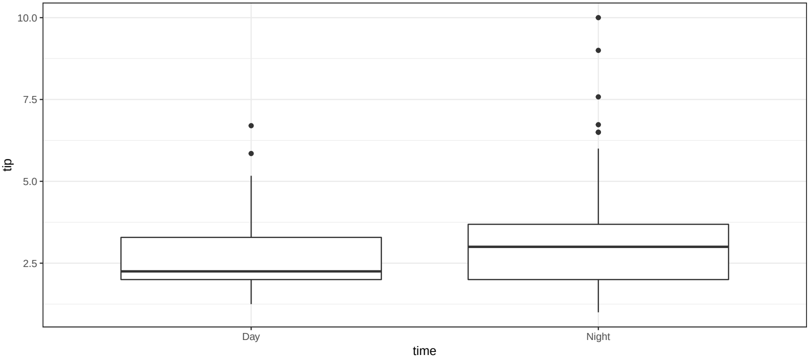
Side-by-side Graphs

- Boxplots tend work better for this since they easily facilitate direct comparisons (ie: median vs. median)



Association

- Recall that two variables are **associated** if the distribution of one variable depends upon the other
  - Thus, substantial differences in *any single summary measure* (medians, Q1, etc.) suggests an association, even if other parts of the distributions are similar



Numeric Summaries

- Boxplots are just a visual representation of several different numeric summaries (minimum, Q1, median, Q3, and maximum)
  - So we can also find and describe associations using side-by-side numeric summaries

time	min	Q1	median	mean	Q3	max
Day	1.25	2	2.25	2.728088	3.2875	6.7
Night	1.00	2	3.00	3.102670	3.6875	10.0

Reporting Associations

- Being able to identify an association is important, but we also need to be able to describe it to others with sufficient precision
    - As an example, we might report an association between tip and time in the Tips dataset by saying:

  "The mean tip at Dinner is 38 cents (0.38 dollars) higher than the mean tip at Lunch"
- In this class, the **difference in means** will be our go-to when reporting an association between two groups
    - That said, nothing prevents us from reporting a *difference in medians* or a *difference in 90th percentiles*

Practice

Using the “Tips” dataset, available by clicking here or on our website, go to https://www.lock5stat.com/StatKey/index.html, and click on the “One Quantitative and One Categorical” menu in the “Descriptive Statistics and Graphs” section

1. Upload the relavent columns from the “Tips” data to create boxplots that show the relationship between smoking status and tip amount
2. Report the *difference in means* for tips given by smokers and non-smokers
3. Report the *difference in medians* for tips given by smokers and non-smokers
4. Which difference do you think is better to report?

Practice (solution)

2. The difference in means is 3.009 - 2.992 $= 0.017$
3. The difference in medians is 3.00 - 2.74 $= 0.26$
4. Because both distributions are skewed right and contain outliers, we should report the difference in medians

Two Quantitative Variables

We've now discussed how to find and report associations in two contexts:

1. Two categorical variables - contingency tables and comparisons row/column proportions
2. One categorical and one quantitative variable - side-by-side graphs and differences in means/medians

We'll now cover the remaining scenario, two quantitative variables

Pearson's Height Data

- Francis Galton and Karl Pearson, two pioneers of modern statistics, lived in Victorian England at a time when the scientific community was fascinated by the idea of quantifying hereditable traits

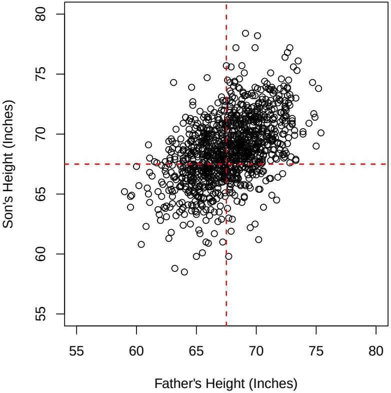
Pearson's Height Data

- Francis Galton and Karl Pearson, two pioneers of modern statistics, lived in Victorian England at a time when the scientific community was fascinated by the idea of quantifying hereditable traits
- Wondering if height is hereditable, they measured the heights of 1,078 fathers and their (fully grown) first-born sons:

Father	Son
65	59.8
63.3	63.2
65	63.3
65.8	62.8
...	...

Pearson's Height Data

Using a scatterplot an association is obvious:



But how do we summarize it?

Pearson's Correlation Coefficient

- Consider two variables, X and Y, and their average values, $\bar{x}$ and $\bar{y}$
- The correlation coefficient, r, measures the strength of a *linear association* between X and Y

$$r_{xy} = \frac{1}{n-1} \sum_i \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Pearson's Correlation Coefficient

- Consider two variables, X and Y, and their average values, $\bar{x}$ and $\bar{y}$
- The correlation coefficient, r, measures the strength of a *linear association* between X and Y

$$r_{xy} = \frac{1}{n-1} \sum_i \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- As you can see, when *above average* values in X are accompanied by *above average* values in Y there is a *positive contribution* to the correlation between X and Y

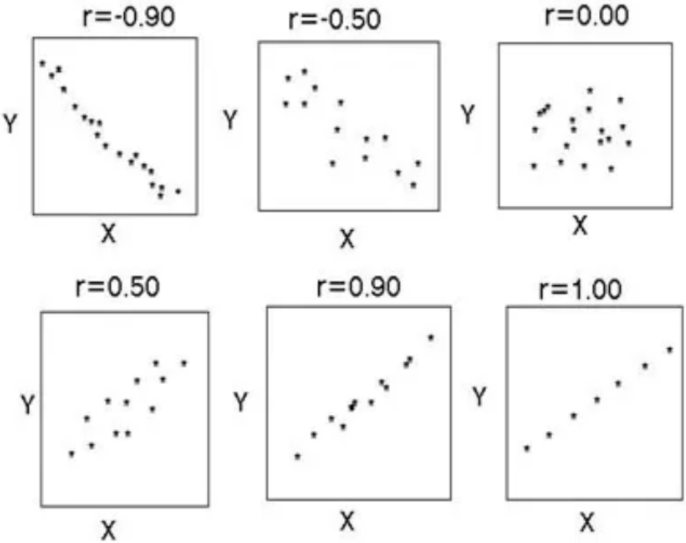
Pearson's Correlation Coefficient

- Consider two variables, X and Y, and their average values, $\bar{x}$ and $\bar{y}$
- The correlation coefficient, r, measures the strength of a *linear association* between X and Y

$$r_{xy} = \frac{1}{n-1} \sum_i \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

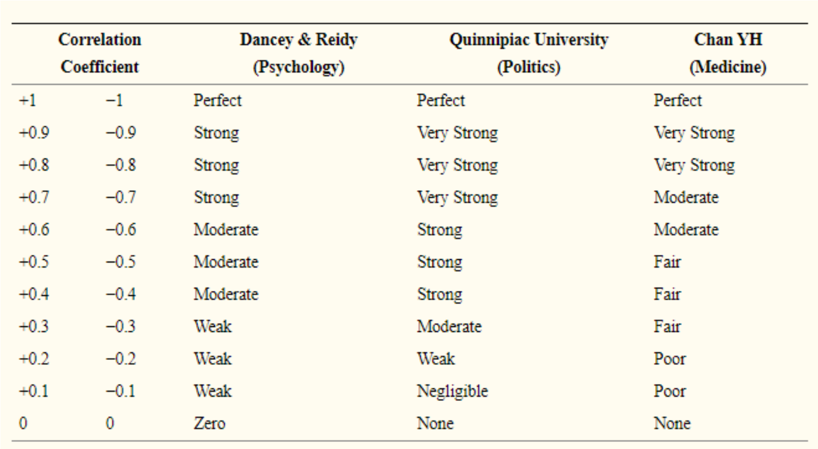
- As you can see, when *above average* values in X are accompanied by *above average* values in Y there is a *positive contribution* to the correlation between X and Y
- When *above average* values in X are accompanied by *below average* values in Y there is a *negative contribution* to the correlation between X and Y

Correlation Coefficient Examples



Strength of Association

Whether a correlation is considered “strong” or “weak” depends on the discipline



Correlation Coefficient		Dancey & Reidy (Psychology)	Quinnipiac University (Politics)	Chan YH (Medicine)
+1	−1	Perfect	Perfect	Perfect
+0.9	−0.9	Strong	Very Strong	Very Strong
+0.8	−0.8	Strong	Very Strong	Very Strong
+0.7	−0.7	Strong	Very Strong	Moderate
+0.6	−0.6	Moderate	Strong	Moderate
+0.5	−0.5	Moderate	Strong	Fair
+0.4	−0.4	Moderate	Strong	Fair
+0.3	−0.3	Weak	Moderate	Fair
+0.2	−0.2	Weak	Weak	Poor
+0.1	−0.1	Weak	Negligible	Poor
0	0	Zero	None	None

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6107969/

Practice

Using the “Tips” dataset, available by clicking here or on our website, go to https://www.lock5stat.com/StatKey/index.html, and click on the “Two Quantitative Variables” menu in the “Descriptive Statistics and Graphs” section

1. Select the proper columns to create a scatterplot with “TotBill” as the X variable and “Tip” as the Y variable
2. Identify the correlation coefficient in the “Summary Statistics” table
3. Use the scatterplot and correlation coefficient to describe the relationship between these two variables

Practice (solution)

2. $r = 0.676$
3. There is a strong, positive relationship between the total bill and the amount tipped (as you'd expect)

Misuse of the Correlation Coefficient

- "Correlation" is one of the most commonly misused terms in world of data analysis
  - In this class, you should only use the word "correlation" to describe *linear relationships* between two quantitative variables
  - That means you shouldn't describe two categorical variables as *correlated*, instead you should describe them as *associated*

Misuse of the Correlation Coefficient

- “Correlation” is one of the most commonly misused terms in world of data analysis
  - In this class, you should only use the word “correlation” to describe *linear relationships* between two quantitative variables
  - That means you shouldn’t describe two categorical variables as *correlated*, instead you should describe them as *associated*
- In the slides that follow, I’ll briefly cover a few other common misuses of the correlation coefficient

Non-linear Relationships and Outliers

From Cook & Swayne's *Interactive and Dynamic Graphics for Data Analysis*:

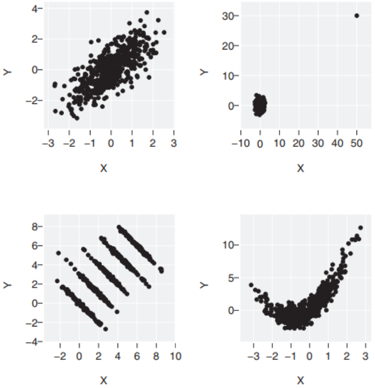
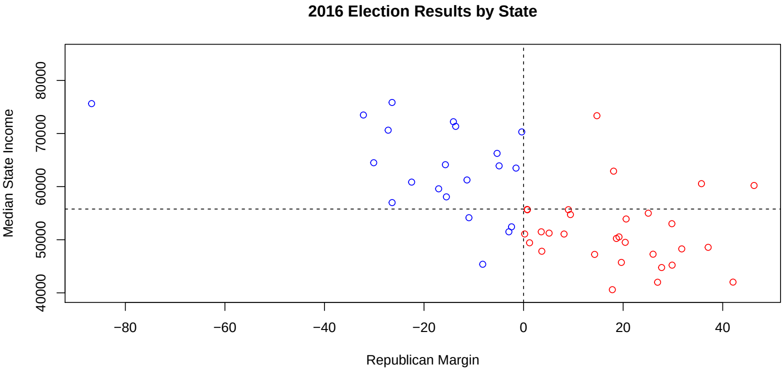


Fig. 6.1. Studying dependence between X and Y. All four pairs of variables have correlation approximately equal to 0.7, but they all have very different patterns. Only the top left plot shows two variables matching a dependence modeled by correlation.

Ecological Correlations

- **Ecological correlations** compare variables at an ecological level (ie: The cases are aggregated data - like countries or states)
  - There's nothing inherently bad about this type of analysis, but the results are often misconstrued
- Let's look at the correlation between a US state's median household income and how that state voted in the 2016 presidential election

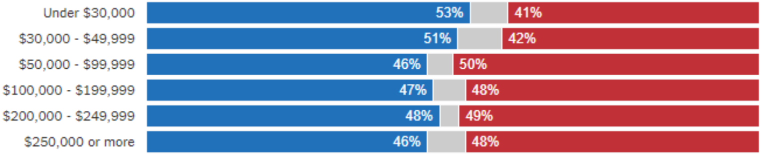
Ecological Correlations



- $r = -.63$, so do republicans earn lower incomes than democrats?

The Ecological Fallacy

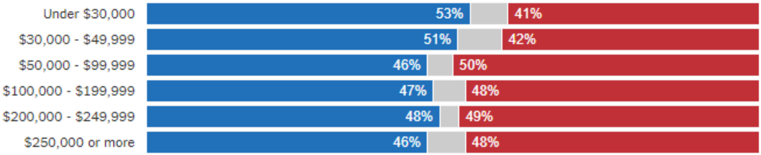
Using 2016 exit polls, conducted by the NY Times (Link), we can get a sense of how party vote and income are related *for individuals*:



- Looking at individuals as cases there is an opposite relationship between political party and income

The Ecological Fallacy

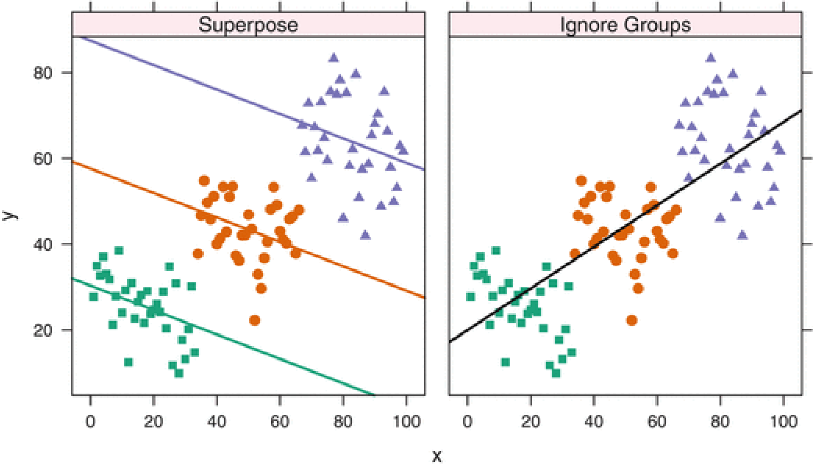
Using 2016 exit polls, conducted by the NY Times (Link), we can get a sense of how party vote and income are related *for individuals*:



- Looking at individuals as cases there is an opposite relationship between political party and income
- This "reversal" is an example of the **ecological fallacy**
  - Inferences about individuals cannot necessarily be deduced from inferences about the groups they belong to
  - The lesson here is we should use data where the cases align with who/what we're aiming to describe

The Ecological Fallacy

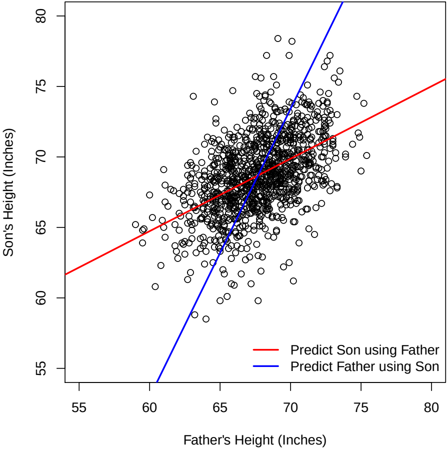
The ecological fallacy is related to ignoring an important grouping variable:



Regression

- The *correlation coefficient* is one method for summarizing the relationship between two quantitative variables
    - Correlation is a **symmetric** statistical method: $r_{x,y} = r_{y,x}$, or it doesn't matter which variable is chosen to be "X" and which is chosen to be "Y"
- Another option is *regression*, which is an **asymmetric** statistical method, meaning the choice of **explanatory** and **response** variables matter

Regression


Son's Height (Inches)
80
75
70
65
60
55
Father's Height (Inches)
55
60
65
70
75
80
Predict Son using Father
Predict Father using Son

Regression Lines

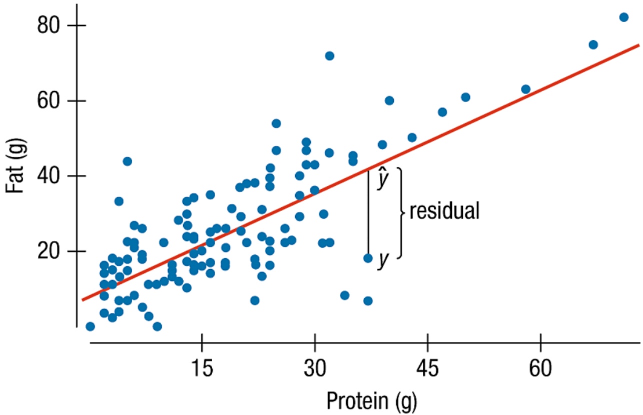
Like any straight line, the regression line relating X and Y is based upon two components, a **slope** and an **intercept**:

$$\hat{Y} = b_0 + b_1 X$$

- in this notation, $\hat{Y}$ is the *predicted value* of the outcome variable
- X is the explanatory variable
- b_0 is the *estimated* intercept, or the predicted value when $X = 0$
- b_1 is the *estimated* slope, or predicted change in the outcome variable for a 1-unit increase in the explanatory variable

Regression Lines

- b_0 and b_1 are estimated from the data such that they minimize the squared **residuals**, or the distances between the predicted to observed outcomes
  - The example below shows the relationship between protein and fat content in Burger King's menu items



Predictions

- The regression line can be used as a predictive tool:

$$\widehat{\text{Fat}} = 8.4 + 0.91 * \text{Protein}$$

- If we wanted an item with 20g of protein, we'd predict it to have $8.4 + 0.91 * 20 = 26.6$ grams of fat

Practice

Using the "Tips" dataset, available by clicking here or on our website, go to https://www.lock5stat.com/StatKey/index.html, and click on the "Two Quantitative Variables" menu in the "Descriptive Statistics and Graphs" section

1. Select the proper columns to create a scatterplot with "TotBill" as the X variable and "Tip" as the Y variable
2. Identify the regression line's intercept and slope in the "Summary Statistics" table
3. What does the *slope* tell you about the relationship between these two variable?
4. What tip does the line predict for a $20 total bill?

Practice (solution)

2. The regression line is: $\widehat{\text{Tip}} = 0.92 + 0.11 * \text{Total Bill}$
3. The slope of 0.11 suggests each \$ increase in the total bill leads to an 11 cent increase in the tip - meaning that people are tipping roughly 11%
4. The predicted tip for a \$20 total bill is given by: $0.92 + 0.11 * 20 = 3.12$

Misuse of Regression

- Much like the correlation coefficient, regression is also very commonly misused
    - The slides that follow will cover a few common misuses

Switching the Explantory and Response Variables

- As mentioned earlier, the choice of explanatory and response variable matters in regression (recall that it doesn't for correlation, which is symmetric)

Switching the Explantory and Response Variables

- As mentioned earlier, the choice of explanatory and response variable matters in regression (recall that it doesn't for correlation, which is symmetric)
- In our Burger King menu example, if we used protein to predict fat: $\widehat{\text{Fat}} = 8.4 + 0.91 * \text{Protein}$
  - A meal with 20g protein is predicted to have 26.6g of fat
- But if we used fat to predict protein: $\widehat{\text{Protein}} = 2.3 + 0.62 * \text{Fat}$
  - A meal with 26.6g of fat is predicted to have 18.8g of protein

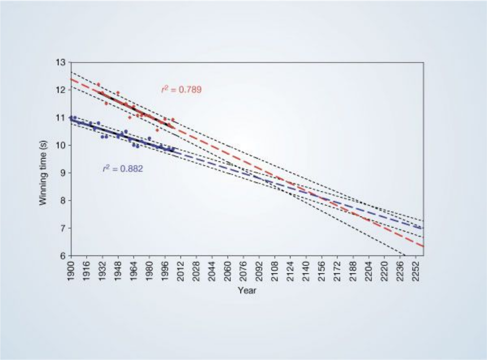
Extrapolation

In 2004, an article was published in *Nature* titled "Momentous sprint at the 2156 Olympics". The authors plotted the winning times of the men's and women's 100m dash in every Olympics, fitting separate regression lines to each. They found that the lines will intersect at the 2156 Olympics, here are a few media headlines:

- "Women 'may outsprint men by 2156' " - BBC News
- "Data Trends Suggest Women will Outrun Men in 2156" - Scientific American
- "Women athletes will one day out-sprint men" - The Telegraph
- "Why women could be faster than men within 150 years" - The Guardian

Extrapolation

Here is a figure from the original publication in Nature:

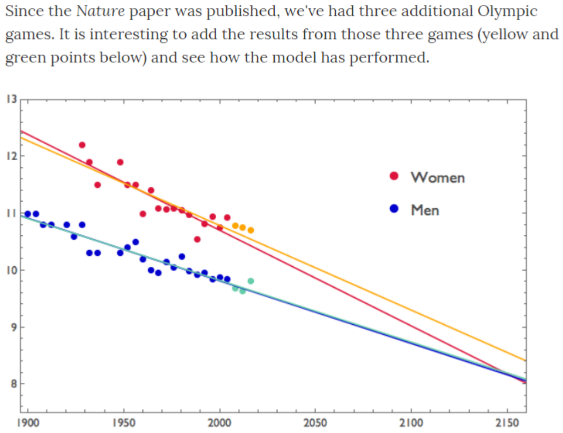


The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

Do you have any problems with the headlines on previous slide?

Extrapolation

It is important not to predict beyond the observed range of your explanatory variable, your data tells you nothing about what is happening outside of its range!



Since the *Nature* paper was published, we've had three additional Olympic games. It is interesting to add the results from those three games (yellow and green points below) and see how the model has performed.

source: https://callingbullshit.org/case_studies/case_study_gender_gap_running.html

Closing Remarks

We've now learned how to find and summarize relationships between various combinations of variables:

- Two categorical variables: Contingency tables and conditional proportions
- One categorical and one quantitative variable: side-by-side graphs and differences in means/medians
- Two quantitative variables: scatterplots, correlation, and regression