Correlation and Regression

Ryan Miller

Pearson's Height Data

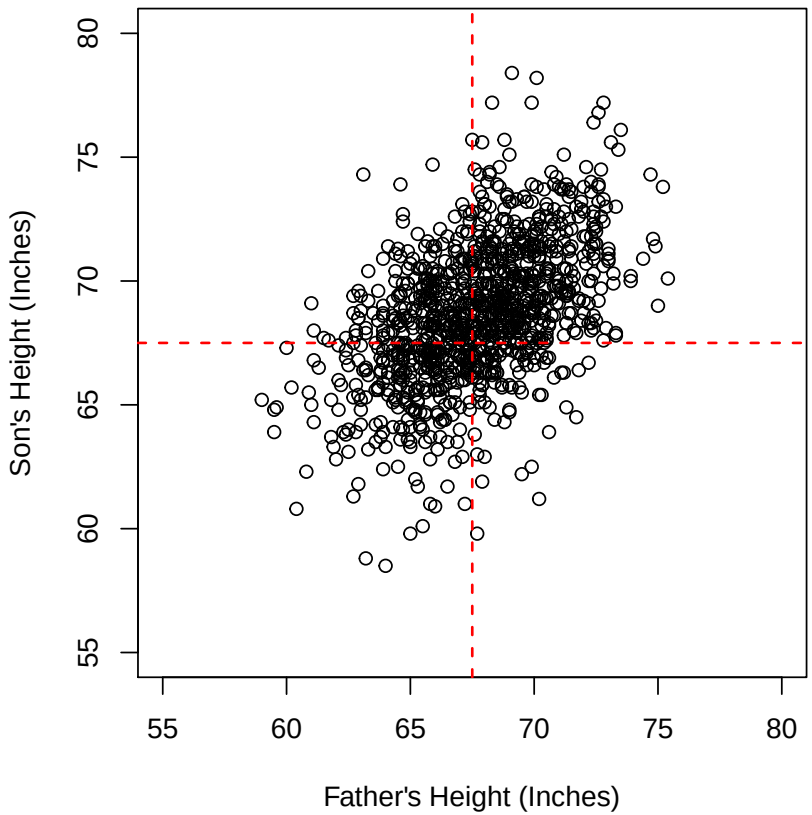
- Francis Galton and Karl Pearson, two pioneers of modern statistics, lived in Victorian England at a time when the scientific community was fascinated by the idea of quantifying hereditable traits
- Wondering if height is hereditable, they measured the heights of 1,078 fathers and their (fully grown)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

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

- Which descriptive statistics or graphs would you use to understand the association between these variables?

Pearson's Height Data

Using a scatterplot the association is obvious:



The correlation coefficient ($r = 0.5$) supports this assessment, though we'll approach this question differently soon

How Strong is it?

Whether a correlation is considered "strong" 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/

Linearity

- The correlation coefficient measures the strength of a *linear association*
  - It is poorly suited for describing non-linear relationships

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

Non-linear Relationships

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

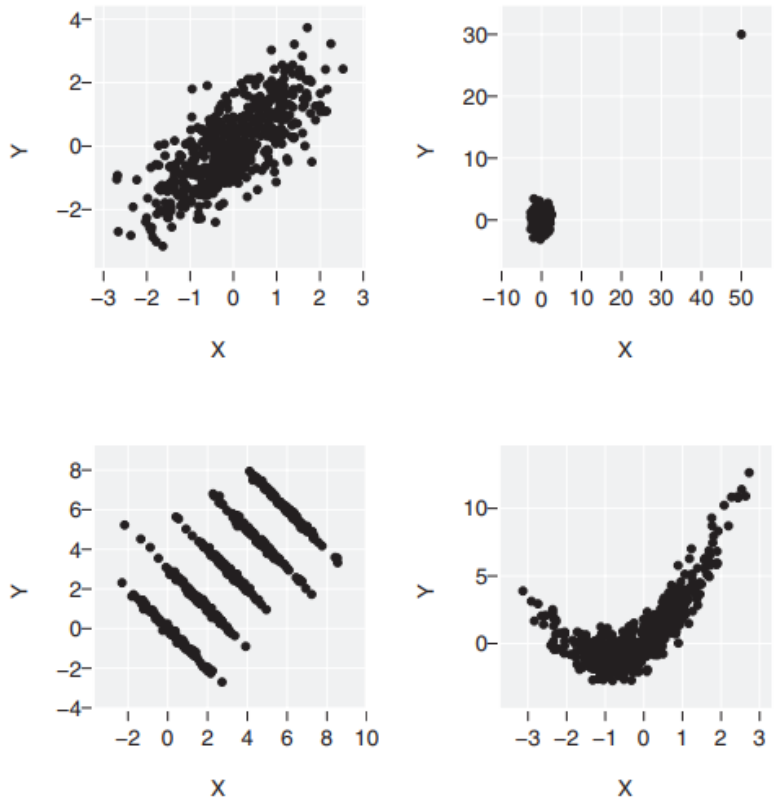
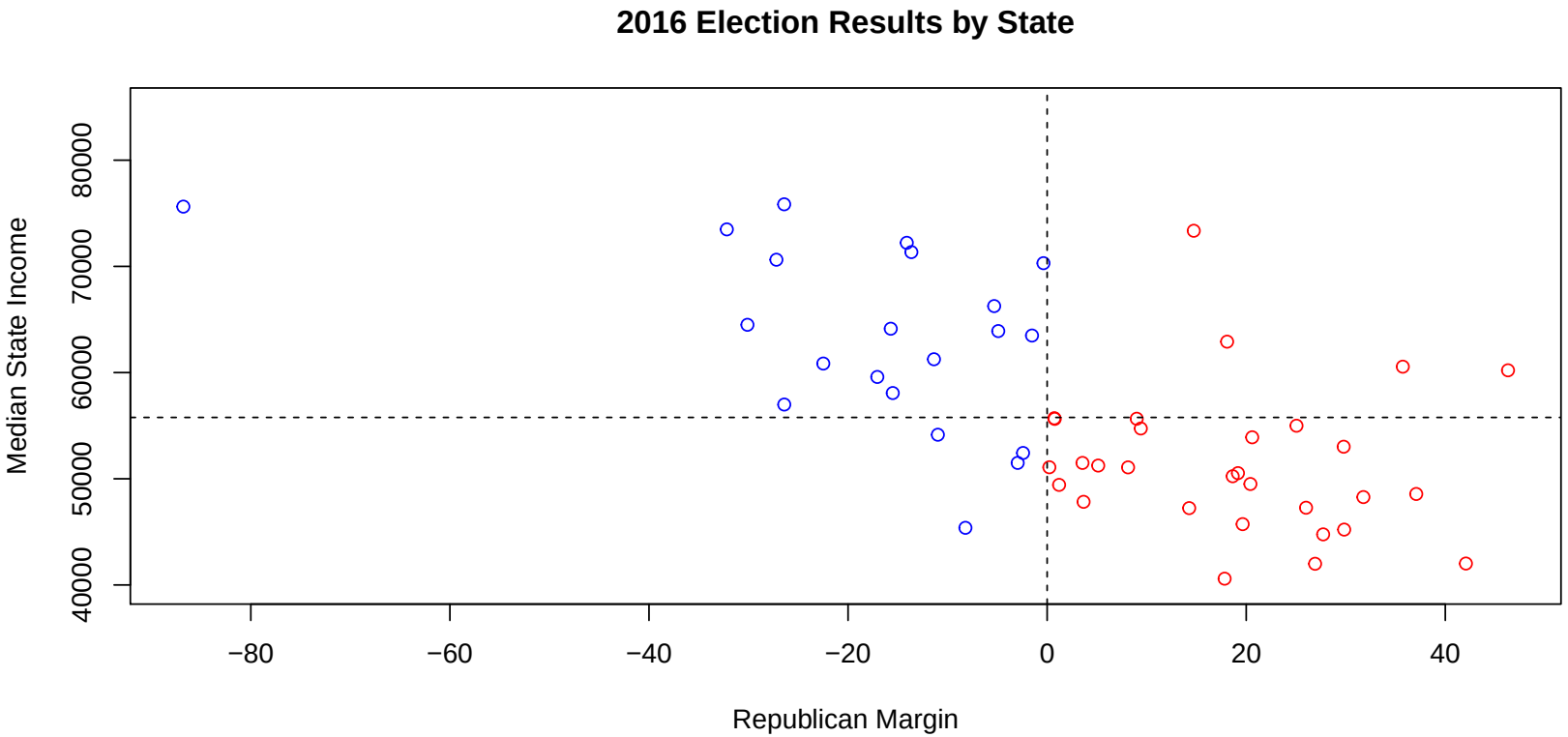


Fig. 6.1. Studying dependence between X and Y. All four pairs of variables have correlation approximately equal to 0.7, but they all have very different patterns. Only the top left plot shows two variables matching a dependence modeled by correlation.

Ecological Correlations

- **Ecological correlations** compare variables at an ecological level (ie: The cases are aggregated data - like countries or states)
- Let's look at the correlation between a US state's median household income and how that state voted in the 2016 presidential election

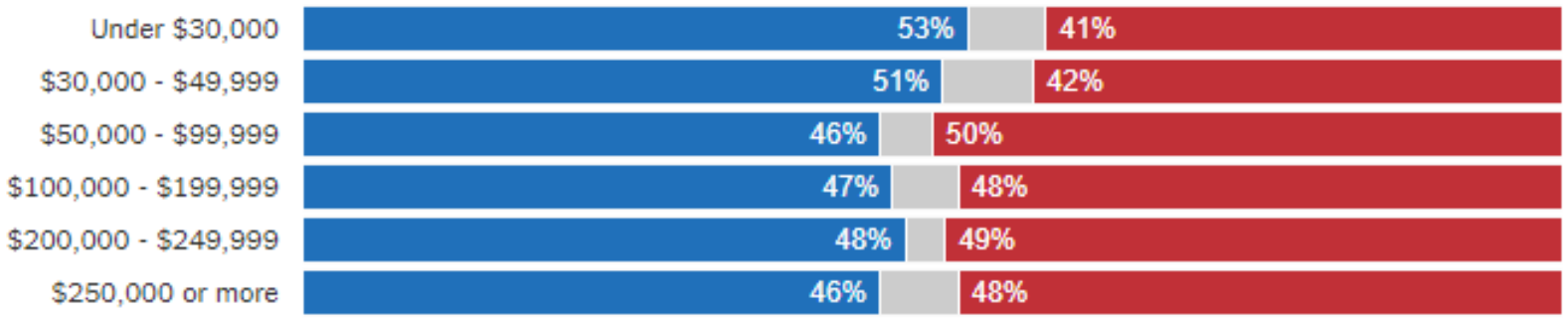
Ecological Correlations



- $r = -.63$, 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
- This "reversal" is an example of the **ecological fallacy**
  - Inferences about individuals cannot necessarily be deduced from inferences about the groups they belong to

Correlation and Predictions

- Suppose we want to use Galton and Pearson's data to make predictions
- What would predict for the height future son for a father who is 67.7 inches tall? (Recall that the average heights were 67.7 inches for fathers and 68.7 inches for sons)
- Since the father is average height, your best prediction is that the son is average height, or 68.7 inches tall

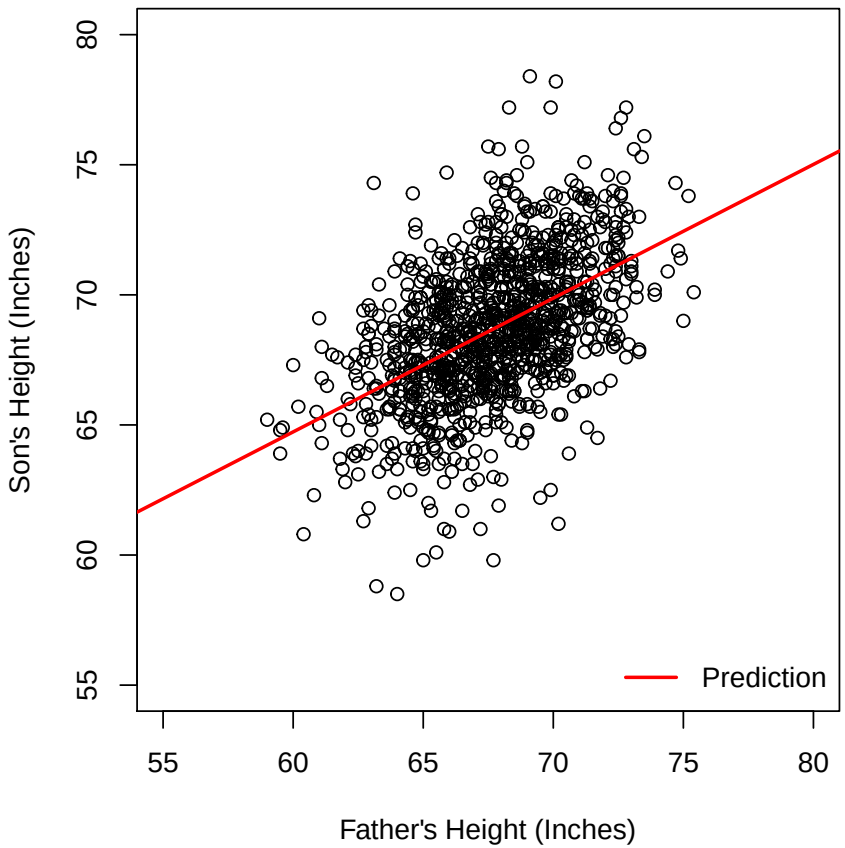
Correlation and Predictions

- How would you predict the son's height if the father were 65.0 inches, or 2.7 inches below the average?
- You'd be wise to predict a below average height for the son, but by how much exactly?
  - One way method is to use correlation coefficient and a principle known called "regression"

1. Standardize the explanatory variable (In this example $z_f = -1$)
2. Use the correlation coefficient to predict how much "regression" occurs (ie: $z_s = z_f * r = -1 * .5$)
3. Unstandardize the prediction to get an answer in the original units (ie: predicted son's height $= \bar{y} + z_s * s_s$)

Using Correlation to make Predictions

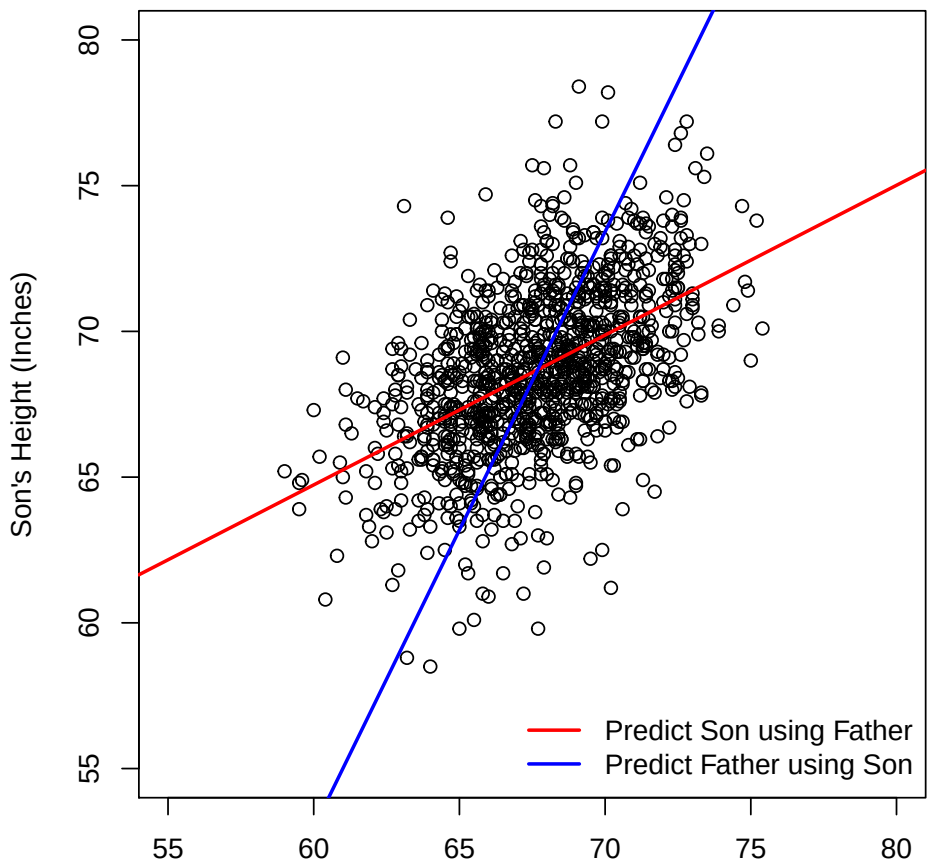
This procedure can make a prediction for *any* father's height:



These predictions form the **Regression Line**

Two Regression Lines

- Regression is an **asymmetric** statistical method: the choice of explanatory and response variables matters
- Correlation is a **symmetric** statistical method: $r_{x,y} = r_{y,x}$



Using a Regression Line to make Predictions

The regression line for the Pearson/Galton Data had the form:

$$\widehat{\text{Son's Height}} = 33.9 + 0.51 * \text{Father's Height}$$

- Using this line, we can predict the Son's Height for a given Father's Height simply by plugging that Father's Height into an equation
- We can also use the regression line as descriptive tool
  - For each 1 inch increase in father's height, we expect a 0.51 increase in son's height
  - The intercept isn't meaningful in this example (but sometimes it can be)

How Regression got its Name

- The correlation coefficient relating two variables is always less than 1 (in absolute value)
- For a 1 standard deviation increase in the explanatory variable, regression will always predict the response variable increases by *less than* 1 standard deviation
- Galton described this phenomenon as: "regression to mediocrity"

The Madden Curse

Article Link: "Is the 'Madden' cover curse still a thing? A look back at 20 years of NFL stars offers a verdict"

- Madden is an iconic videogame whose cover features a different NFL player each year, usually a player who performed exceptionally well in the previous season
- Frequently, the player featured on the Madden cover suffers from a decline in play or sustains an injury in their next season (see the article)
- Is the "Madden Curse" real? What might be a more statistically sound explanation?

Regression to Mediocrity

- Each player featured on the Madden cover was selected because they had exceptional season
- Performance in the subsequent season is correlated with that of the prior season, but the correlation is nowhere near 1
- The best prediction is for these players to regress
- The NFL is such that seasons near the league's statistical averages are not generally regarded as "good"
  - In 2017, the 16th rated passer was Tyrod Taylor, with 2799 yds, 14 tds, 4 ints
  - The 16th rusher was Lamar Miller with 888 yds, 3 tds

Extrapolation

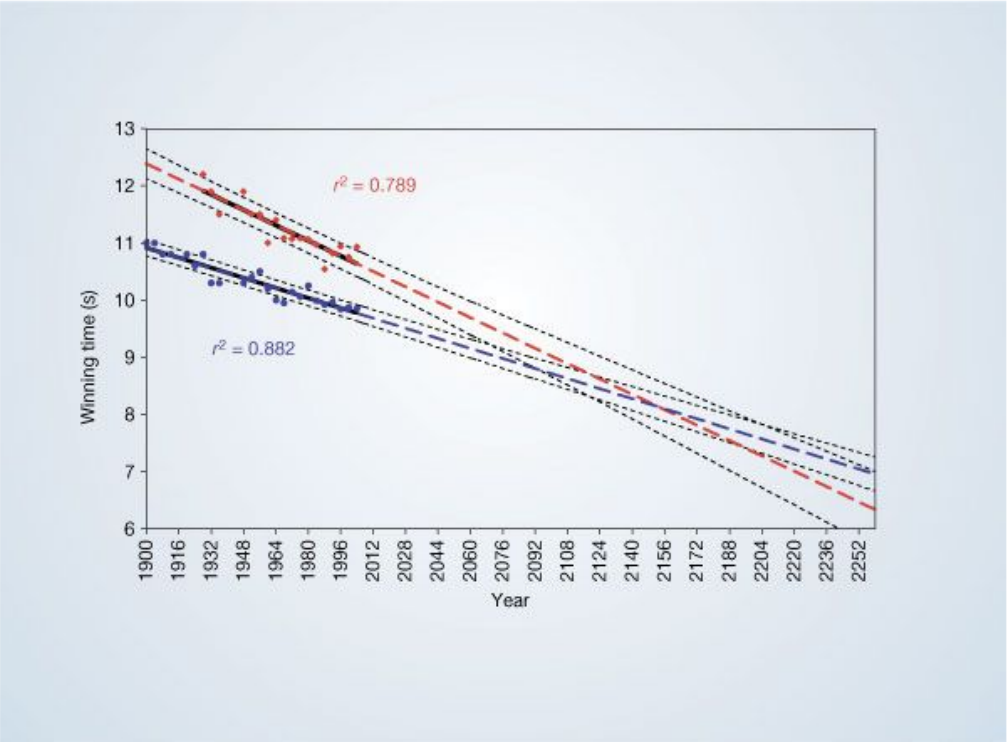
In 2004, an article was published in *Nature* titled "Momentous sprint at the 2156 Olympics". The authors plotted the winning times of the men's and women's 100m dash in every Olympics, fitting separate regression lines to each. They found that the lines will intersect at the 2156 Olympics, here are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

Do you have any problems with these conclusions?

Extrapolation

Here is a figure from the original publication in Nature:

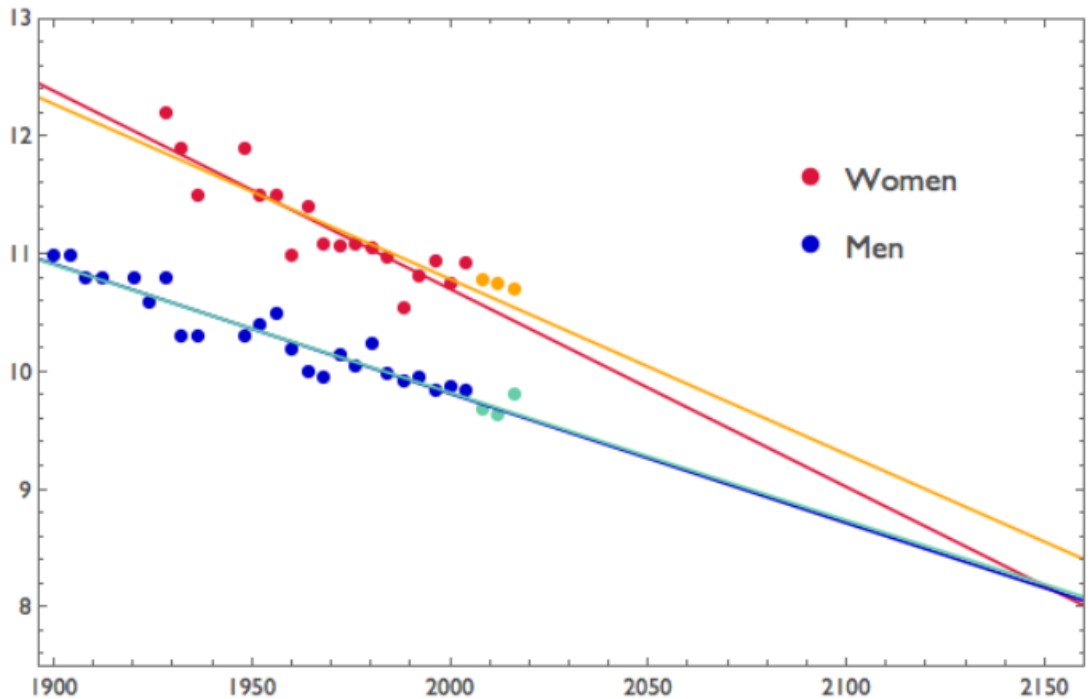


The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of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

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

Correlation and Regression Takeaways

- The correlation coefficient is a *symmetric* summary measure that describes the relationship between two quantitative variables
- The correlation coefficient only captures a *linear relationship*
- Beware of conclusions that are based upon ecological correlations
- Regression is an *asymmetric* approach to describing the relationship between two quantitative variables
- Avoid using regression to make judgements beyond the range of your data

The Next Steps

- We will revisit the topic of regression when learning about *statistical models* later in the semester
- For now, consider regression to be a descriptive tool (we need to learn about *statistical inference* before we're ready for statistical modeling)

Conclusion

Right now you should...

1. Know how to interpret a correlation coefficient and how to avoid some common misuses
2. Understand regression, how it is similar/different from correlation, and how it can be misused

These notes cover Section 2.5 and Section 2.6 of the textbook, I encourage you to read through the section and its examples