Z-Scores, Correlation, and Regression

Ryan Miller

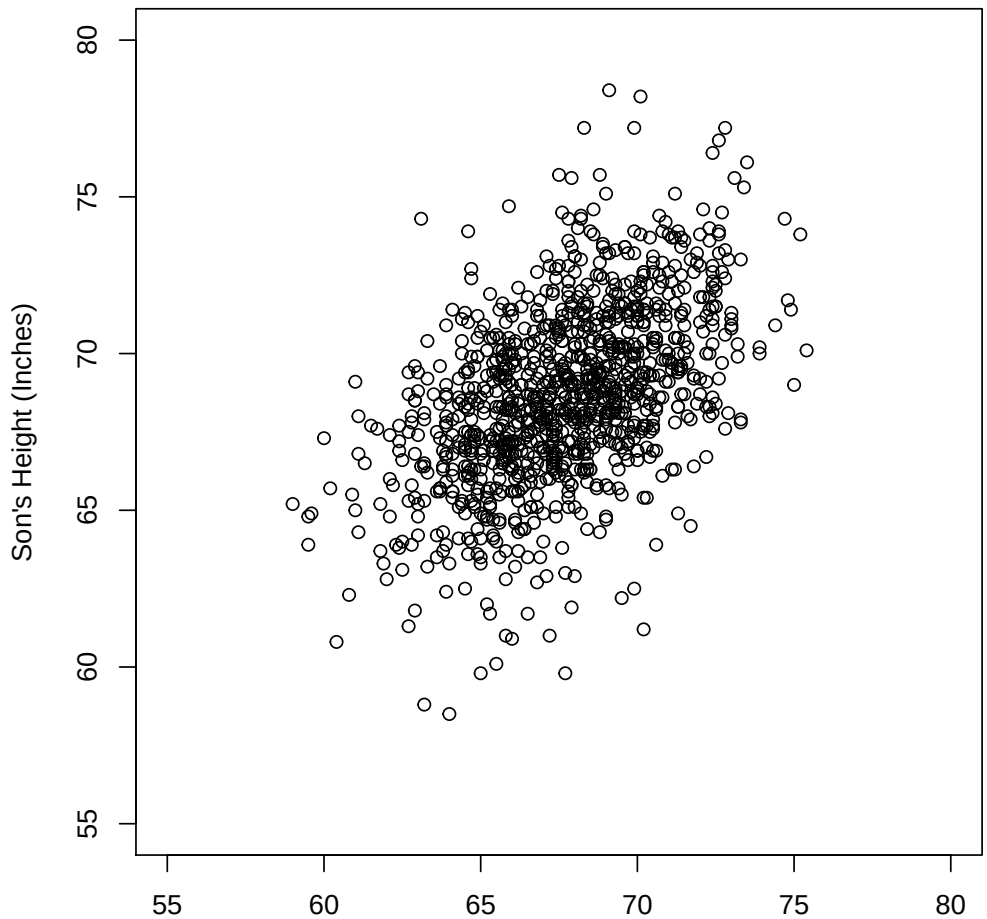
Pearson's Height Data

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

Father	Son
65	59.8
63.3	63.2
65	63.3
65.8	62.8
61.1	64.3
63	64.2
...	...

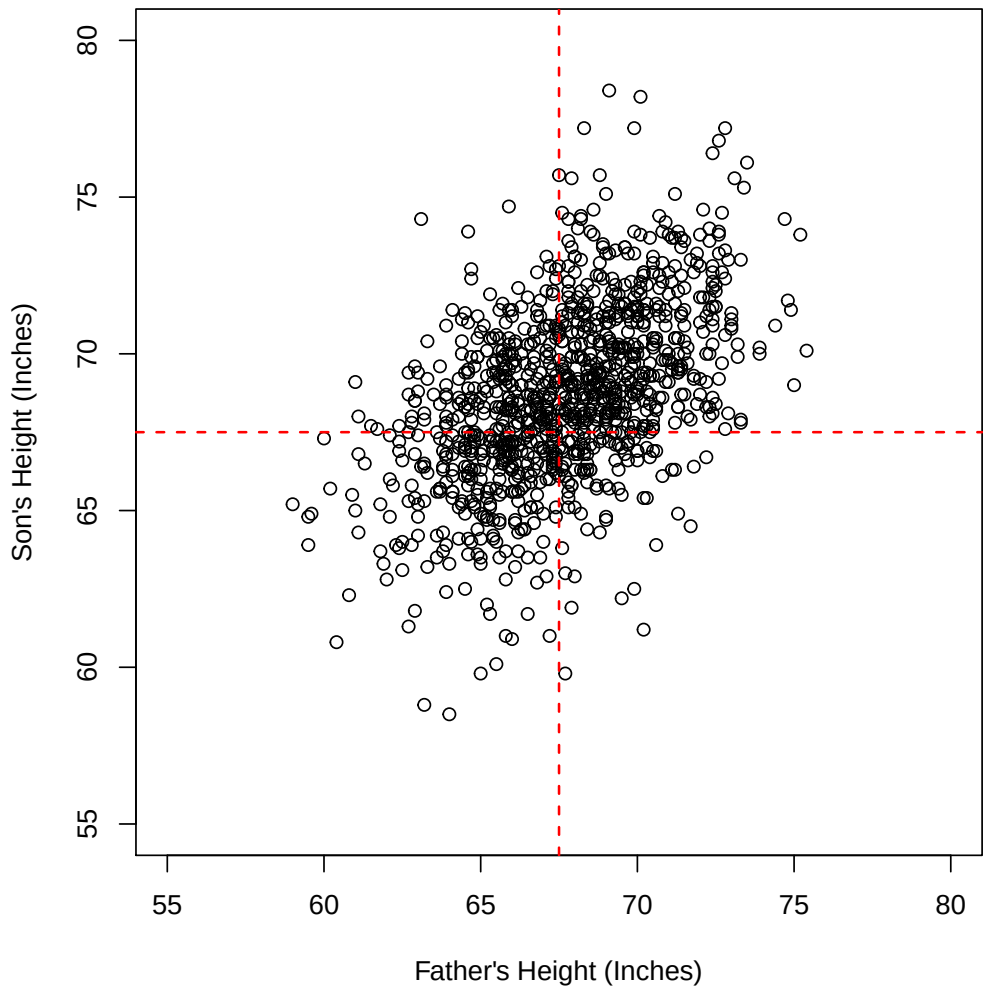
Scatterplots

The best way to visualize the two quantitative variables contained in Galton and Pearson's data is with a **scatter plot**... is height hereditable?



Scatterplots

If we divide the scatter plot into quadrants, the heritability of height is obvious:



Standardization

Tall fathers tend to have tall sons, although there are plenty of exceptions. But exactly how strong is this association?

To summarize an association between two quantitative variables we need to consider:

1. A "large" value for each variable depends upon that variable's mean
2. Both variables are on different scales with potentially different units

Let's see how to account for these using Pearson's height data

Standardization

We need to put both variables on an equal playing field, let's focus on the first case in Pearson's data, here the father is 65.0 inches tall

- The average height of fathers in the sample is 67.7 inches
  - We could describe the first father as 2.7 inches below average
- The standard deviation of fathers in the sample is 2.8 inches
  - So another way of describing the first father is to say that his height is about 1 standard deviation below the average

Standardization

Formalizing the previous example, we can **standardize** the i^{th} case's value for a variable x using the following calculation:

$$z_i = \frac{x_i - \bar{x}}{s_x}$$

Here z_i is the **z-score** for the i^{th} case, $\bar{x}$ is the sample mean, and s_x is the sample standard deviation of x

If we were working with an entire population and knew μ and σ, we would use them in place of $\bar{x}$ and s

Why Standardization is Useful

- Suppose you're told that the concentration of urea in your blood is 50 mg/dl above average, what do you conclude?
- Suppose you're told that the concentration of urea in your blood is 4 standard deviations above average, what do you conclude?

Standardization allows non-experts to better interpret quantitative variables!

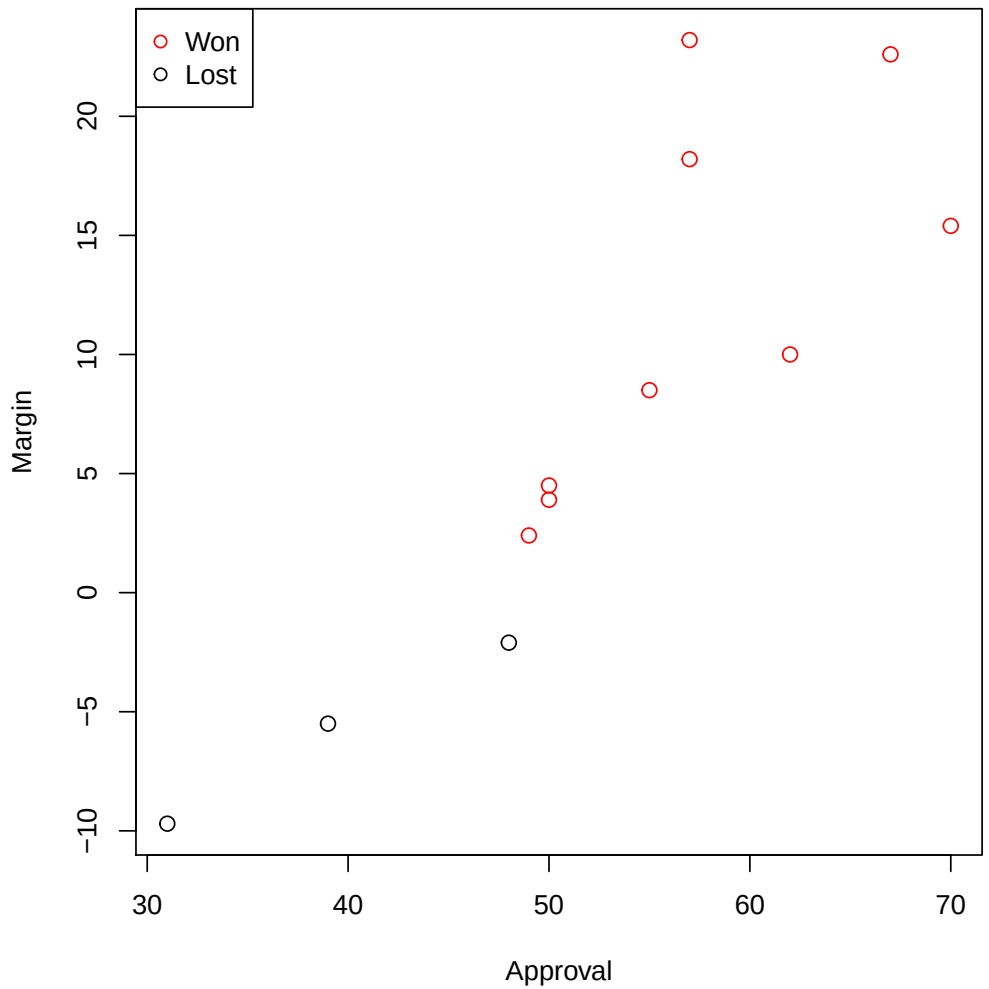
Practice

- Load the "ElectionMargin" data into Minitab
- These data cover all US presidential elections since 1940 where an incumbent was running for re-election and include:
  - the year
  - the candidate
  - the candidate's approval rating at the time of the election
  - the candidate's margin of victory or defeat
  - the election result

- Create a scatterplot visualizing the association between approval rating and margin of victory/defeat
  - Do you see a relationship? How strong is it?
  - What approval rating appears necessary to win re-election?
- George W. Bush won re-election with a 49% approval rating. Calculate and interpret a z-score to compare Bush's approval rating with other all other incumbents. Repeat this comparison for all incumbents who won re-election

Practice - Solutions

There appears to be a strong relationship, an approval rating of close to 50% seems necessary for re-election



Practice - Solutions

1. $\bar{x} = 52.9$, $s = 11$, $z_{\text{GWB}} = \frac{49-52.9}{11} = -0.35$
2. $\bar{x}_{re} = 57.3$, $s_{re} = 7.6$, $z_{\text{GWB}} = \frac{49-57.3}{7.6} = -1.1$

GW Bush's approval was 0.35 standard deviations below the average incumbent. His approval was 1.1 standard deviations below the average incumbent who won re-election.

Correlation

- The summary statistic measuring the strength and direction of linear association between two quantitative variables is called the **correlation coefficient** (sometimes called Pearson's correlation coefficient)
- We denote sample correlation using r and population correlation using ρ (the Greek letter 'rho')
- Sometimes we include subscripts that indicate the variables we are comparing:

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

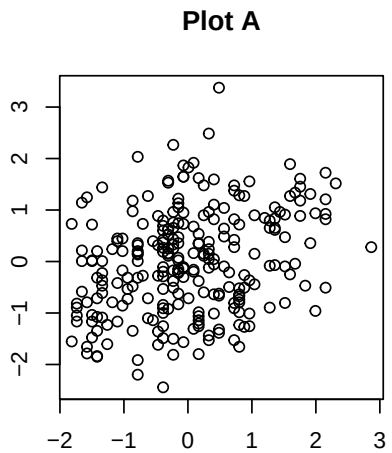
- Correlation is essentially just *the average product of each pair of z-scores*

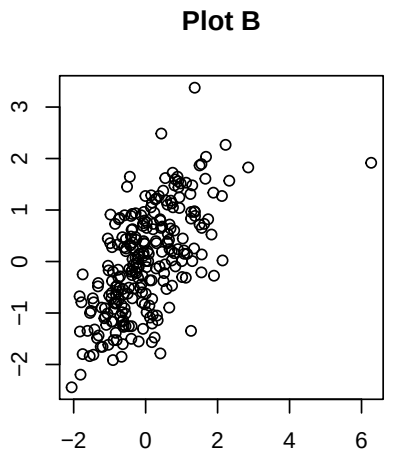
Note: $n - 1$ is used here rather than n in order to cancel with the denominators of s_x and s_y

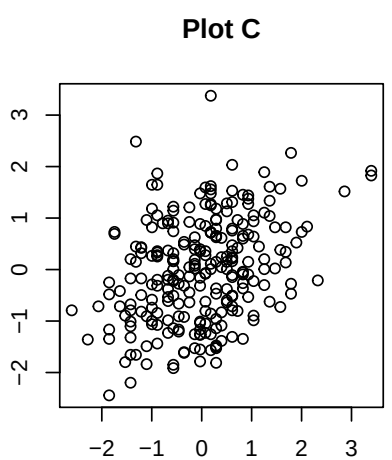
Correlation - Bodyfat Percentage for Men

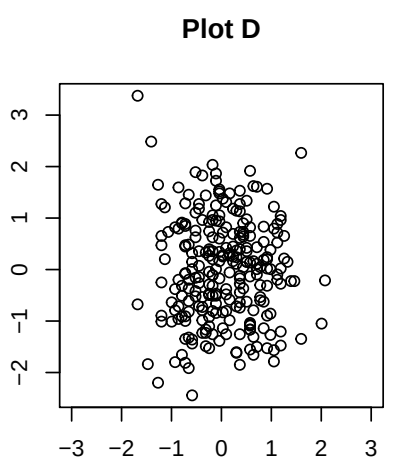
Correlations: 0.61, -0.09, 0.29, 0.35

Y = bodyfat percent, X = Age, Height, Weight, Wrist Circumference

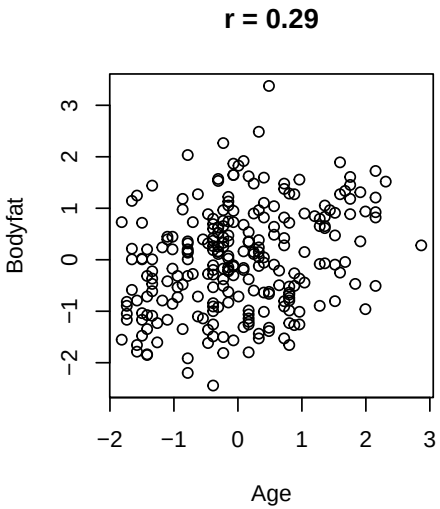


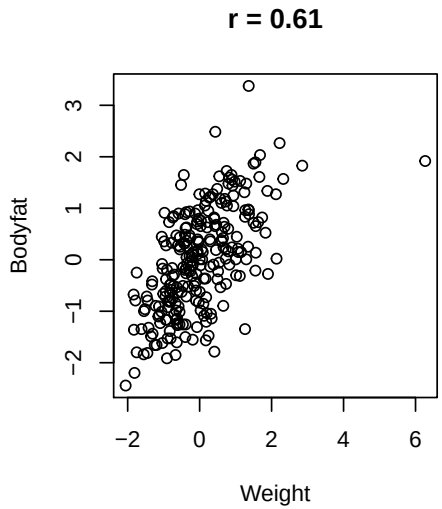


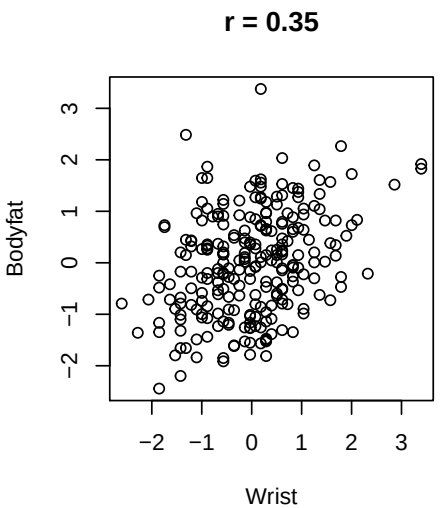


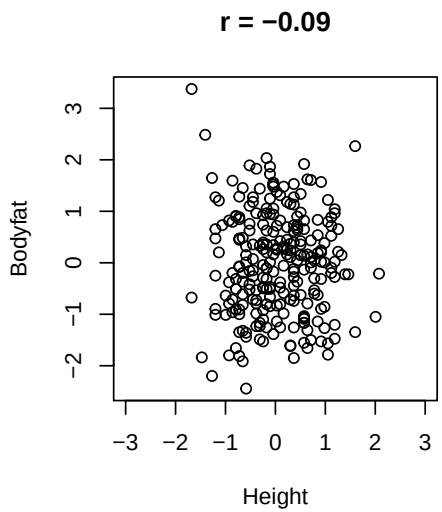


Correlation Examples - Solutions









Correlations can be Misleading!

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

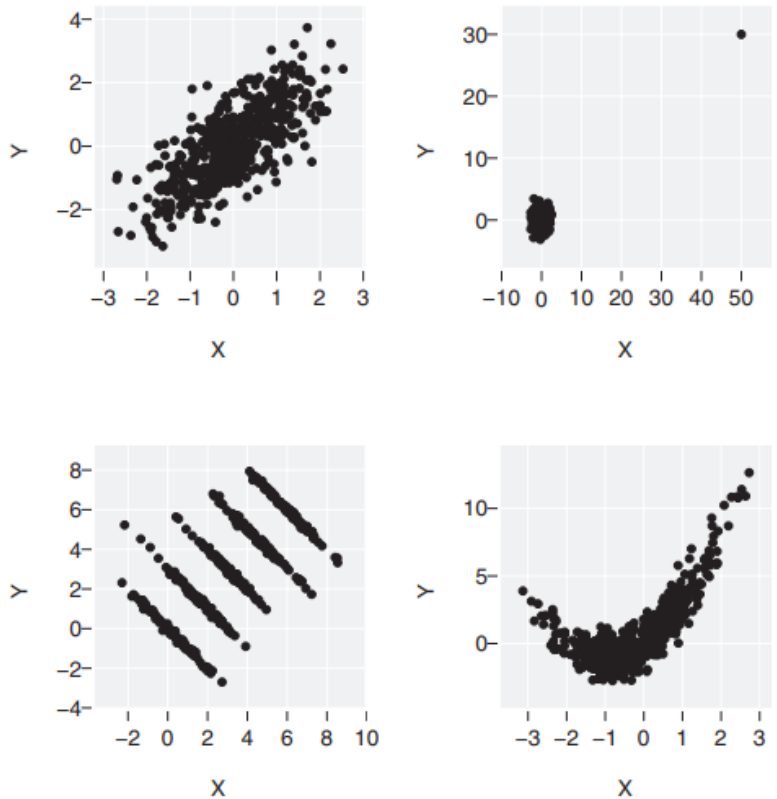
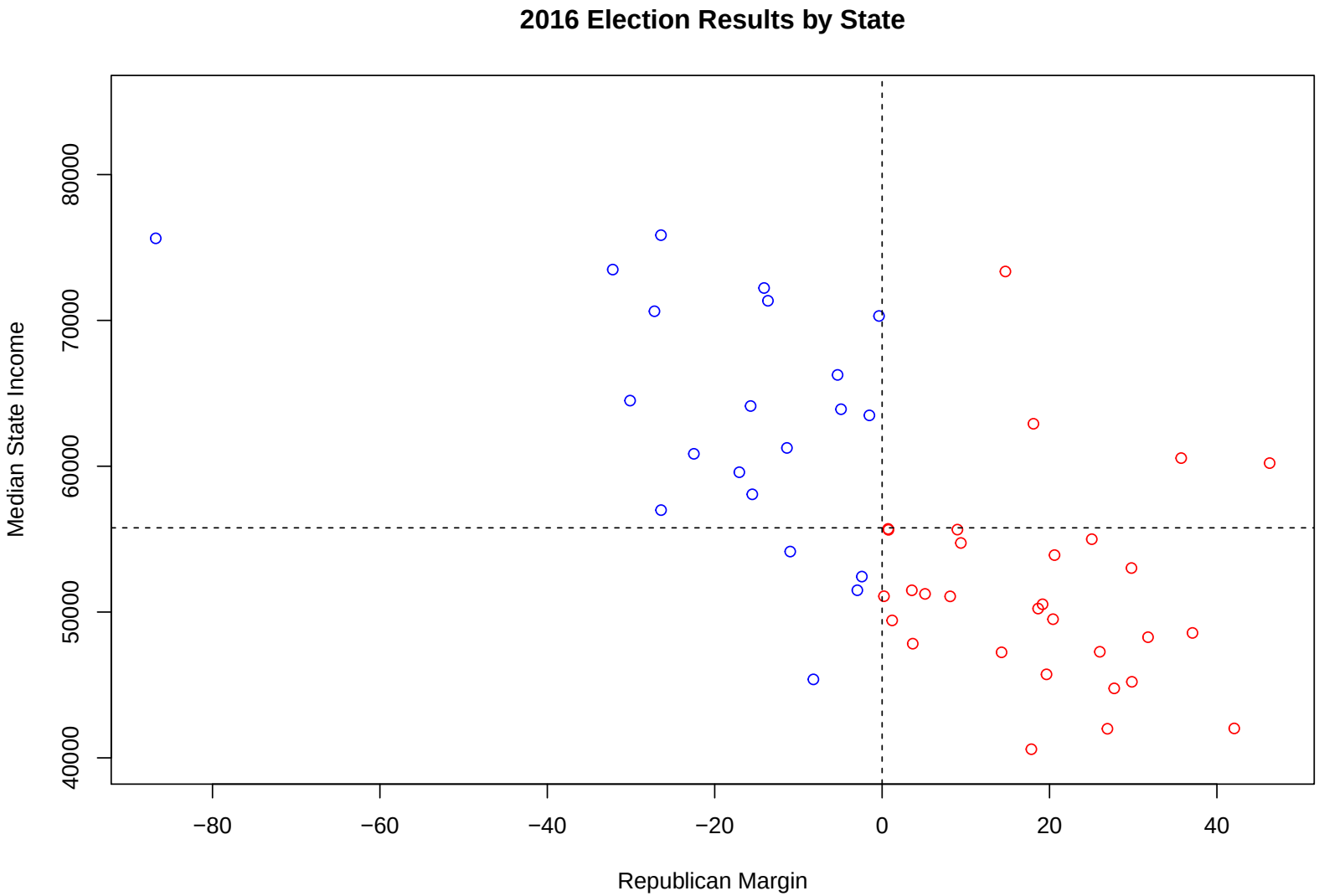


Fig. 6.1. Studying dependence between X and Y. All four pairs of variables have correlation approximately equal to 0.7, but they all have very different patterns. Only the top left plot shows two variables matching a dependence modeled by correlation.

Ecological Correlation

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

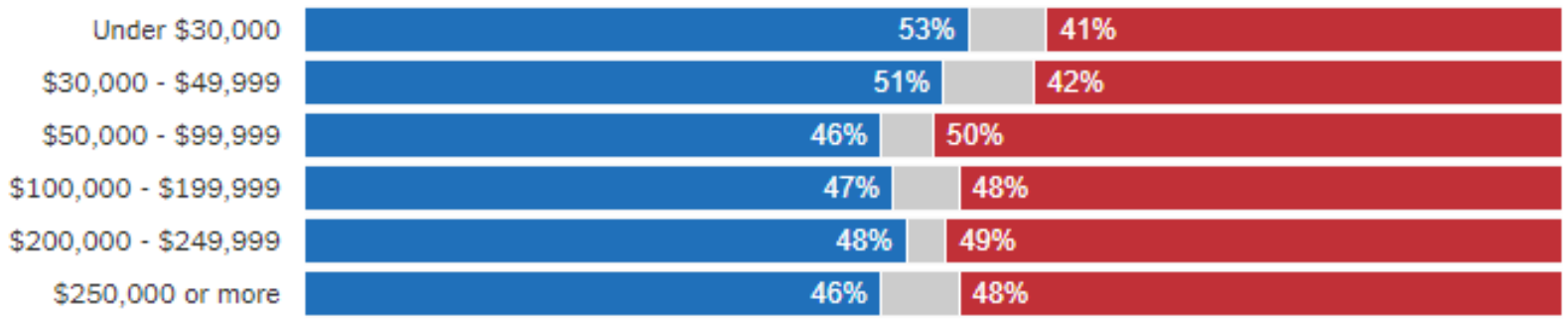
Ecological Correlations



$r = -.63$, a strong relationship where states that are more Republican tend to have lower median incomes

Ecological Correlations

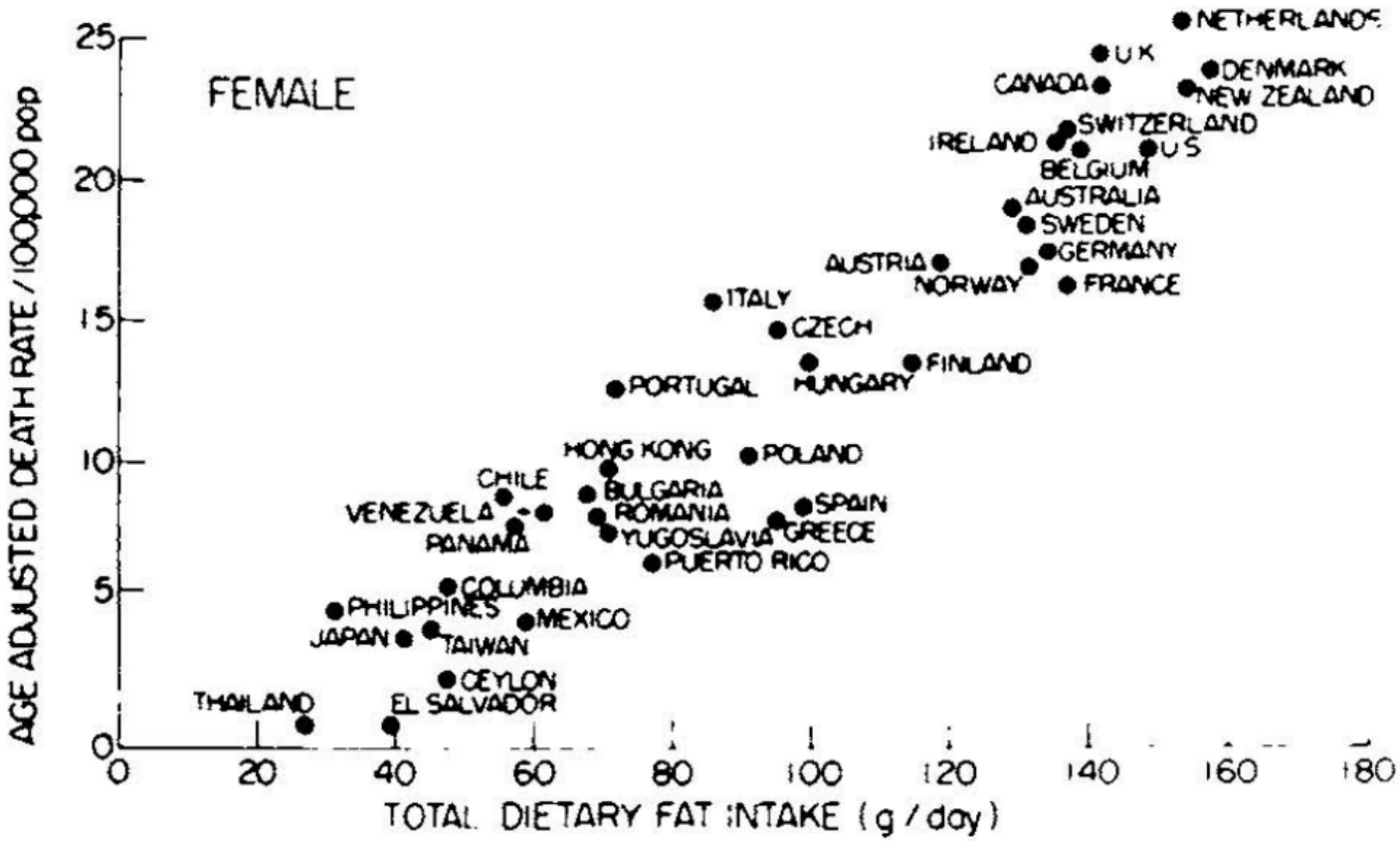
From 2016 exit polls, conducted by the NY Times (Link), we can get a sense of how party vote and income are related for individuals:



- Looking at individuals as cases, we see the opposite relationship between political party and income
- The ecological fallacy can be particular harmful in medical settings - we are interested in the health of individuals, not states/countries/counties

Ecological Correlations

From an article by Carroll in *Cancer Research* (1975):



- Studying aggregated cases (like countries or states) is okay if you're trying to make decisions at that level, it is *not* okay if you want to make decisions at the individual level

Correlation is not Causation

At this point we should re-emphasize the now famous phrase:

Correlation is not causation

In fact, if you search hard enough you can find numerous variables that are highly correlated but very clearly have no causal relationship:

Link: Spurious Correlations

Using Correlation to make Predictions

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

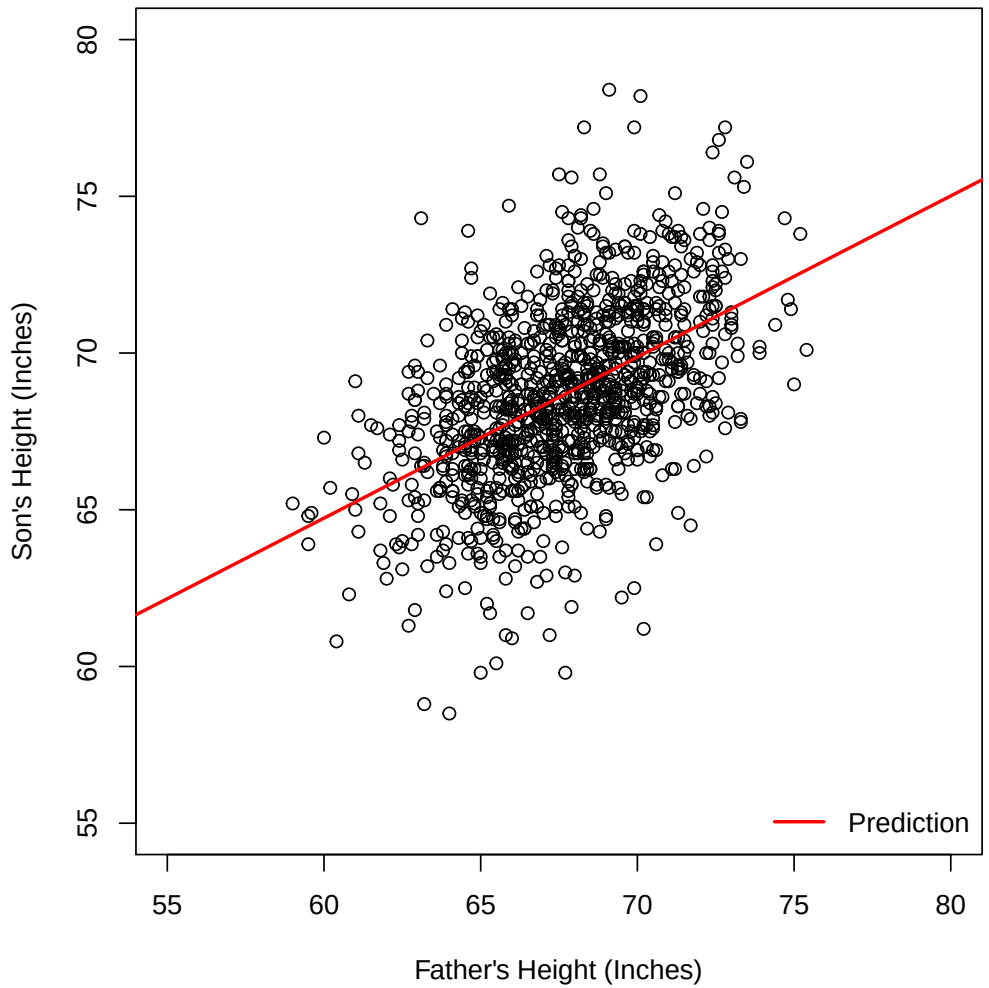
Using Correlation to make Predictions

- 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?
- Part of the answer is *standardization*
  - 65.0 inches is exactly 1 standard deviation below the average for father's height
  - But we know that father's height and son's height aren't perfectly correlated, so we shouldn't predict the son will be exactly 1 standard deviation below average

It turns out that because of the correlation of 0.50 between father/son, the "best" prediction we can make is that the son will be $0.5 * 1$ standard deviations between below average height. The standard deviation for son's heights is 2.8, so we should predict this son will be 67.3 inches tall

Using Correlation to make Predictions

We can use this approach to obtain a prediction for *any* father's height:



Making Predictions

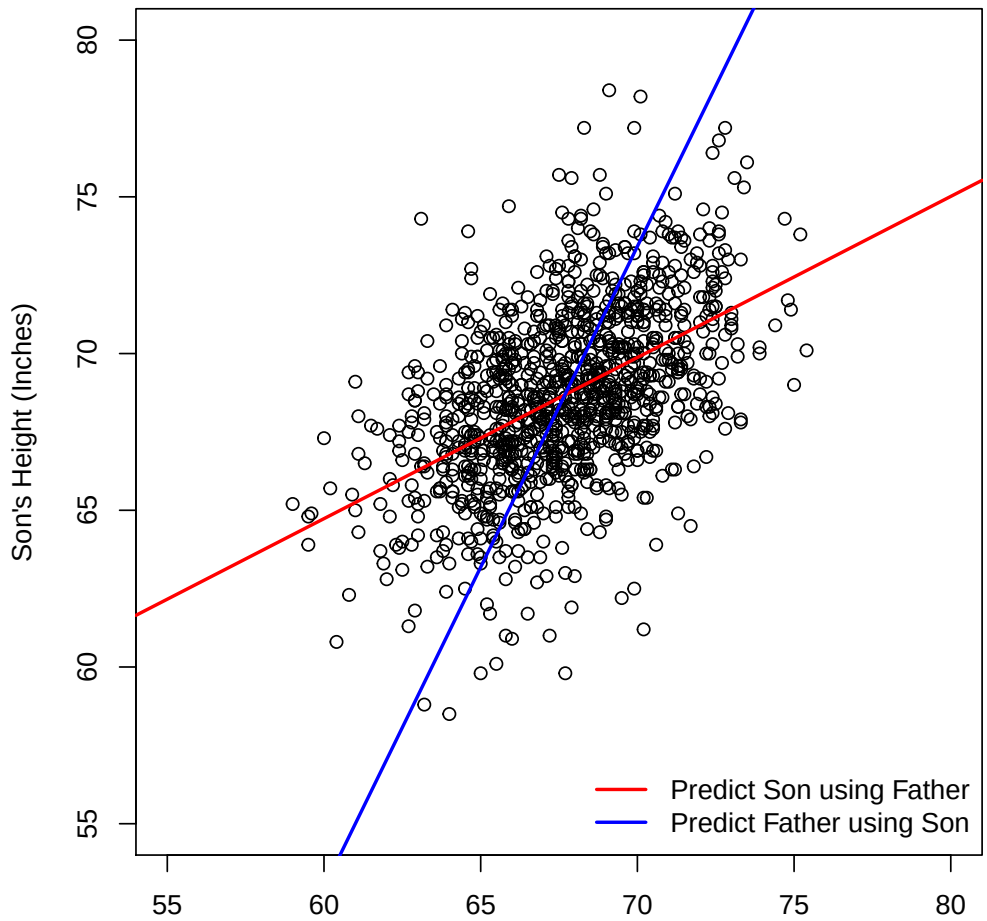
Let's practice by using son's height (Y) to predict father's height (X), some relevant from the data are listed below:

- $\bar{y} = 68.7$
- $\bar{x} = 67.7$
- $s_y = 2.8$
- $s_x = 2.7$
- $r = .5$

Predict the father's height for a son who is 67.3 inches tall

Two Regression Lines

For a father who is 65 inches, we predicted the son would be 67.3 inches tall. But for a son who is 67.3 inches tall, we predicted that the father is 67 inches tall?



Explanatory and Response Variables

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

In general, regression lines have the form:

$$\widehat{\text{response}} = a + b * \text{explanatory}$$

The predicted value of the response variable is a linear function of the explanatory variable

Using a Regression Line to make Predictions

The regression line for the Galton Data had the form:

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

Using this line, we can the predict Son's Height for a given Father's Height simply by plugging the Father's Height in to this equation

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 no where 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

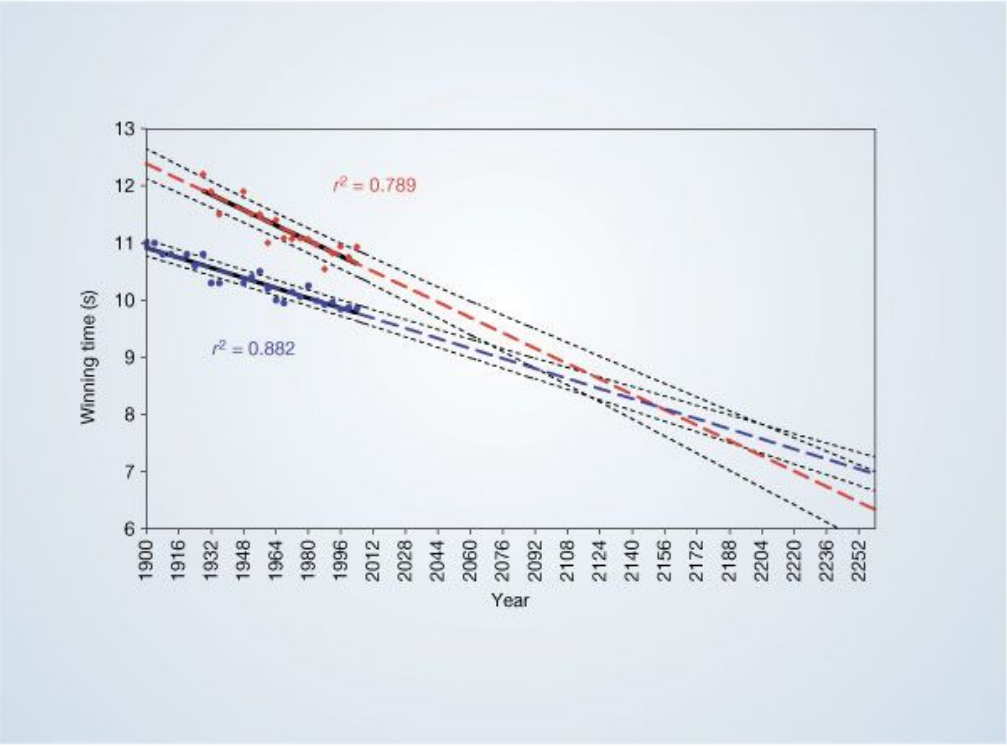
In 2004, an article was published in *Nature* titled "Momentous sprint at the 2156 Olympics". The authors plotted the winning times of the men's and women's 100m dash in every Olympics, fitting separate regression lines to each. They found that the lines will intersect at the 2156 Olympics, here are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

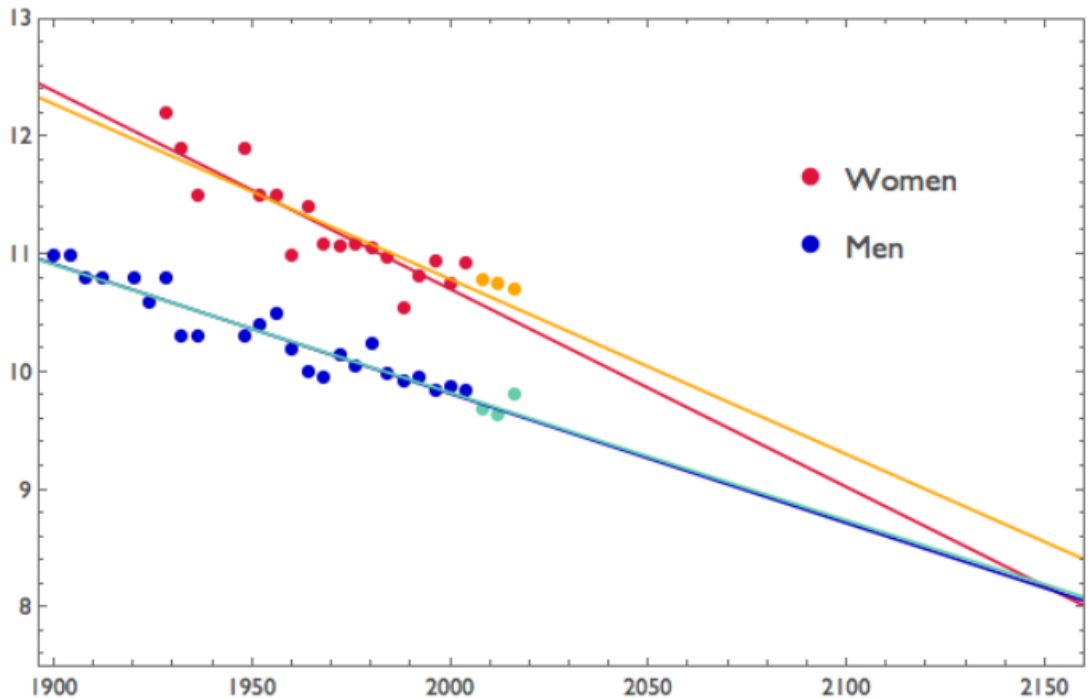


The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of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



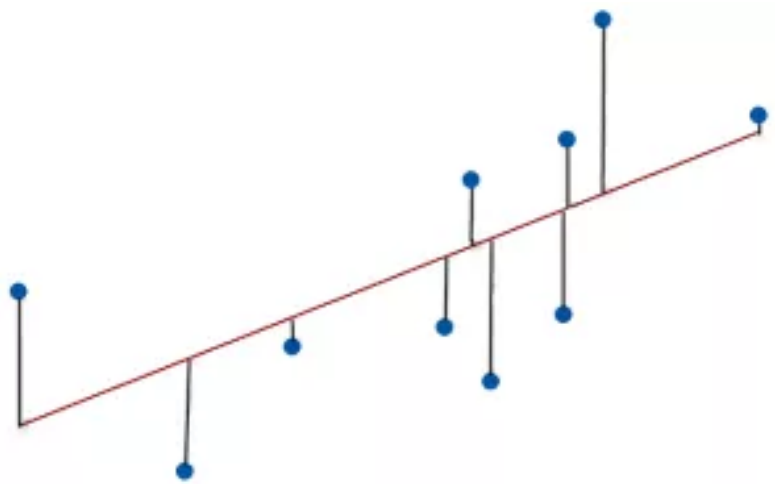
source: https://callingbullshit.org/case_studies/case_study_gender_gap_running.html

Residuals

- Suppose we use a regression to predict the response value for a case already in our dataset
- The difference between the predicted value and the actual observed value for that case is a **residual**

$$\text{Residual for case } i = \text{Observed} - \text{Predicted} = y_i - \hat{y}_i$$

- On a scatter plot, residuals are the vertical deviation from the observed data to the regression line



Least Squares

Until now I've been rather vague about how the regression line provides the "best" predictions. . .

- The regression is found by *minimizing the sum of the squared residuals*

$$\text{Minimize} : \sum_i (y_i - \hat{y}_i)^2$$

The regression line is the "best" in the sense that no other line can provide predictions resulting in smaller squared residuals

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 predictions beyond the range of your data

Conclusion

Right now you should...

1. Be able to calculate a z-score and use it to assess how far an observation is from the mean
2. Understand correlation, how it is calculated, what it tells us, and what various correlations look like
3. Understand regression, how it is similar/different from correlation, what it tells us, what it doesn't tell us

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