Correlation (part 2)

Ryan Miller

Introduction

- The last presentation introduced Pearson's **correlation coefficient** as summary measure used to describe the relationship between two quantitative variables
- This presentation will cover several misconceptions and common mistakes when applying the correlation coefficient

Mistake #1 - Non-linear Relationships and Outliers

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

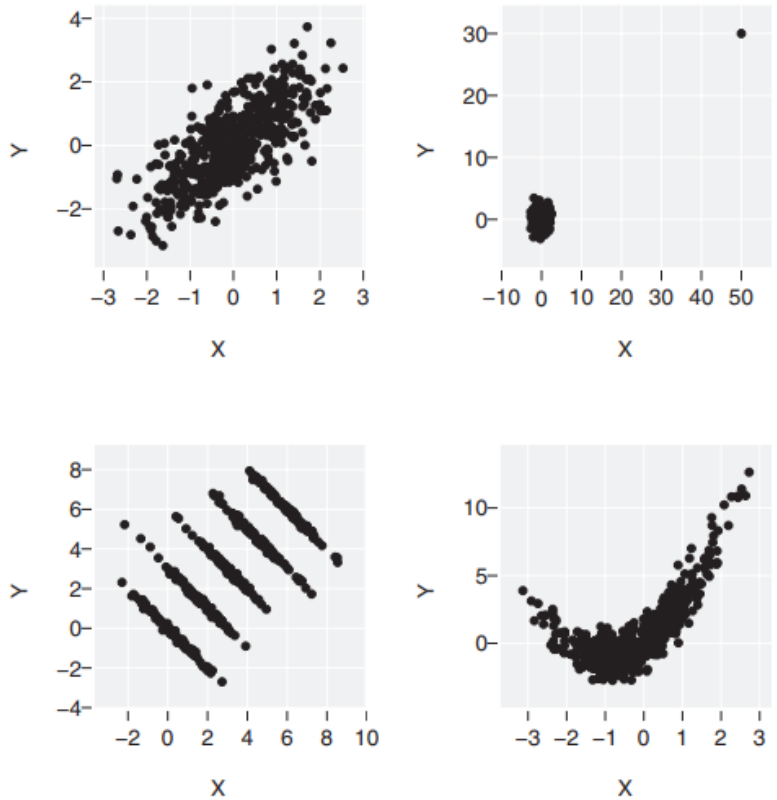


Fig. 6.1. Studying dependence between X and Y. All four pairs of variables have correlation approximately equal to 0.7, but they all have very different patterns. Only the top left plot shows two variables matching a dependence modeled by correlation.

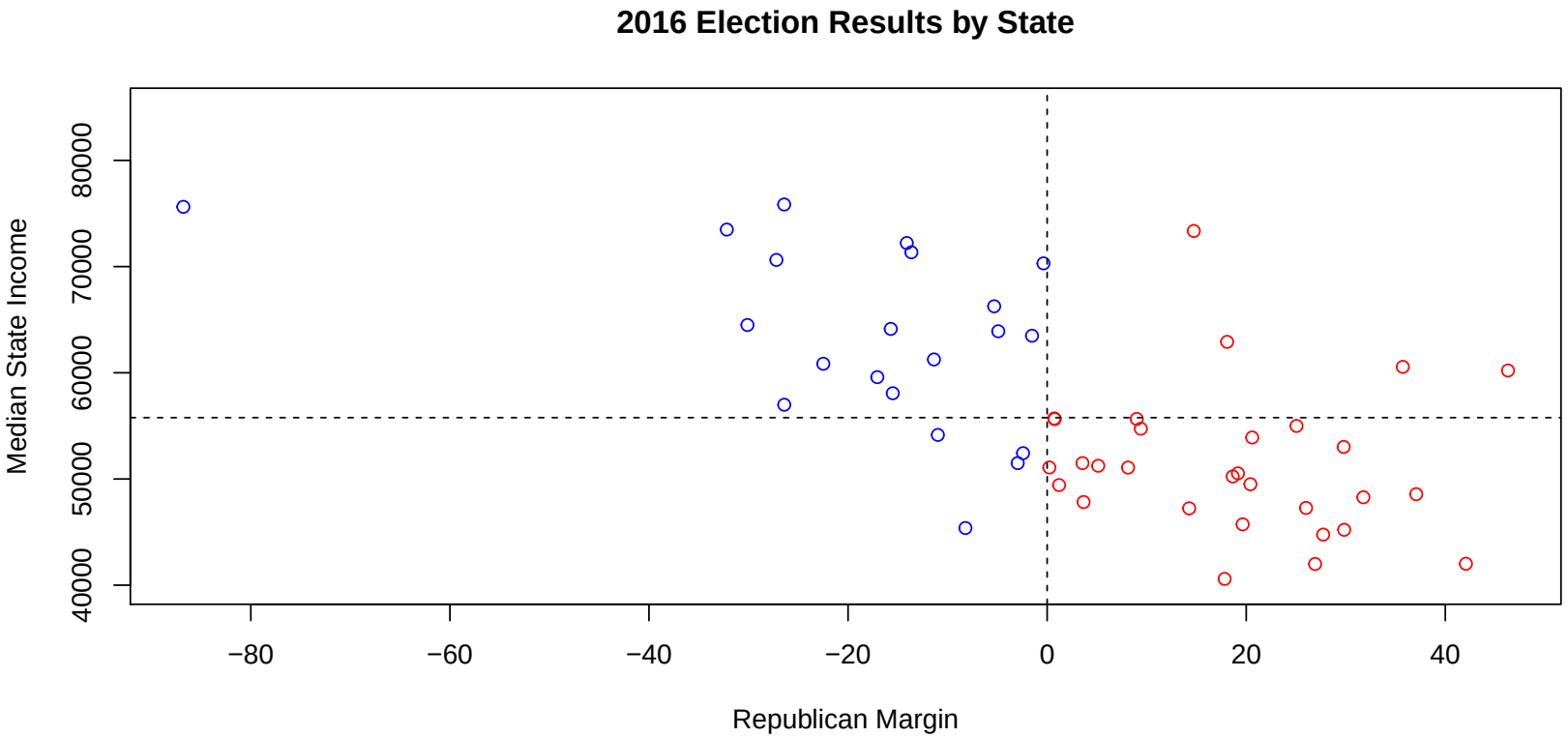
Advice

- Always check the scatterplot before blindly jumping to the correlation coefficient
- Do not report the correlation coefficient in situations where it can be misleading (outliers, non-linear relationships, omitted variables)

Mistake #2 - Ecological Correlations

- **Ecological correlations** compare variables at an ecological level (ie: The cases are aggregated data - like countries or states)
  - There is 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

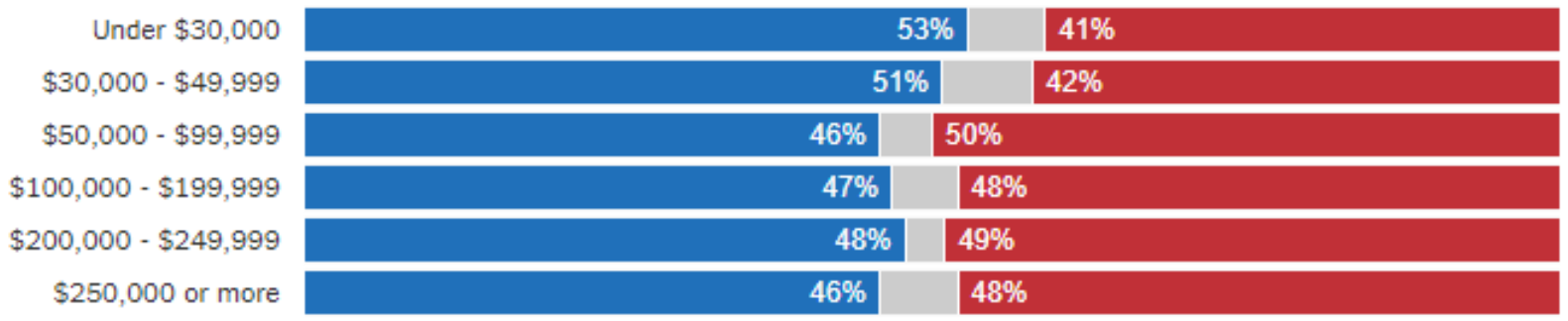
Ecological Correlations



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

The Ecological Fallacy

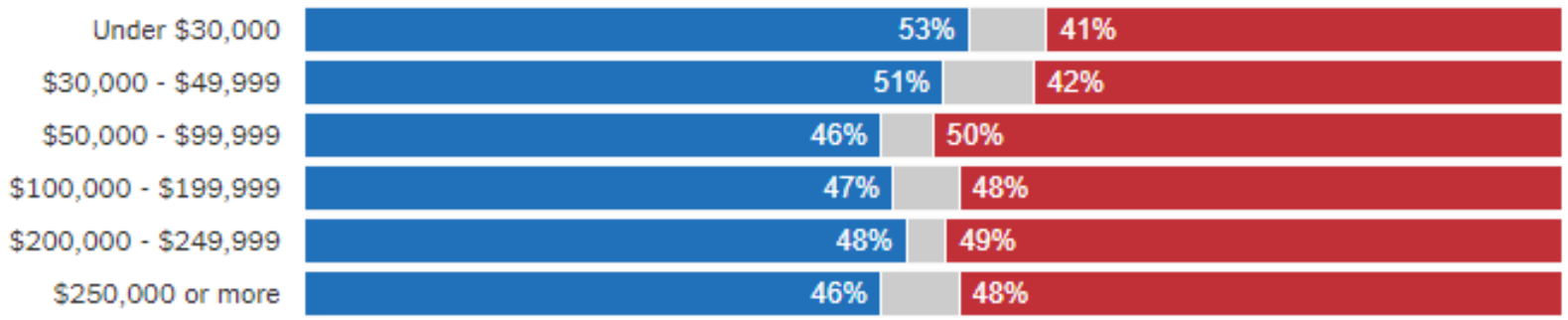
Using 2016 exit polls, conducted by the NY Times (Link), we can get a sense of how party vote and income are related *for individuals*:



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

The Ecological Fallacy

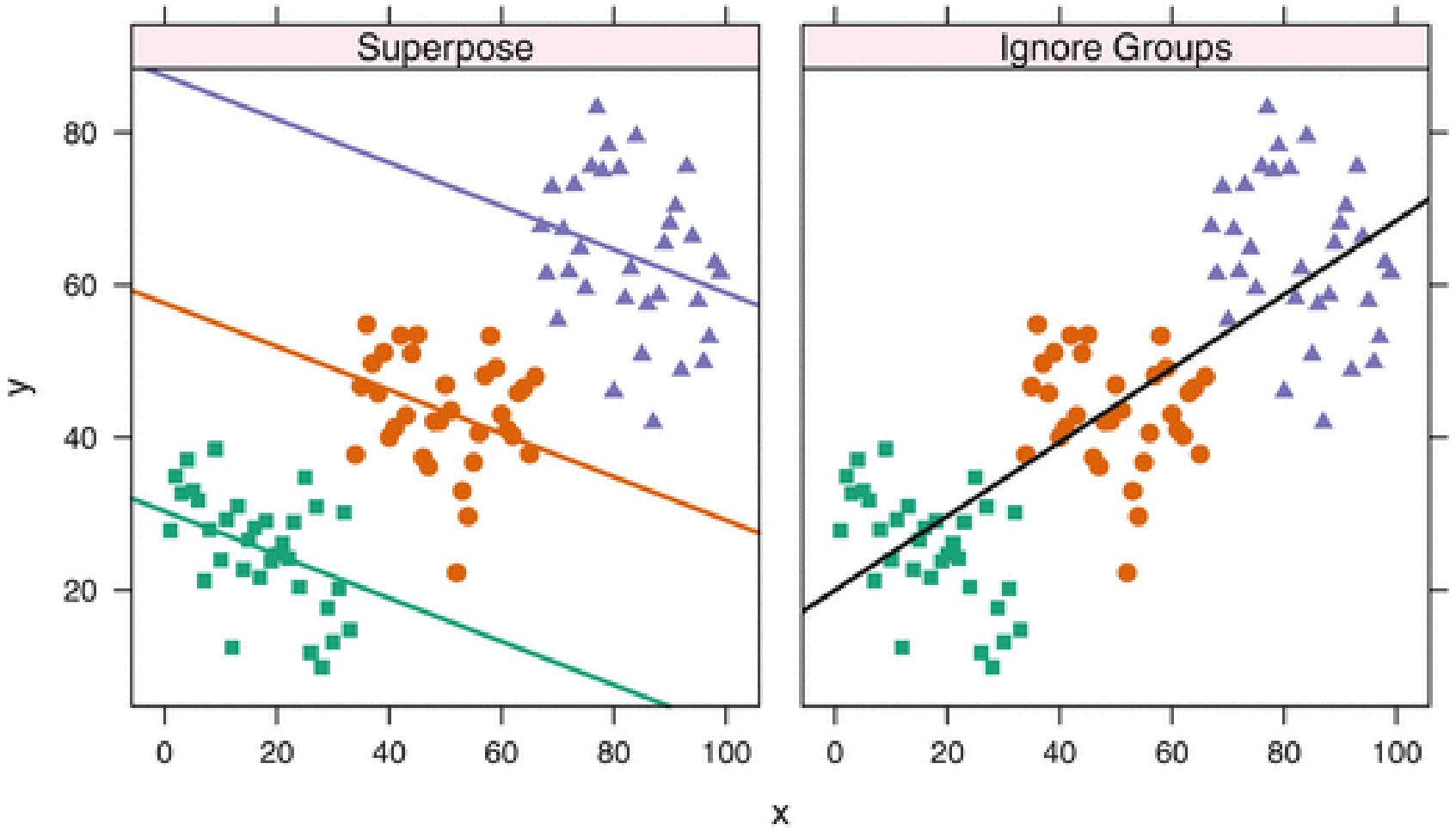
Using 2016 exit polls, conducted by the NY Times (Link), we can get a sense of how party vote and income are related *for individuals*:



- 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

Ecological Fallacy

The ecological fallacy can result from ignoring an important grouping variable:



Advice

- Always base your analysis around cases you're actually interested in
  - For example, analyze states when you're actually interested in talking about people
- Always explore your data thoroughly by considering scatterplots that color the points by group