Prediction vs. Bias: A Debate

Lucy Bernholz
Conference Co-Host; Senior Research Scholar, Stanford PACS; Director, Stanford Digital Civil Society Lab
@p2173

Andrew Means
Conference Co-Host; Head, beyond.uptake
@meansandrew

Kristian Lum
Lead Statistician, Human Rights Data Analysis Group
@KLdivergence

Candace Thille
Founding Director, Open Learning Initiative, Carnegie Mellon University and Stanford University
@StanfordEd

#DoGoodData
1

START WITH THE

LEARNER
DIVERSITY OF THE LEARNER
DIVERSITY OF THE LEARNER

- background knowledge
- relevant skills
- future goals
DIVERSITY OF THE LEARNER

- background knowledge
- relevant skills
- future goals

- attributions—how learners explain the causes of experiences
THE INNOVATION
PERSONALIZED AND ADAPTIVE INSTRUCTION
OARS

EDA: Examining Relationships

Learning Objectives in this Module

Interpret the value of the correlation coefficient, and be aware of its limitations as a numerical measure of the association between two quantitative variables.

mastered by: 36 / 84

In the special case of linear relationship, use the least squares regression line as a summary of the overall pattern, and use it to make predictions.

mastered by: 11 / 84
Interpret the value of the correlation coefficient, and be aware of its limitations as a nun.

**Skills Required for this Learning Objective**

**Estimating r**
- (learners who HAVE DEMONSTRATED MASTERY of this skill)
- (learners who HAVE NOT DEMONSTRATED MASTERY of this skill)
- mastered by: 36 / 84 (requires at least 3 attempts)

**Interpreting regression**
- (learners who HAVE DEMONSTRATED MASTERY of this skill)
- (learners who HAVE NOT DEMONSTRATED MASTERY of this skill)
- mastered by: 11 / 84 (requires at least 3 attempts)
Reconsider the strength of the linear relationship. \( r = 0.678 \) Reconsider the direction and the strength of the linear relationship. \( r = 0.845 \) Reconsider the direction and the strength of the linear relationship. You may find it easier to look at the scatterplots for all 6 questions first. You will notice that the correlation coefficient is used only once. For each scatterplot, first determine the direction of the relationship, and then determine the strength of the relationship.

attempted by: 15 / 84

correct on first attempt: 1 / 15

correct on last attempt: 13 / 15

We compute the correlation between gestation period and longevity and find that \( r = 0.663 \).

Based on these findings, what is the strength of the relationship between gestation period and longevity?

Weak and positive While you are correct that the relationship is positive, an \( r \) value of 0.633 is not considered weak. Moderate and positive The relationship is positive and moderate, as \( r = 0.633 \) increases so does longevity, and an \( r \) value of 0.633 is considered moderate. Strong and positive While you are correct that the relationship is positive and strong, an \( r \) value of 0.633 is not considered strong. Moderate and negative While you are correct that an \( r \) value of 0.633 is considered moderate, a negative relationship indicates that as values in one variable increase values in the other variable also increase, while the other variable decrease. What are range of the values for \( r \) that make the strength of the relationship weak, moderate, or strong?

attempted by: 38 / 84

correct on first attempt: 35 / 38

correct on last attempt: 37 / 38

Looking at the scatterplot you can see that there is an outlier in both longevity (40 years) and gestation (645 days). Note: This outlier corresponds to the above data point.

What do you think will happen to the correlation (0.663) if we remove this outlier?

Increase After removing the outlier, the correlation decreases from 0.663 to 0.519 because the data point had been in the same linear direction as the other data points. A decrease in the strength of the relationship between longevity and gestation period. In this case, it would be most informative to report both correlations. Decrease After removing the outlier, the correlation decreases from 0.663 to 0.519 because the data point had been in the same linear direction as the other data points. A decrease in the strength of the relationship between longevity and gestation period. In this case, it would be most informative to report both correlations. Decrease After removing the outlier, the correlation decreases from 0.663 to 0.519 because the data point had been in the same linear direction as the other data points. A decrease in the strength of the relationship between longevity and gestation period. In this case, it would be most informative to report both correlations. Decrease After removing the outlier, the correlation decreases from 0.663 to 0.519 because the data point had been in the same linear direction as the other data points. A decrease in the strength of the relationship between longevity and gestation period. In this case, it would be most informative to report both correlations. Decrease After removing the outlier, the correlation decreases from 0.663 to 0.519 because the data point had been in the same linear direction as the other data points. A decrease in the strength of the relationship between longevity and gestation period. In this case, it would be most informative to report both correlations. Decrease After removing the outlier, the correlation decreases from 0.663 to 0.519 because the data point had been in the same linear direction as the other data points. A decrease in the strength of the relationship between longevity and gestation period. In this case, it would be most informative to report both correlations.

attempted by: 37 / 84

correct on first attempt: 15 / 37
Knowledge Modeling

$h_0 \rightarrow h_1 \rightarrow h_2$

From Possibilities to Responsibilities

#DoGoodData

Data on Purpose
3

THE

SCIENCE
• guided learning
• complex cognitive tasks
• targeted hints & feedback
Social-Psychological Science

Mindset
Social belongingness
Stereotype Threat
A PART OF

HIGHER EDUCATION

#DoGoodData
“Improvement in post secondary education will require converting teaching (and courseware, platform & analytic system development) from a solo sport to a community-based research activity.”

• Herbert Simon 1991 (modifications 2015)
“Without a complete revolution...in our approach to teaching... we cannot go beyond (current levels) of productivity”  
William Baumol, 1967

Our Message: Such a revolution is possible happening

Our Question: Who will lead it?
START WITH AND END WITH THE LEARNER
Registrant Data

Is this data biased?

A data set is non-representative or biased* with respect to a population of interest if not all elements of the population had the same probability of appearing in the dataset.

*This is not a value judgment about the dataset or those who compiled it. A “statistically biased” dataset does not require that anyone acted with bias in the colloquial sense of the word.
Is this data biased?

A data set is *non-representative* or *biased* with respect to a population of interest if not all elements of the population had the same probability of appearing in the dataset.

*This is not a value judgment about the dataset or those who compiled it. A “statistically biased” dataset does not require that anyone acted with bias in the colloquial sense of the word.*
Registrant Data

Is this data biased?

A data set is non-representative or biased* with respect to a population of interest if not all elements of the population had the same probability of appearing in the dataset.

*This is not a value judgment about the dataset or those who compiled it. A “statistically biased” dataset does not require that anyone acted with bias in the colloquial sense of the word.
Machine Learning, Statistical Modeling, Predictive Analytics, etc.

- logistic regression
- k-means clustering
- random forest
- support vector machines
- kernel density estimation
- reinforcement learning
- neural nets
- generalized linear model
- nearest neighbors
- boosting
- deep learning
- ensemble models
- principal components analysis
Police Data

Are police records and unbiased sample of all crimes?

• Variation in reporting rates
  » NCVS indicates that reporting rates vary substantially by demographic characteristics, i.e. some crimes are more likely than others to be reported to police depending on who was victimized.
  » In this case, the bias derives not from the police themselves but from the community the police serve.

• Variation in police attention
  » Crimes that are committed in areas that are highly patrolled by police are more likely to be discovered by police than those committed in less patrolled areas.
  » Police are not tasked with collecting a random sample, so bias in the data may come from legitimate police strategy.

• Variation in rates of enforcement for similar criminal behavior
  » While white and black populations use marijuana at similar rates, blacks are arrested for marijuana possession at a rate several times that of whites.*
Machine Learning, Statistical Modeling, Predictive Analytics, etc.

Ways to learn patterns & structure in data

logistic regression  k-means clustering
random forest  support vector machines
kernel density estimation  reinforcement learning
neural nets  generalized linear model
boosting  deep learning
ensemble models  principal components analysis

**Which model you pick defines how you learn the patterns and the types of patterns we can learn.**
What is predictive policing?

Predictive policing uses police records to learn patterns in the occurrence of crime. Using these patterns, the computer then predicts the most likely locations of future crimes where crime will be detected in the future.

Additional police are dispatched to the locations with the highest predicted rate of crime.
Drug Crimes in Oakland, CA

Data collected and cleaned by OpenOakland.org

Drug Crimes in Oakland, CA

Image Copyright, 2013, Weldon Cooper Center for Public Service, Rector and Visitors of the University of Virginia (Dustin A. Cable, creator)
Drug Crimes in Oakland, CA
Drug Crimes in Oakland, CA

Percent with Targeted Policing

Percent Using Drugs

[Maps and bar charts showing data on drug crimes in Oakland, CA, comparing white, black, and other groups.]
What if…

When police are sent to a location, they find a little more crime than they would have?
Audience Q&A

Lucy Bernholz
Conference Co-Host; Senior Research Scholar, Stanford PACS; Director, Stanford Digital Civil Society Lab
@p2173

Andrew Means
Conference Co-Host; Head, beyond.uptake
@meansandrew

Kristian Lum
Lead Statistician, Human Rights Data Analysis Group
@KLdivergence

Candace Thille
Founding Director, Open Learning Initiative, Carnegie Mellon University and Stanford University
@StanfordEd