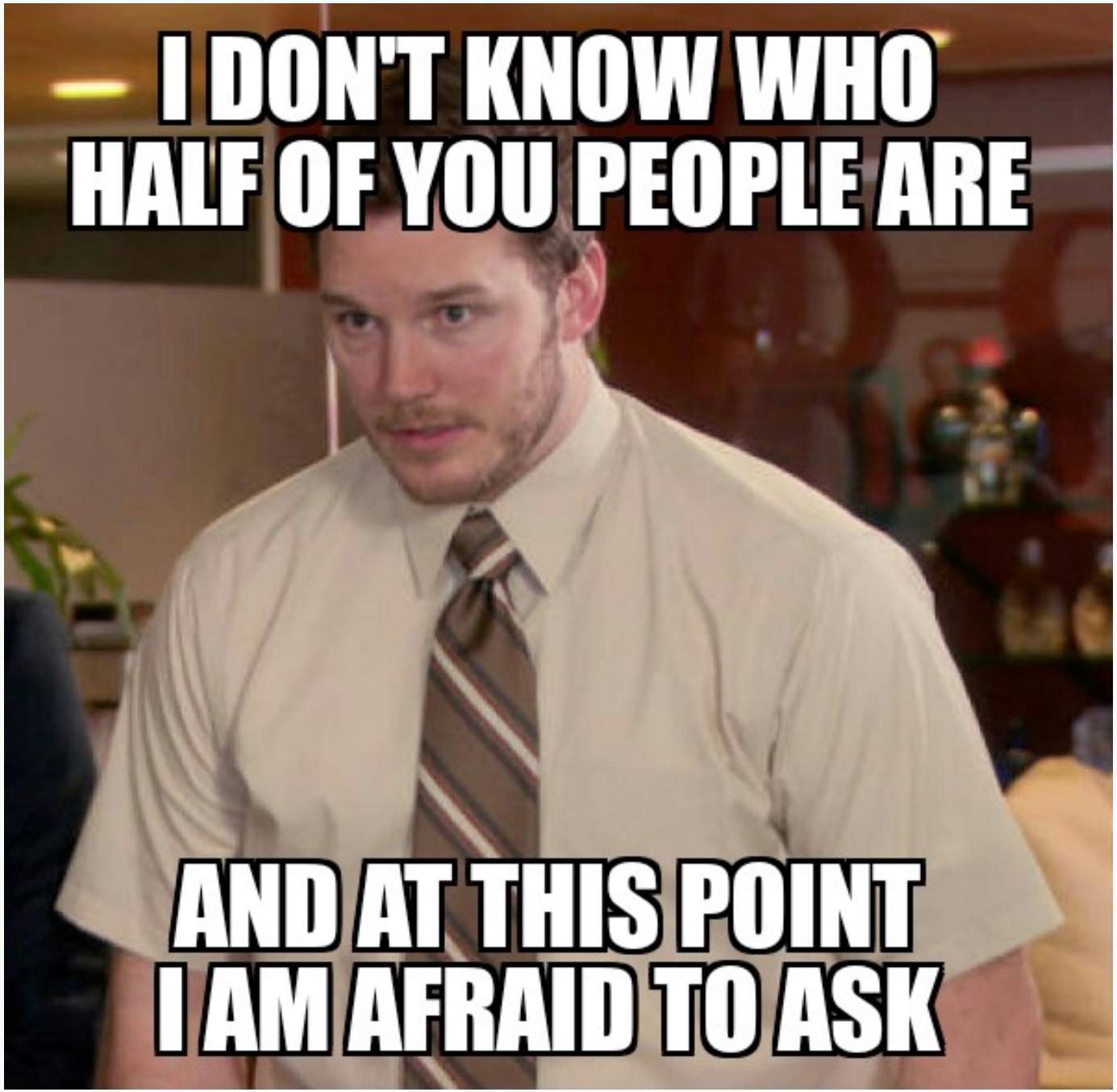
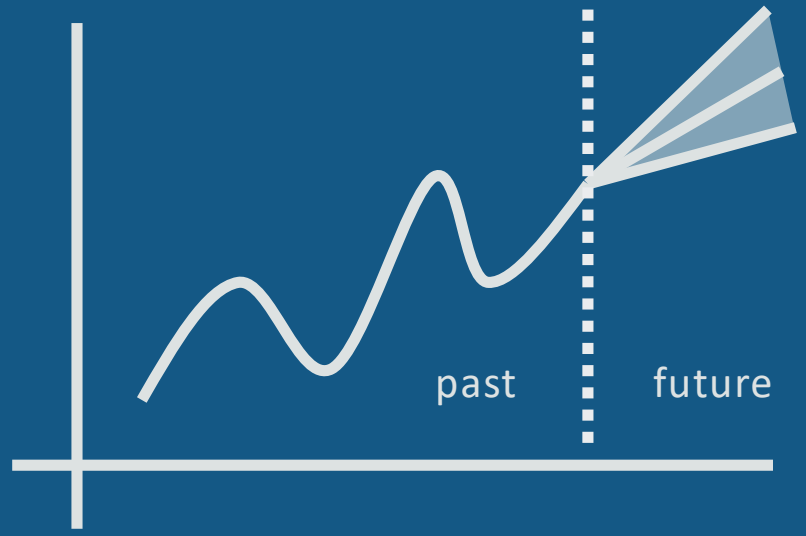


I’m kidding! I know all of you

EDUC 7610

Chapter 7

Prediction with Regression



Fall 2018

Tyson S. Barrett, PhD

“Two Cultures”

1. One culture that is about **understanding effects** (the focus of the majority of the class)
2. The other is about **prediction** – high prediction is good even if the model isn’t all that interpretable (the focus of this chapter)

Extra credit opportunity – read this paper and write a half page response to it

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:

There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

There are two different approaches toward these goals:

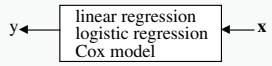
The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

response variables = f(predictor variables, random noise, parameters)

Leo Breiman is Professor, Department of Statistics, University of California, Berkeley, California 94720-4735 (e-mail: leo@stat.berkeley.edu).

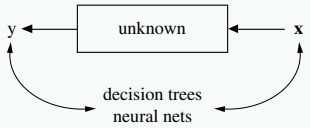
The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on **x** to predict the responses **y**. Their black box looks like this:



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

In this paper I will argue that the focus in the statistical community on data models has:

- Led to irrelevant theory and questionable scientific conclusions;

Predictive Modeling and Machine Learning Examples

Google's Assistant

https://www.youtube.com/watch?v=D5VN56jQMWM

Tesla's Autopilot

https://www.youtube.com/watch?v=cbKA-MLoZLM

Prediction with Regression

- Many machine learning/predictive modeling techniques are built on regression (e.g., lasso, ridge, elastic net, polynomial regression)
  - Others (random forests, neural networks, support vector machines) are more complex and less interpretable
- We'll do the simplest type – linear regression for prediction

Why prediction with regression?

Sometimes the problem is about predicting an outcome

(e.g., predicting risk of substance abuse, predicting future income, etc.)

- Many clinical and practical applications

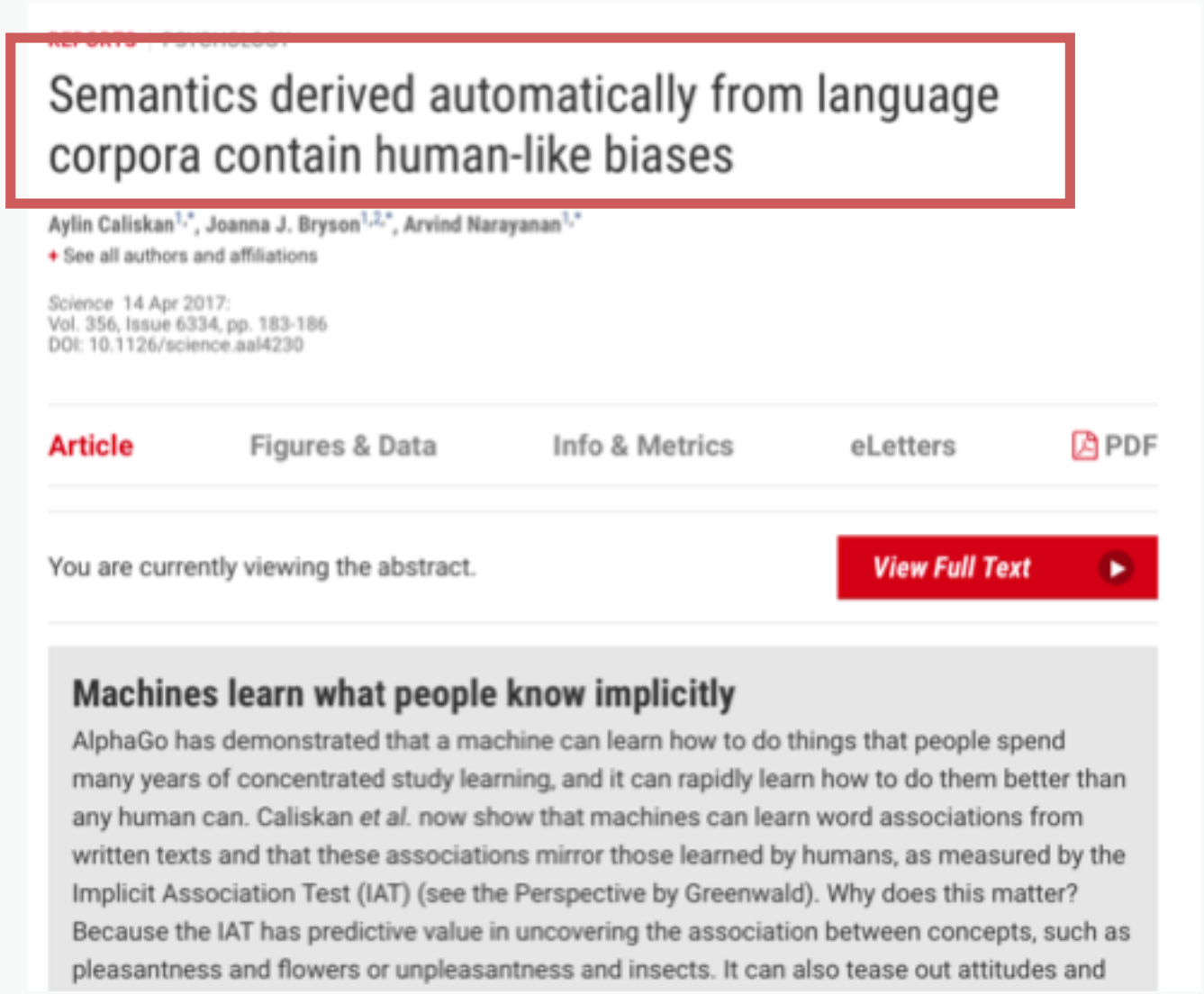
Using regression (or some other algorithm) can remove some bias and error from more subjective predictions or decisions

- Regression, although simple, is fairly accurate in predicting outside cases
- Best models generally use information that can be gathered cheaply and effectively

Prediction

http://science.sciencemag.org/content/356/6334/183

Although Darlington and Hayes say the predictive modeling is not biased, we have to understand that it can be biased if the data are biased



Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan[1,*], Joanna J. Bryson[1,2,*], Arvind Narayanan[1,*]
+ See all authors and affiliations

Science 14 Apr 2017:
Vol. 356, Issue 6334, pp. 183-186
DOI: 10.1126/science.aal4230

Article | Figures & Data | Info & Metrics | eLetters | PDF

You are currently viewing the abstract. View Full Text

Machines learn what people know implicitly

AlphaGo has demonstrated that a machine can learn how to do things that people spend many years of concentrated study learning, and it can rapidly learn how to do them better than any human can. Caliskan *et al.* now show that machines can learn word associations from written texts and that these associations mirror those learned by humans, as measured by the Implicit Association Test (IAT) (see the Perspective by Greenwald). Why does this matter? Because the IAT has predictive value in uncovering the association between concepts, such as pleasantness and flowers or unpleasantness and insects. It can also tease out attitudes and

How do we know we have a good model?

Common Problems

- Over-fitting
- Lack of generalizability
- Finding the right predictors
- Low predictive power

Potential Solutions

- Cross-validation
  - 5 or 10-fold CV
- Leave-One-Out
- Variable selection
- Larger sample

Over-fitting Model

Definition

Some aspects of the sample will represent the population
but other parts are only representative of the sample

Potential Solutions

Use cross-validation or leave-one-out
Use a random sample
Use less complexity in the model

Over-fitting Model

Cross-Validation

1. Split the sample into X number of subsamples
2. Fit the model to X – 1 parts and then predict the X^{th} part with that model
3. Assess the accuracy of that prediction
4. Do X times

Leave-one-out is a special case of cross-validation

Lack of Generalizability

Definition

The model lacks in its ability to generalize to the population or other samples

Potential Solutions

Use cross-validation or leave-one-out
Use a random sample
Use less complexity in the model

Finding the right predictors

Definition

We often don't care about causal relationships, we just want good prediction so we want to use the fewest number of the best predictors

Potential Solutions

Use theory and/or empirical information
Use automatic selection methods (Lasso, elastic net, backward/forward stepwise)

Low Predictive Power

Definition

The model doesn’t predict the cross-validated outcome with enough accuracy

Potential Solutions

Larger sample size
Use different predictors
Improve measurement (reduce measurement error)

Possible Predictor Configurations

The way the predictors are related affect how well a group of predictors can predict the outcome

Four important configurations:

- Complementarity = when unique contribution of the combined set of predictors exceeds the sum of the individual contributions –
  - $R^2 > (r^2_{YX_1} + r^2_{YX_2})$
- Suppression = any case in which a variable receives a sig negative weight when it has a positive or zero correlation with Y
- Independence = predictors are not correlated
- Redundancy (partial or full) = when variables share information about Y

Some Final Thoughts

1. The model accuracy is just an estimate so we can use "resampling" methods (repeating the modeling multiple times with a random selection of the sample) to find out how much variability there is in that estimate

2. The principles we've discussed apply to other machine learning techniques as well (we'll get some practice with some others in the chapter example)