JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Prediction: Using statistics to put your money where your mouth is

Rafael A. Irizarry

What distinguishes science from religion?

We can predict!

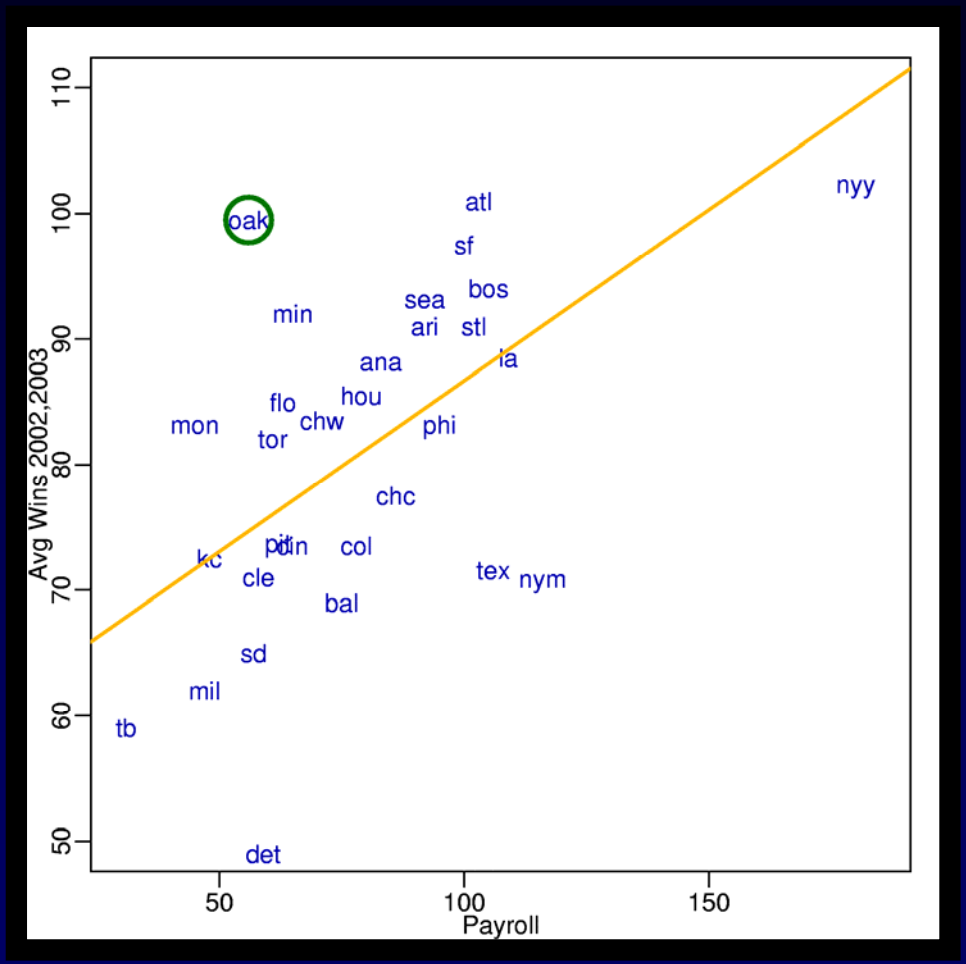
Examples

Religion	Science
Earth is flat	Earth is round (Eratosthenes of Cyrene 220 BC)
Sun orbits earth	Earth orbits sun (Copernicus circa 1500)
Wine turns into blood	Tastes like wine (me 1984)

Two approaches

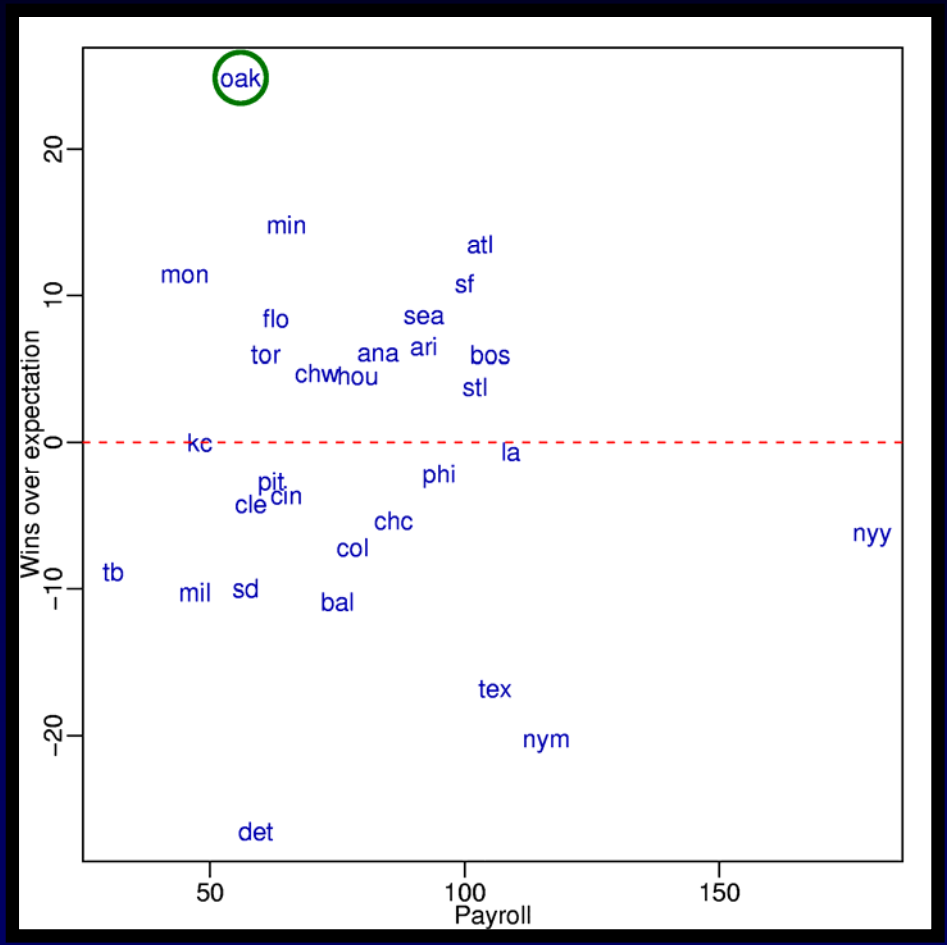
- **Physical models**
  - **A physicist can predict any two objects will hit ground at same time**
  - **A chemist predicts *Na* + *Cl* tastes salty**
  - **An astronomer can predict the next eclipse**
  - **An M.D. can predict that if you eat cyanide you die**
  - **An engineer can predict a bridge won't fall**
- **Stochastic models**
  - **I can predict that if you toss 10,000 coins you will see between 45-55% heads**
  - **I can predict Vegas will make money**
  - **I can predict the Oakland A's will win more games than any other team with similar payroll**

MLB Wins versus Payroll



Oakland A's ignore "experts" and use statistics instead

MLB residuals versus Payroll



Many Problems in Science

Sometimes we want to understand nature

Sometime we don’t really care

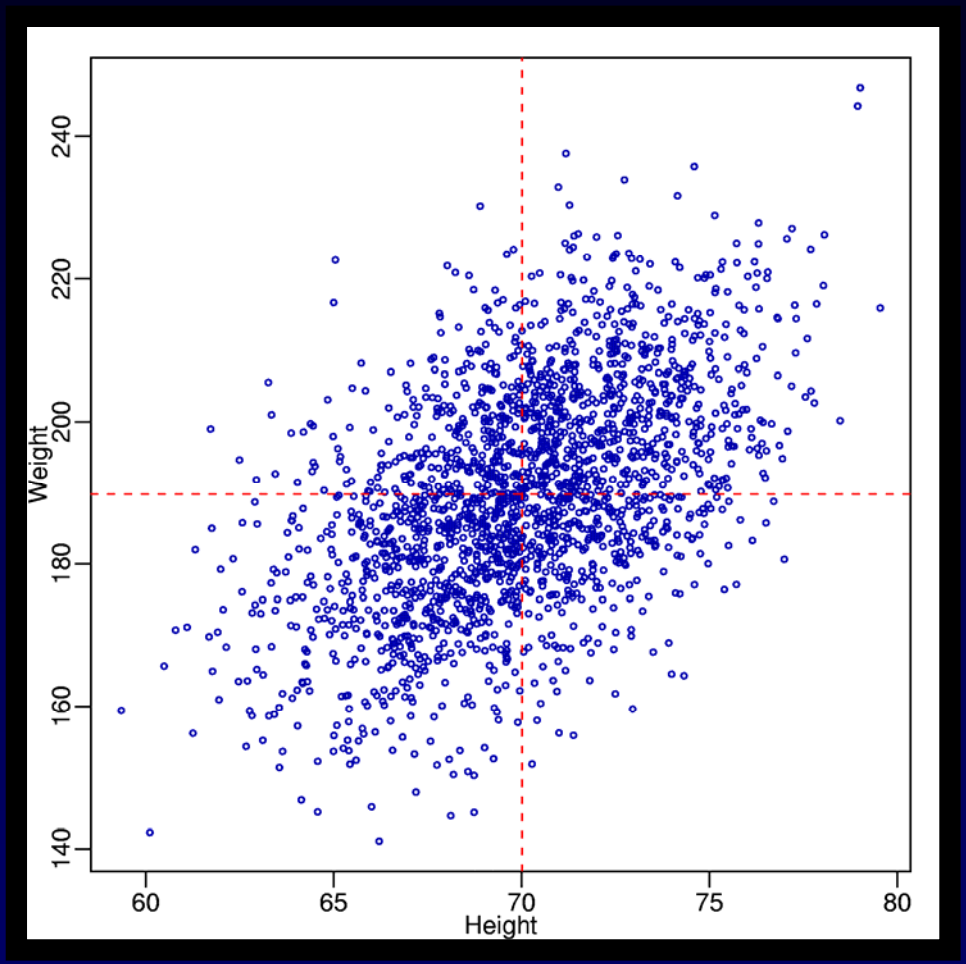
We are always happy if we can predict Y

Most common approach

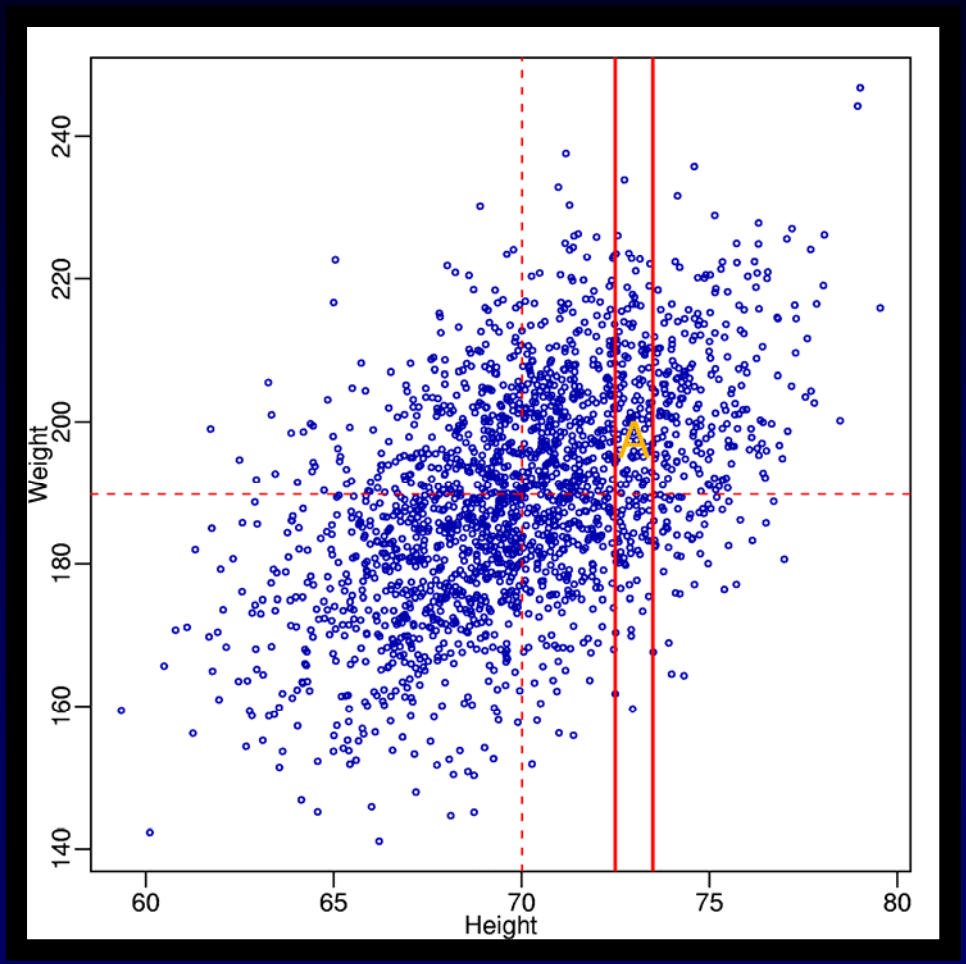
- Use parametric statistical model

- Fit the model, interpret parameters, predict Y given X

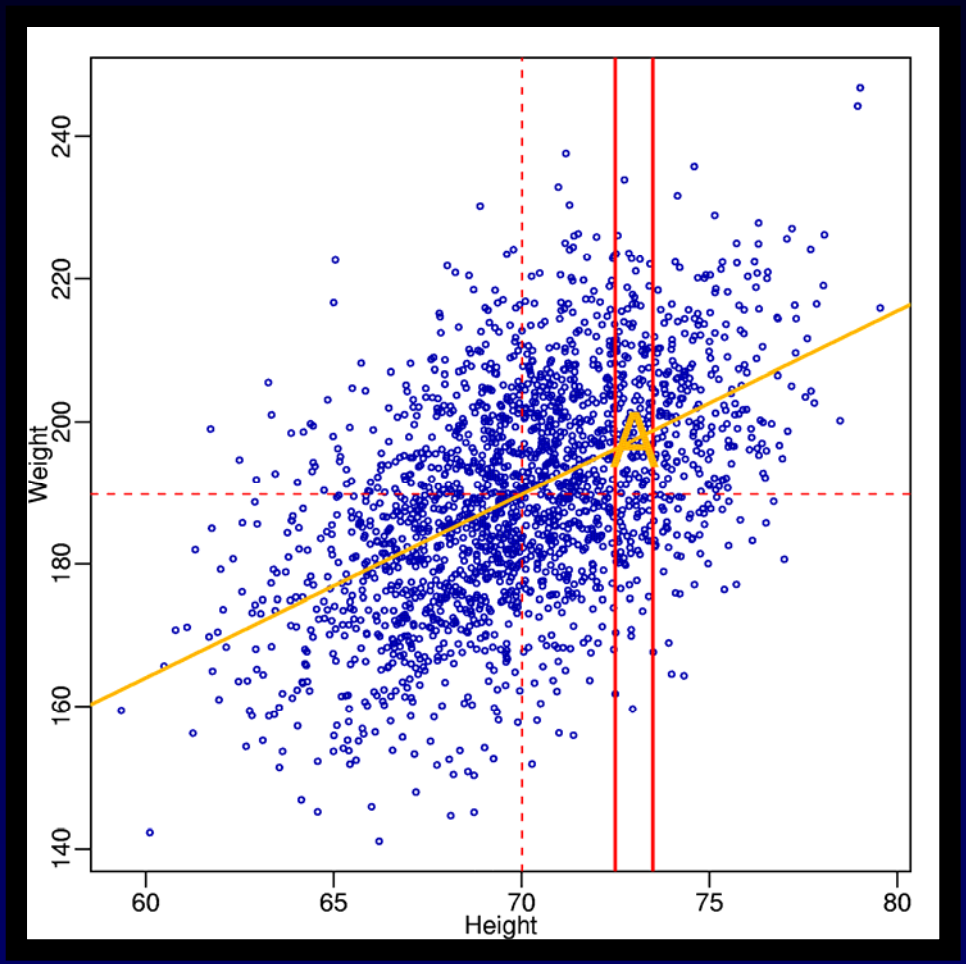
Example



Example



Example



Regression

- **Model: height and weight are normal and correlated then regression line gives *best* predictor**
  - ***E[Y | X] = Avg Y + (SD of Y / SD of X) (correlation) (X - Avg X)***
- **But this is only the case if model is correct**
- **Regression is now used in applications where its hard to tell if assumptions hold**

When are parametric models used?

- **Used lots:**
  - **Behavioral Sciences, Psychology, Epidemiology, Economics**
- **Not used much or at all:**
  - **Finance, fraud detection, zip code reading, face/voice recognition**
- **Lack of proper assessments causes unwarranted optimism about models**

Example

- Create a binary outcome and 6 covariates for 25 individuals
- Make everything completely uncorrelated
- Fit a regression models
- A likely result*
  - AIC chooses a model with 4 covariates
  - One covariate has $p < 0.05$
  - If we predict outcomes we get 80% right!
- Is this a good model? How would we know in real life?

*For one simulation. Code to reproduce is available upon request

Over-fitting

- How can can we predict 80% when there is no information in the covariates?
- Important fact: If we assess the fit of model on the same data we fitted, it will appear better than it really is.
- A fair assessment would happen on a new data set… which we rarely can get… but we can fake it!

Cross-validation

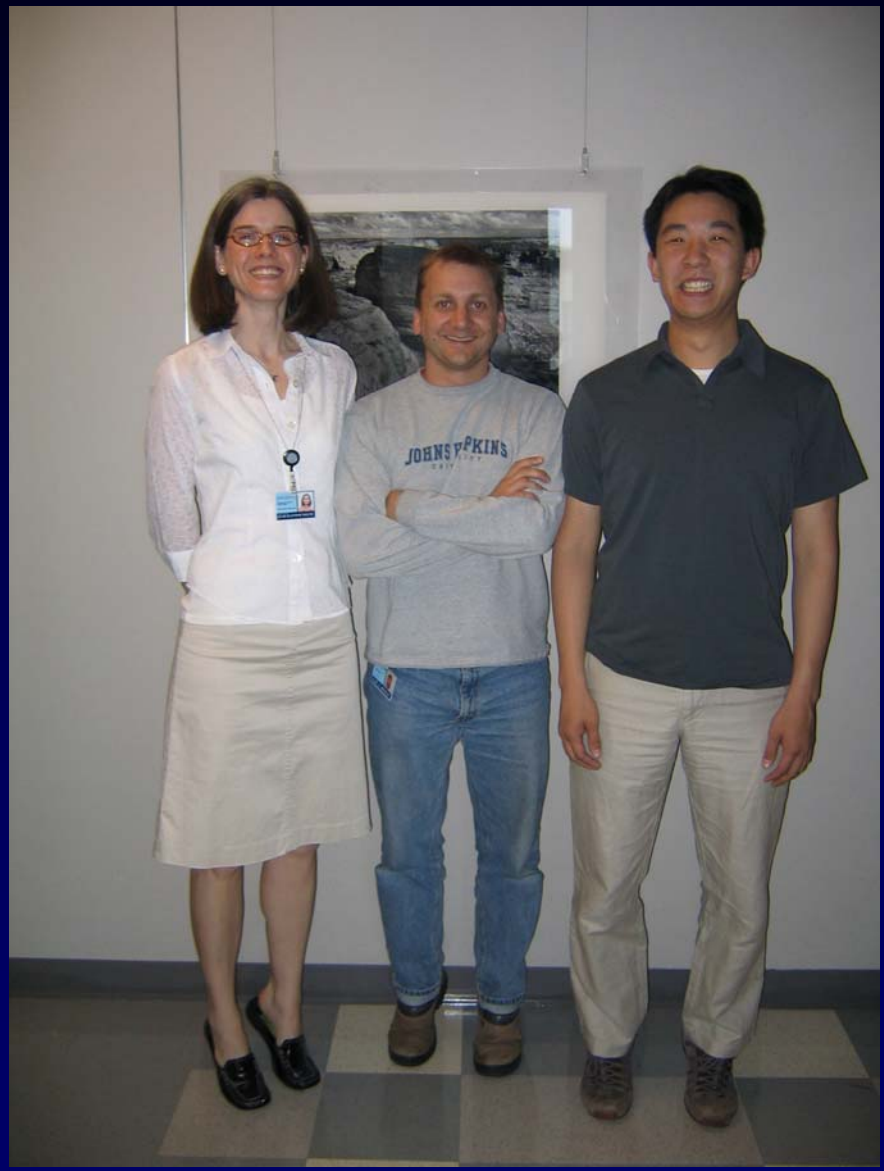
- Leave out 10% of data at random (test set)
- Fit model on the remaining 90% (train set)
- See how well our fit predicts on the test set
- Repeat above various times
- In our example our CV error is 50% !

Example of over-fitting

Height	Gender	Swede	Age	IQ	Hair color	Weight	Region	Blood Type	Asian
5’8’’	M	Yes	35	118	Blond	150	Midwest	AB	No
6’1	F	No	28	120	Brown	110	Midwest	O+	No
6’1	M	No	29	118	Black	190	NE	A	Yes

- We want to predict height
- Easily find a model that with perfect R^2
- An example says, on average:
  - Women are 1 inch taller than men
  - Swedes are 5 inch shorter than non-Swedes

My Random Sample



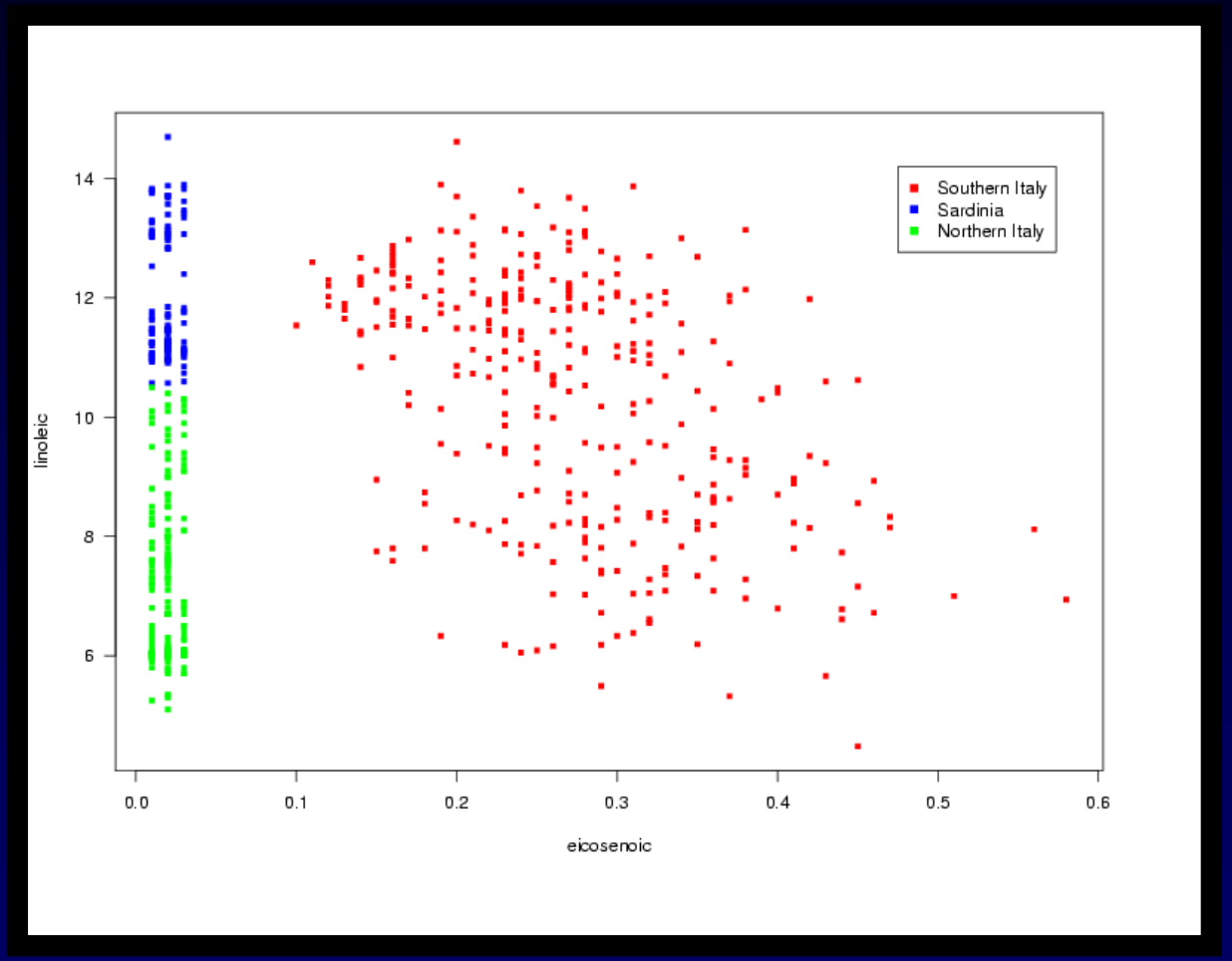
Hard Problems

- Many covariates… easy to over-fit
- We care mostly/only about predicting outcome
- Examples abound
- Statisticians have developed some of the best methods… despite few of us working on these problems
- CV is an essential tool

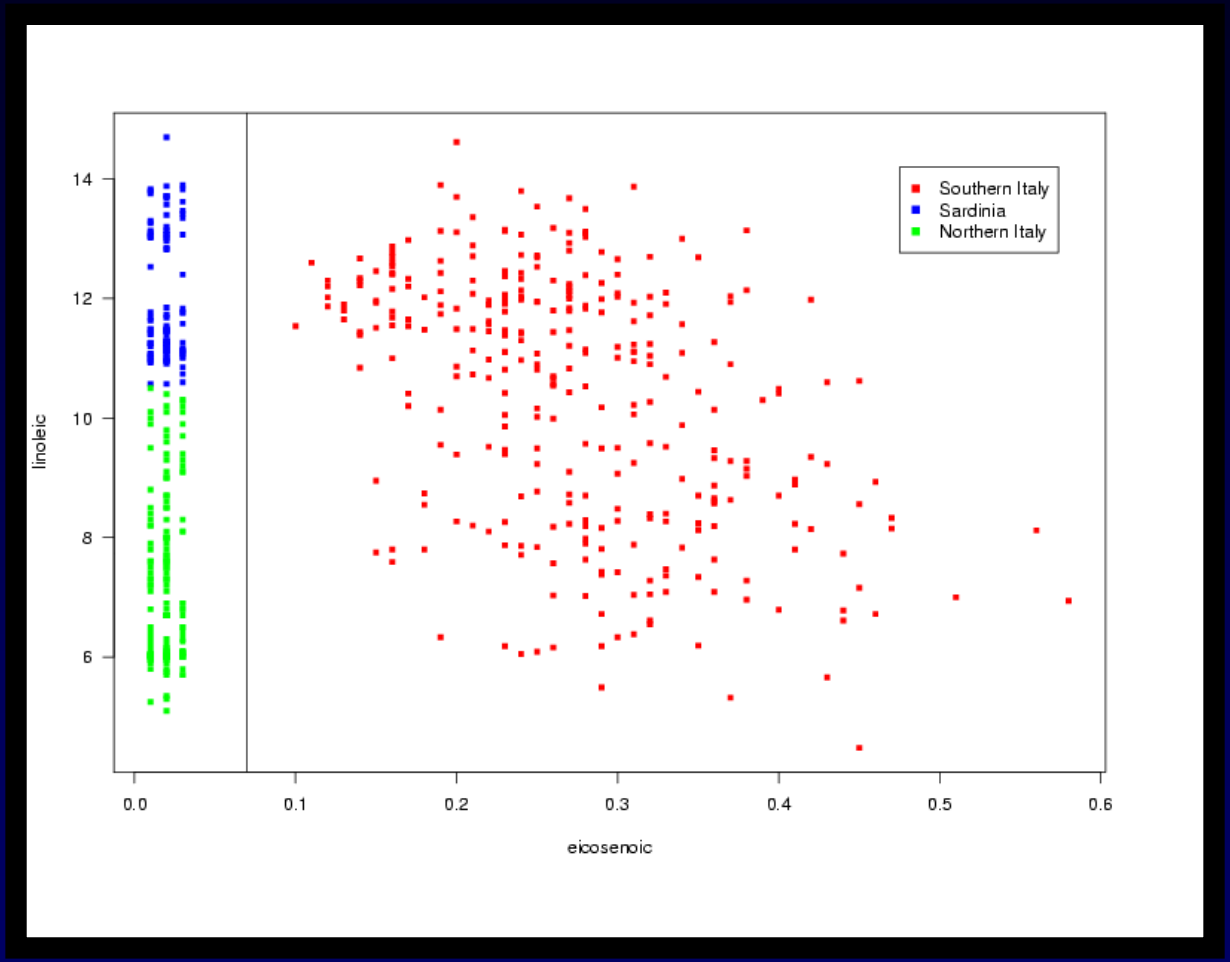
Algorithmic Approach

CART
Random Forests
Logic Regression
Linear Discriminant Analysis

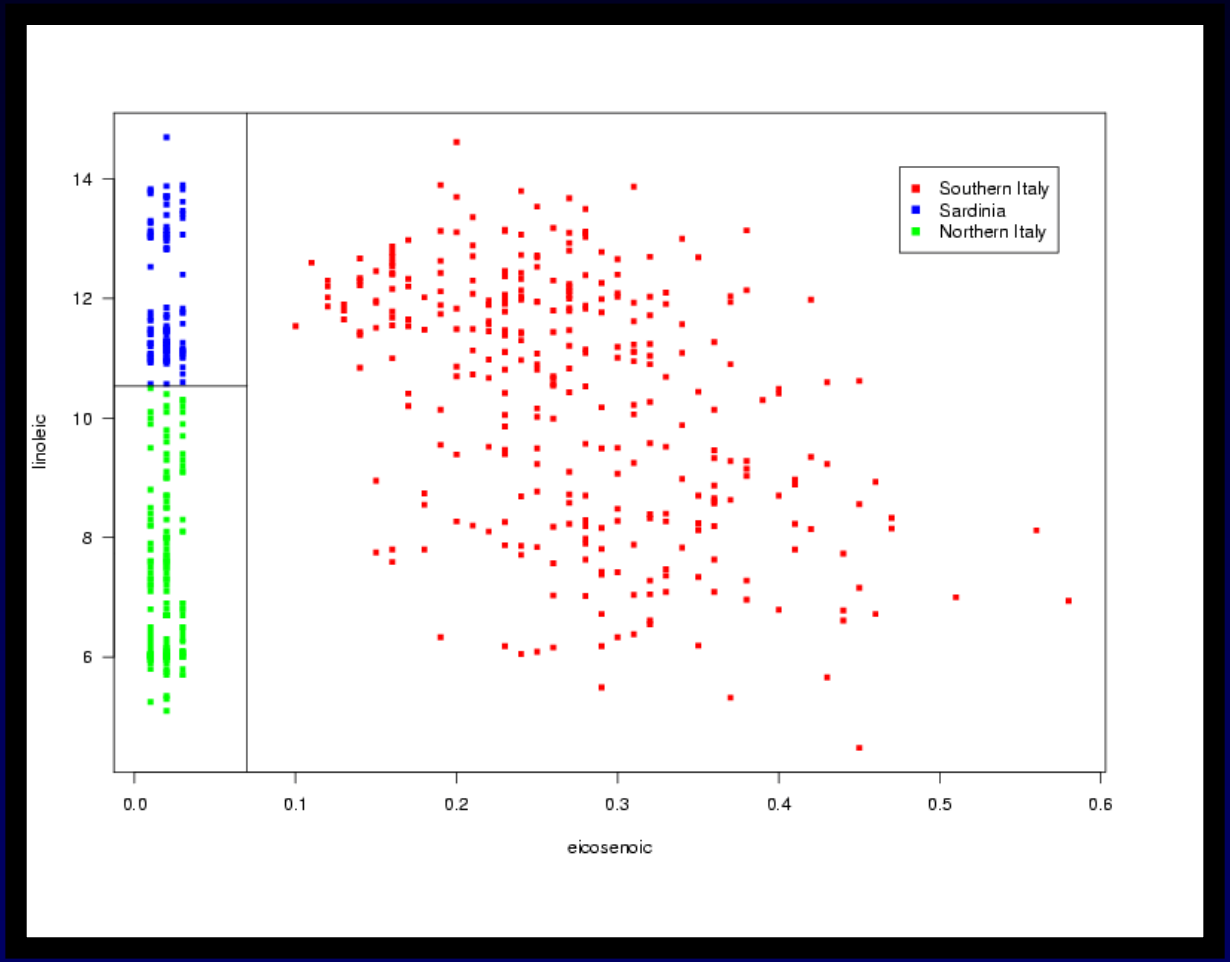
CART: Olive Example



CART: Olive Example



CART: Olive Example



But how do we decide what branches to use/keep? Pick best predictor!

Conclusion

- When interpreting a p-value consider the alternative hypothesis: “My model is wrong”
- If you can, use cross-validation to assess models
- There are many methods designed specifically for prediction

Some harsh quotes

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.

Leo Breiman

The whole area of guided regression is fraught with intellectual, statistical, computational, and subject matter difficulties

Mosteller and Tukey