

# Overfitting and Optimism in Models

Steyerberg, Chapter 5

Dave Glidden  
6 October 2009

“Left to our own devices we are not very good at picking out patterns from a sea of noisy data. To put it another way, we are all too good at picking out non existing patterns that happen to suit our purposes”

Efron and Tibshirani (1993)

# Outline

- Model overfitting
- Model optimism
- Bias/Variance trade off
- Bootstrap, Cross-Validation, Shrinkage
- Big Idea

# Lee Model

- Cohort of community-dwelling elders
- Predictors: age, co-morbidities, functional status: > 48 predictors
- Follow-up: mortality over 4 years
- Who is at the highest risk of death?

Lee, SJ et al. JAMA. 2006;295:801-808.

# We want to know

- What factors are predictive?
- What are the most predictive factors?
- How predictive is the model ?

# One Approach

- Choose the most significant predictors
- Backward selection
  - all variables in model
  - remove one with largest p-value
  - until all  $p < 0.05$  (or some other cutoff)

*Pure stepwise is always a bad strategy. This is for illustration.*

# “Best” Model

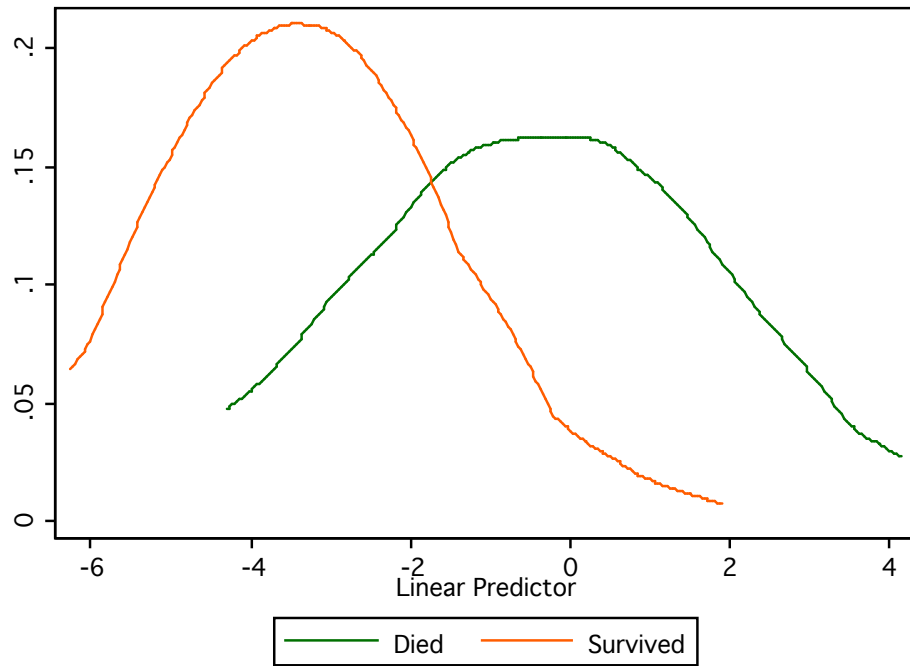
```
Logistic regression                               Number of obs   =       385
                                                  LR chi2(8)      =      103.25
                                                  Prob > chi2     =       0.0000
Log likelihood = -87.233708                    Pseudo R2      =       0.3718
```

dead	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
hrsdm	2.937781	1.405445	2.25	0.024	1.150275	7.503041
bmicat	3.30297	1.403299	2.81	0.005	1.43636	7.595317
hrsmemory	45.69148	64.06265	2.73	0.006	2.926788	713.3116
hrsarth	.1429438	.0869329	-3.20	0.001	.0434007	.4707968
hrscad	4.252373	1.974888	3.12	0.002	1.71126	10.56688
hrslung	4.829491	2.826925	2.69	0.007	1.533393	15.2107
meals	5.390685	2.847899	3.19	0.001	1.914054	15.18217
walkjog	9.540012	4.048986	5.31	0.000	4.152224	21.91882

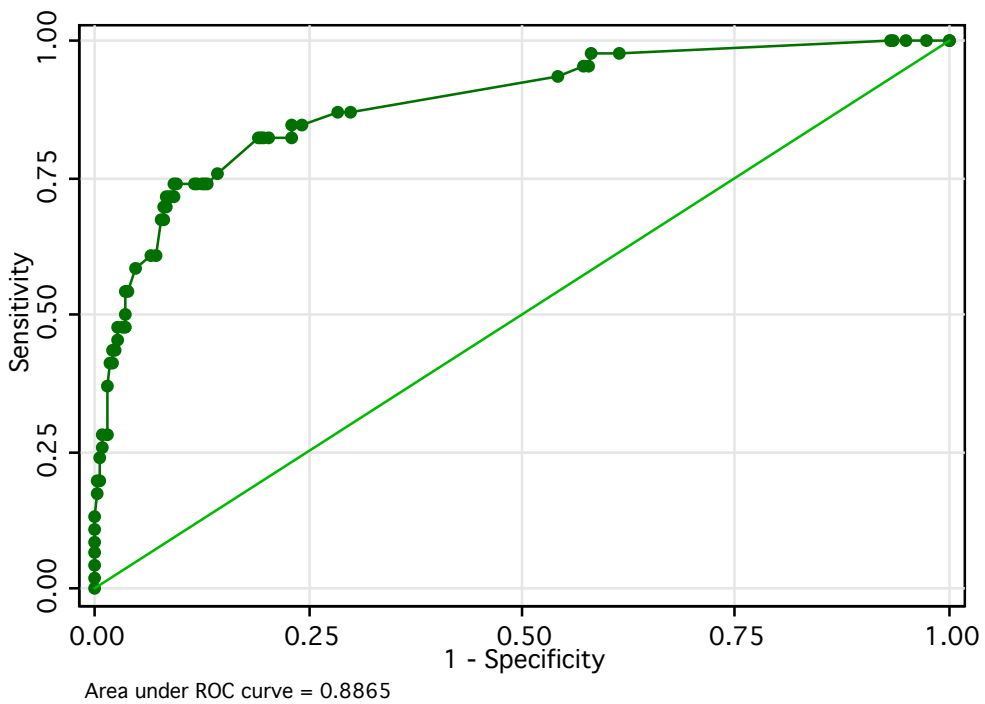
## Making a Prediction

- Get a OR for each variable
- $\log(\text{OR}_{\text{hrsdm}}) = \beta_{\text{hrsdm}}$
- Score given by linear predictor  
score =  $x_1\beta_1 + x_2\beta_2 + \dots + x_p\beta_p$
- Calculate score for each person
- Use scores to distinguish those who die

# Score by Survival



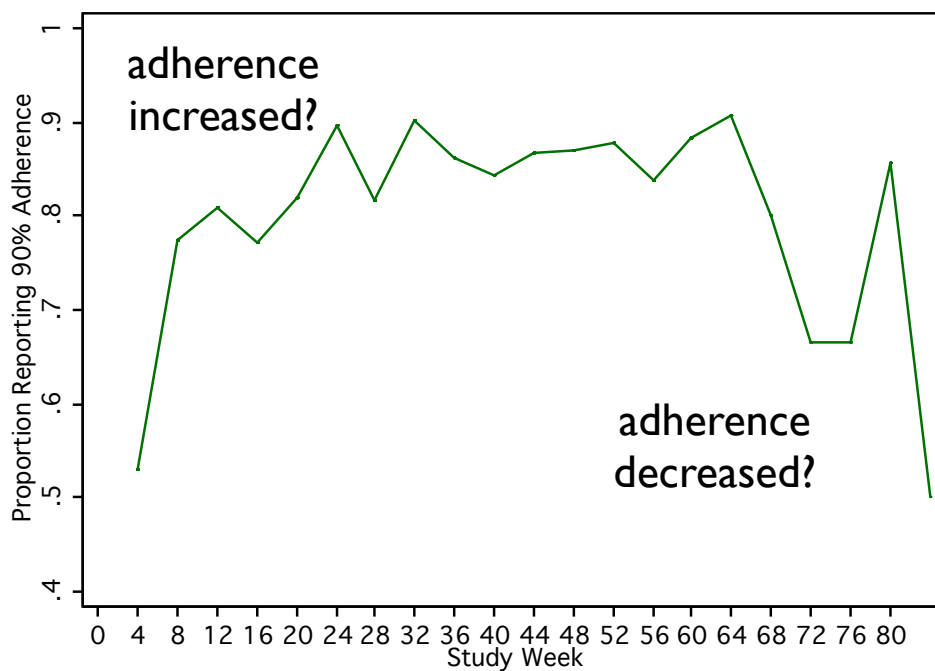
# ROC Curve



# Examining Adherence

- Clinical trial  
*daily oral medication*
- Every 4 weeks, adherence recalled  
*outcome: > 90% adherence*
- Calculate fraction by week
- What is pattern of adherence over time?

## Adherence by Week



# Profiling Tx Centers

- Transplant centers in US
- Reviewed for mortality outcomes  
*per federal regulations*
- Centers diverse  
*huge case-mix issues*
- Identify centers with high mortality

# Simulated Data

- 100 centers  
*20 patients each center*
- Widely varying mortality  
*0 to 30% of patients died*
- Overall set to 10% mortality  
none/minor variation across centers
- Calculate mortality at each center  
*set aside case-mix issues*

# Questions

- Is mortality model in elders good?
- Is adherence pattern accurate?
- Can we identify high mortality centers?
  - Reflect the truth underlying the data?
  - Can they be replicated?

## Model Replication

mortality model

Variable	Odds Ratio	
	test data	full data
hrsdm	2.9	1.8
bmicat	3.3	2.2
hrsmemory	45.7	2.7
hrsarth	7.0	1.3
hrscad	4.2	1.4
hrslung	4.8	2.3
meals	5.4	2.7
walkjog	9.5	3.8

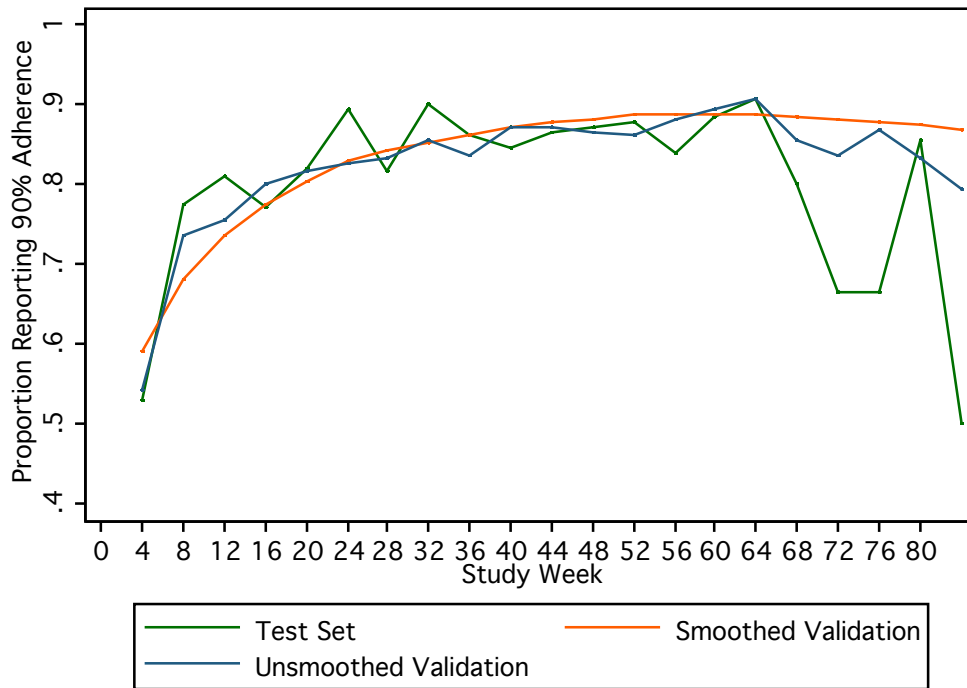
# Poor Validation

- Odds Ratios consistently overestimated  
consistent overestimation -> bias
- Many estimates outside CI
- More extreme OR -> more overestimated
- Artifact of model selection procedure  
*retained only if  $p < 0.05$*

# Testimation Bias

The combination of testing a hypothesis and then estimating a quantity, resulting in estimation of an effect which is stronger than is true. One example the process of estimating regression coefficients only among significant predictors.

# Adherence Pattern



## Results

- Early increase replicates  
*later decline doesn't*
- Many quirks don't check out  
*appear to be "noise"*
- Quirks suggest interesting questions
- Smoothing: chops of peaks, fills in valleys

# Overfitting

*Overfitting* is a broad term that refers to entering too many terms into the model -- for the amount of available data.

This produces unstable, variable estimates. The extreme ones *tend* to be both spurious and also *tend* to get the most focus.

Hence, results are frequently unreplicable.

# Regression to the mean

If you use a set of data to select a series of people and predictors which appear to be most extreme

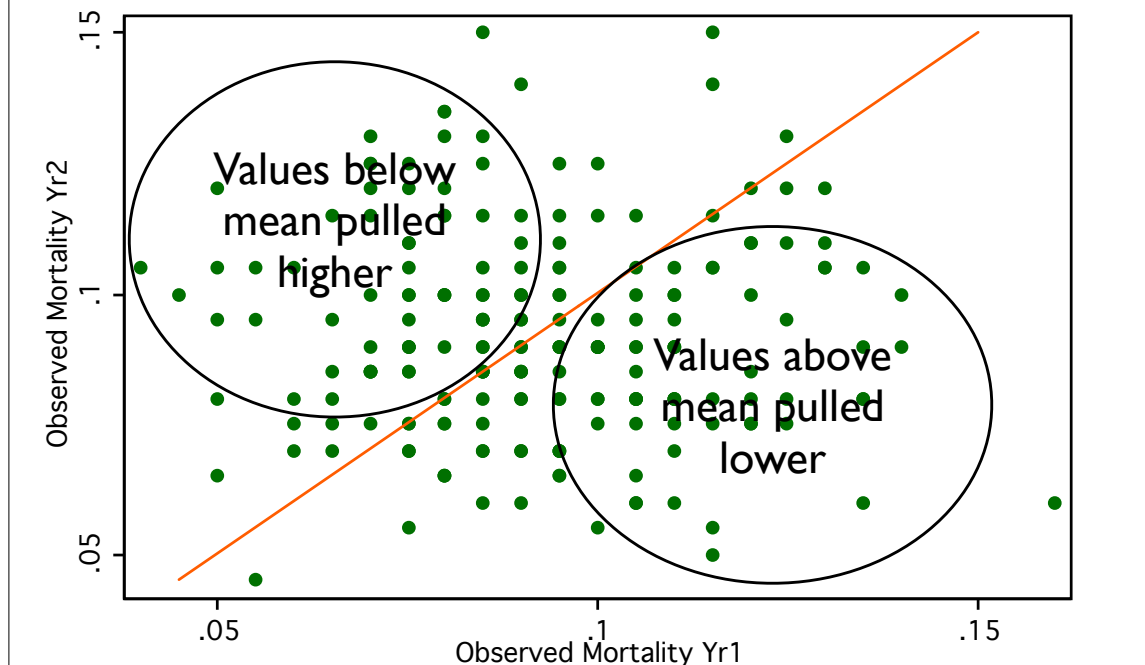
On re-measuring/validation these will tend, on average, to the mean of the population from which they came

Sometimes big, sometimes not but always a tendency

# Regression to the Mean

- Transplant Example
- More variation within than between centers
- Calculate mortality rate 2 years in row
- Compare

## Year 1 v. Year 2



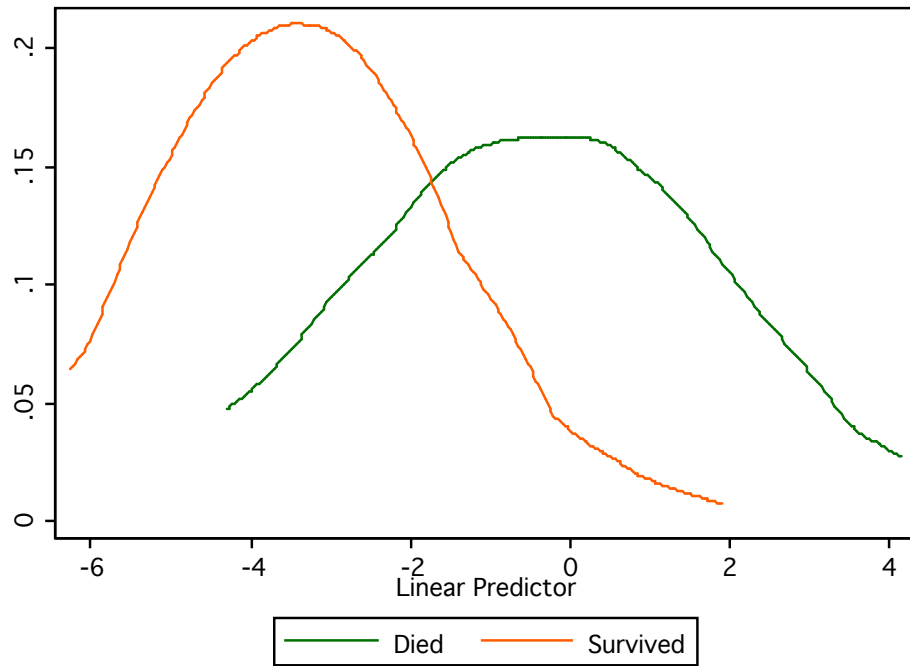
# Why is that?

- Data = signal + noise
- Hard to parse them in data
- Noise appears to fit data so well
- Can't identify patterns and test them well  
*model optimism*

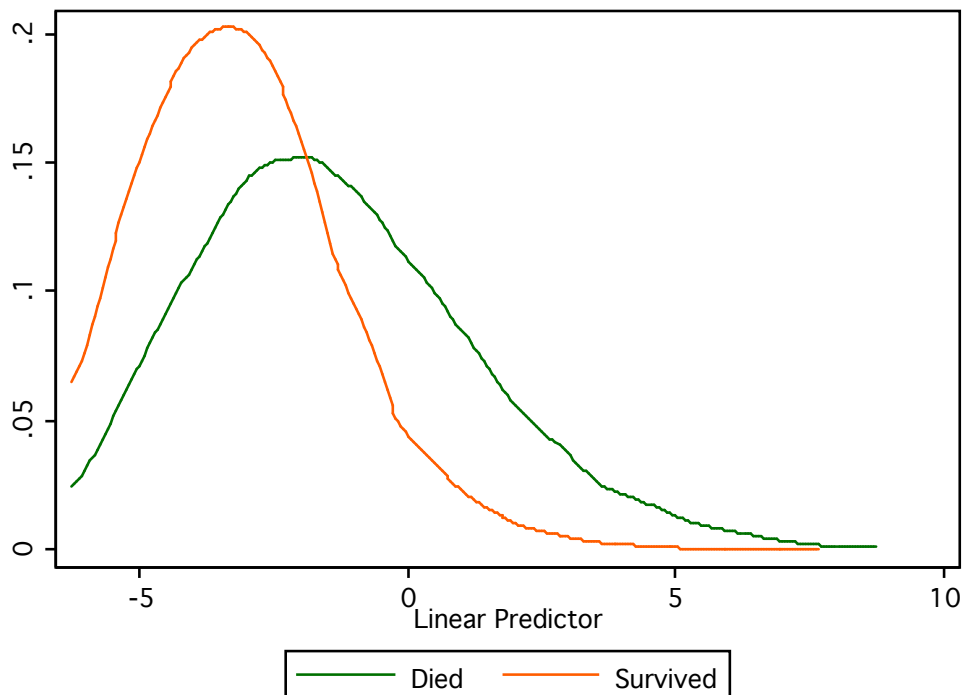
## Model Optimism

Models which have been suggested by the available data give consistently better measures of fit and agreement on data on which they were built compared with their true performance -- making it difficult to assess if we are overfitting to the available data.

# Score by Survival



# Score in Validation



# So What?

- Results can't be replicated
- Costing
  - money
  - time
  - energy
- Committing type I errors

The screenshot shows the PLOS MEDICINE website interface. At the top, there is a search bar and navigation links: Home, Browse Articles, About, For Readers, and For Authors and Reviewers. The article title is "Why Most Published Research Findings Are False" by John P. A. Ioannidis, marked as an "ESSAY" and "OPEN ACCESS". Below the title are tabs for "Article", "Related Content", and "Comments: 21". The "Abstract" section is visible, starting with the word "Summary". The abstract text discusses the increasing concern that most current published research findings are false, depending on study power, bias, and the ratio of true to no relationships. A sidebar on the right contains a "Jump to" menu with links to "Abstract", "Modeling the Framework for False Positive Findings", "Bias", "Testing by Several Independent Teams", "Corollaries", "Most Research Findings Are False for Most Research Designs and for Most Fields", "Claimed Research Findings May Often Be Simply Accurate Measures of the", and "Bi".

**PLOS MEDICINE**  
a peer-reviewed open-access journal published by the Public Library of Science

Search article

Home Browse Articles About For Readers For Authors and Reviewers

ESSAY OPEN ACCESS

## Why Most Published Research Findings Are False

Article Related Content Comments: 21

John P. A. Ioannidis

### Abstract [Top](#)

#### Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and

To add a note, highlight some text. [Hide notes](#)  
Make a general comment

**Jump to**  
[Abstract](#)  
[Modeling the Framework for False Positive Findings](#)  
[Bias](#)  
[Testing by Several Independent Teams](#)  
[Corollaries](#)  
[Most Research Findings Are False for Most Research Designs and for Most Fields](#)  
[Claimed Research Findings May Often Be Simply Accurate Measures of the](#)  
[Bi](#)

# How severe is bias?

- It depends
- Increases with
  - large number of predictors
  - large number of unassociated predictors
  - smaller sample size

# Mortality Data

- 46 variables with 46 deaths
- Search over large number of variables
- Likely most are related:  
*prior probably probably good*
- What is the effect of sample size?  
try 46 variables with 680 deaths

# Model Validation

dead	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
hrsfall	1.58984	.1713431	4.30	0.000	1.287111	1.96377
bath	1.997538	.3074196	4.50	0.000	1.477392	2.700811
couple	.6053514	.0601164	-5.05	0.000	.4982831	.735426
bmicat	2.338523	.2224735	8.93	0.000	1.940722	2.817864
hrslung	1.882528	.3194919	3.73	0.000	1.349834	2.625442
hrsCHF	2.315742	.469566	4.14	0.000	1.556291	3.445794
money	1.777714	.2574839	3.97	0.000	1.338363	2.361293
map	1.143892	.1337196	1.15	0.250	.9096634	1.438433
shop	1.380721	.1924961	2.31	0.021	1.050591	1.814588
hrsca	2.082971	.2482573	6.16	0.000	1.649048	2.631074
hrsdm	1.528729	.18003	3.60	0.000	1.213638	1.925625
climb	1.587402	.1873359	3.92	0.000	1.259601	2.00051
hrspain	.6914175	.0833919	-3.06	0.002	.5458545	.8757979
walkjog	2.735907	.328657	8.38	0.000	2.161966	3.462214
gender	.4782791	.0488091	-7.23	0.000	.3915751	.5841815

Area under ROC = 0.81

# Strongest Predictors

Variable	Odds Ratio	
	test data	validation data
bmicat	2.1	2.3
hrsca	2.6	2.1
gender	2.5	2.1
walkjog	2.0	2.7

# Adherence Issues

- Test data: small sample size  
*n=100 week 4, n=58 week 40, n=7 week 80*
- 20 parameters to be estimated
- Adherence likely similar at nearby weeks  
*motivation for smoothing*

# Assessing Overfitting and Optimism

# Resampling Methods

- Modern methods for complex problems
- Involve creating “random” datasets
- Two important ones: bootstrap, cross-validation

## Cross-Validation

- Correcting for model-based optimism
- Split data into a series of random subsets  
*usually 10 of them*
- one-at-time, leave subset out
- Build model on 90%  
*use other 10% for validation*

# Stata Code

- `gen u = uniform()`  
*makes random # and puts in "u"*
- `xtile cat=u, nq(10)`  
*creates "cat" 10 groups based on "u"*
- `logistic dead predictors if cat!=k`  
*xtile cat l = u if dead == l, nq(10)*
- `predict prediction if cat==k`  
*save prediction for kth group*

# How it works

- Uses data to build and validate model
- 90% builds model, 10% validates it
- Every observation gets used in 1 test set
- Provides an internal validation

# Bootstrap

- Data of sample size  $n$
- Repeatedly sample a dataset of size “ $n$ ”
- From original data
- Sampled “with replacement”  
*Some obs sampled > 1 in a dataset  
others completely left out*
- Seems ad-hoc:  
*justification very mathematical*

# Using Bootstrap

- Allows us to calculate uncertainty without using complex math
- Helps to calculate standard error
- But can also assess and correct for biases
- Very useful technique

# Bootstrap for Prediction

- Building a model based on data
- Draw bootstrap (BS) sample
- Build predictive model on BS sample
- Predict original (not resampled) data

# Mitigating Overfitting

- Limit the scope of the model  
*fewer predictors, fewer  $df$*
- Pull in most extreme coefficients
- smoothing and shrinkage
- Discussed in future lectures