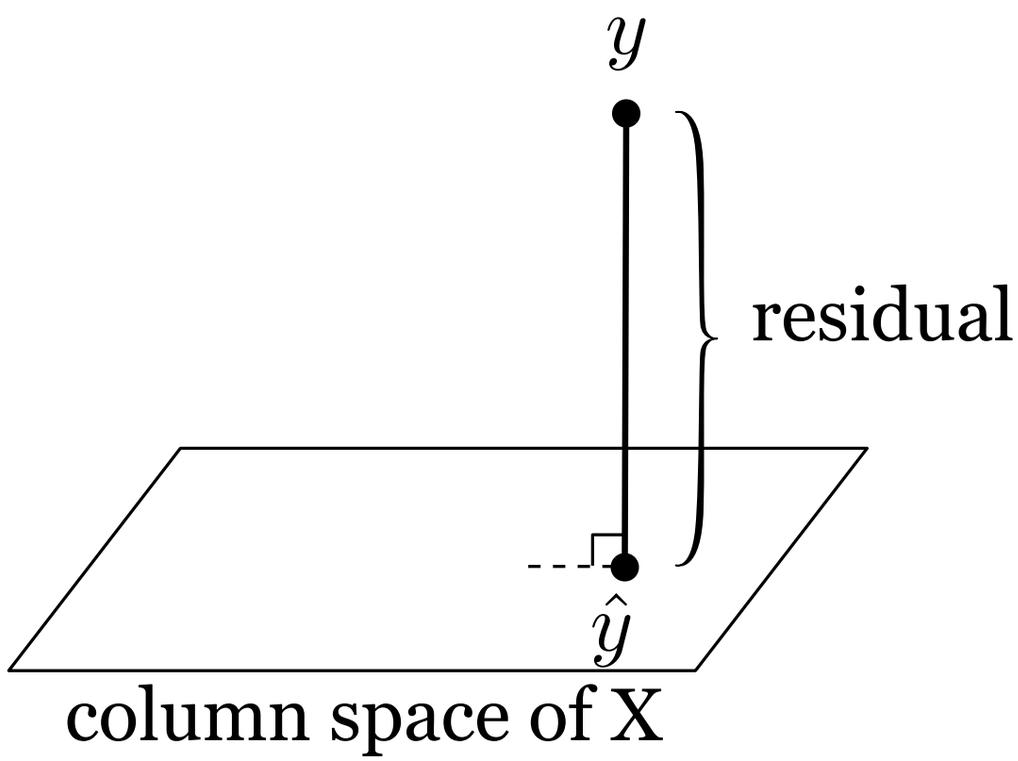
CS109/Stat121/AC209/E-109
Data Science
Bias and Regression

Hanspeter Pfister, Joe Blitzstein, and Verena Kaynig



This Week

- HW1 due tonight at11:59 pm (Eastern Time)
- HW2 posted soon

Census: Everybody's moving into their parents' basements

By Brad Plumer June 20, 2012 Follow @bradplumer

Ever since the financial crisis hit, Americans have found it harder and harder to live on their own. According to a new report (pdf) from the Census Bureau, the number of "shared households" increased by a whopping 2.25 million between 2007 and 2010:

Daniel Sherrett, 28, prepares dinner with his mother as part of his deal to live at home. Parents and children are sharing homes for longer than expected. (Michael Temchine/The Washington Post)

> In spring 2007, there were 19.7 million shared households. By spring 2010, the number of shared households had increased by 11.4 percent, while all households increased by only 1.3 percent.

This number does not include co-habitating or married couples. Rather, it's specifically a measure of the growing fraction of Americans who are either living with roommates or shacking up with relatives. And the bulk of the increase came from kids who are living at home with their parents: "Between 2007 and 2010, the number of adult children who resided in their parents'

Most Read Business

1 Here is everything we know about whether gentrification pushes poor people out 

2 Honey isn't as healthy as we think 

3 These are the hardest places for minimum wage workers to live 

4 Why Americans dress so casually 

5 Is your adviser truly protecting your retirement? 

The Most Popular All Over

The Atlantic
The Rise of Victimhood Culture 

Magazine

1657 COMMENTS

It's Official: The Boomerang Kids Won't Leave

By ADAM DAVIDSON JUNE 20, 2014

Census Data from the Current Population Survey (CPS)

"It is important to note that the CPS counts students living in dormitories as living in their parents' home."

– Census Bureau, http://www.census.gov/prod/2013pubs/p20-570.pdf

publication bias: replication issues
p-values: need $p < 0.05$ to publish

Some Forms of Bias

- selection bias
- publication bias (file drawer problem)
- non-response bias
- length bias

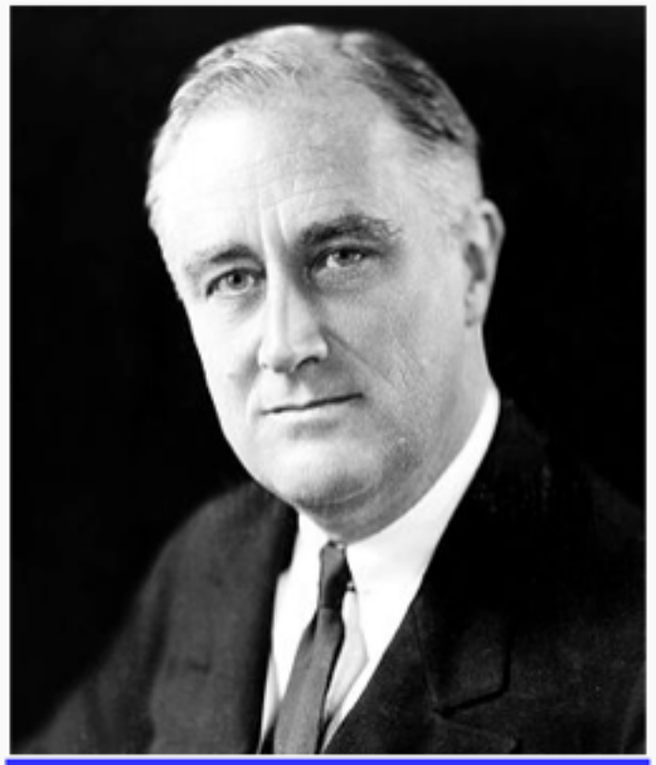
systemic differences between those who respond and don't respond to a survey.

1936 Presidential Election, Landon vs. FDR

1932 ← **November 3, 1936** → 1940

531 electoral votes of the Electoral College

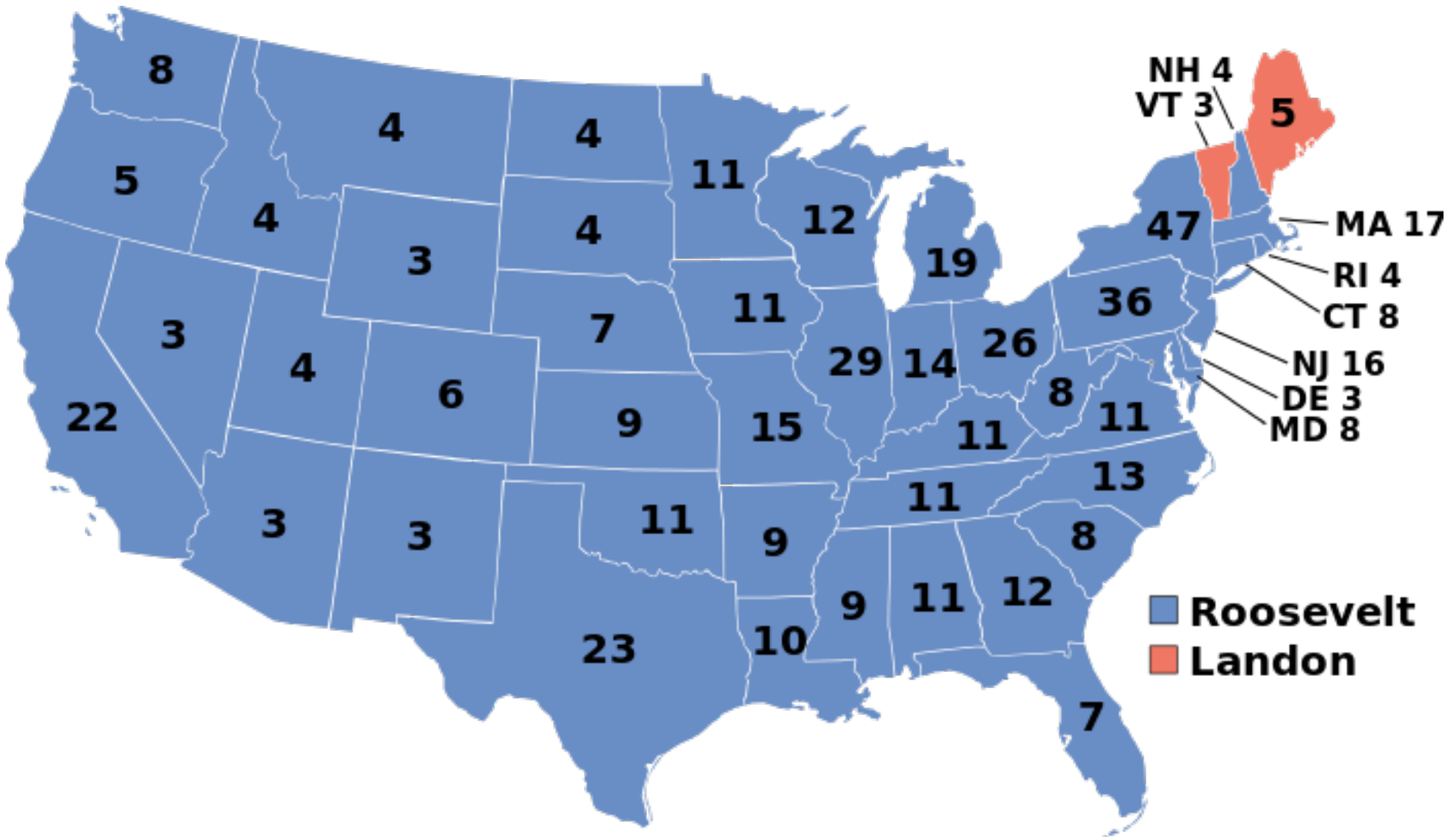
266 electoral votes needed to win



Nominee	**Franklin D. Roosevelt**	Alf Landon
Party	Democratic	Republican
Home state	New York	Kansas
Running mate	**John N. Garner**	Frank Knox

1936 Presidential Election, Landon vs. FDR

Literary Digest predicted Landon
would win with 370 electoral votes,
based on sample size of 2.4 million.



source: https://en.wikipedia.org/wiki/United_States_presidential_election,_1936

1936 Presidential Election, Landon vs. FDR

Literary Digest got responses from 2.3 million out of 10 million people surveyed.

To collect their sample, they used 3 readily available lists:

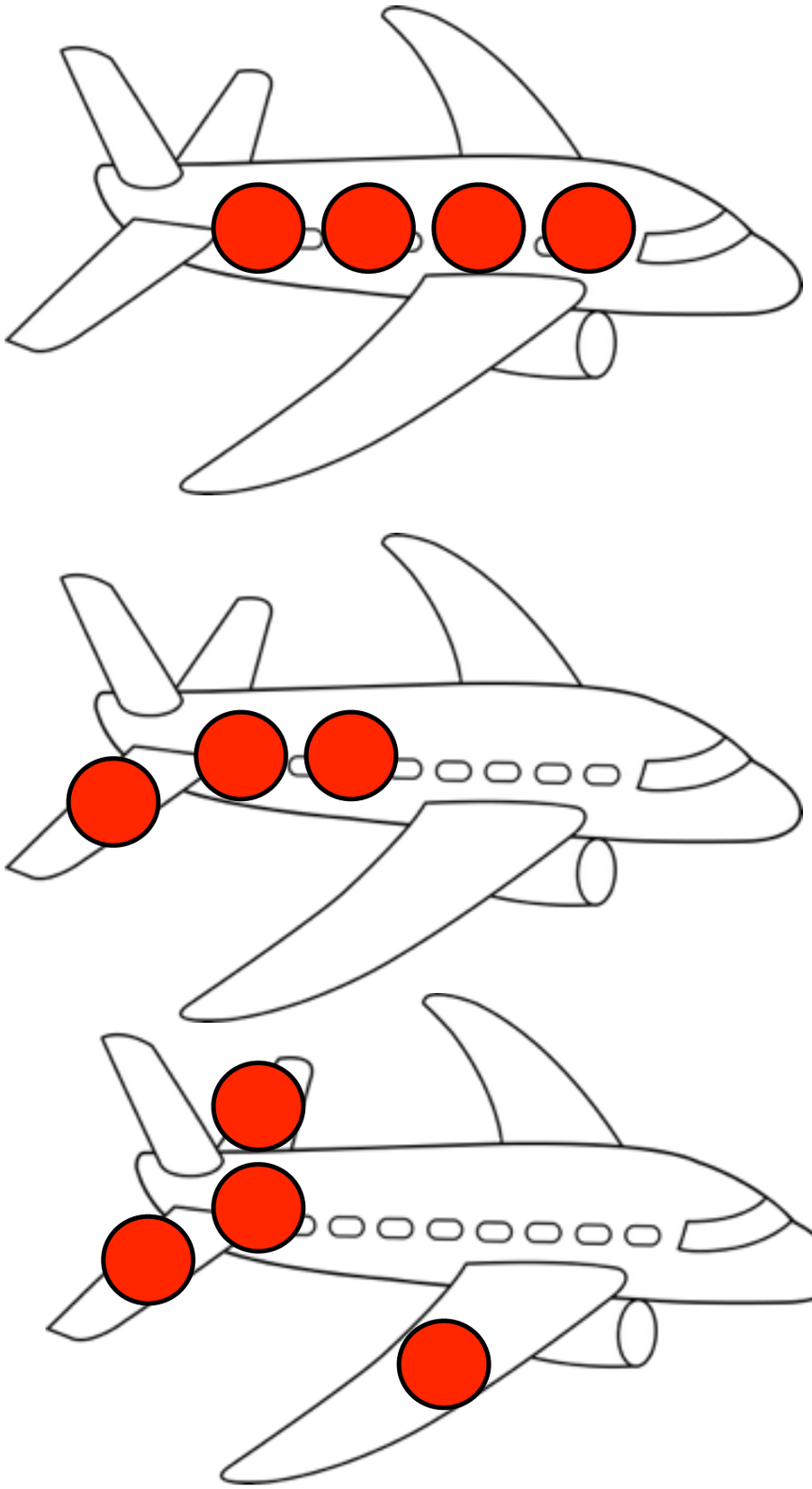
- readers of their magazine
- car registration list
- phone directory

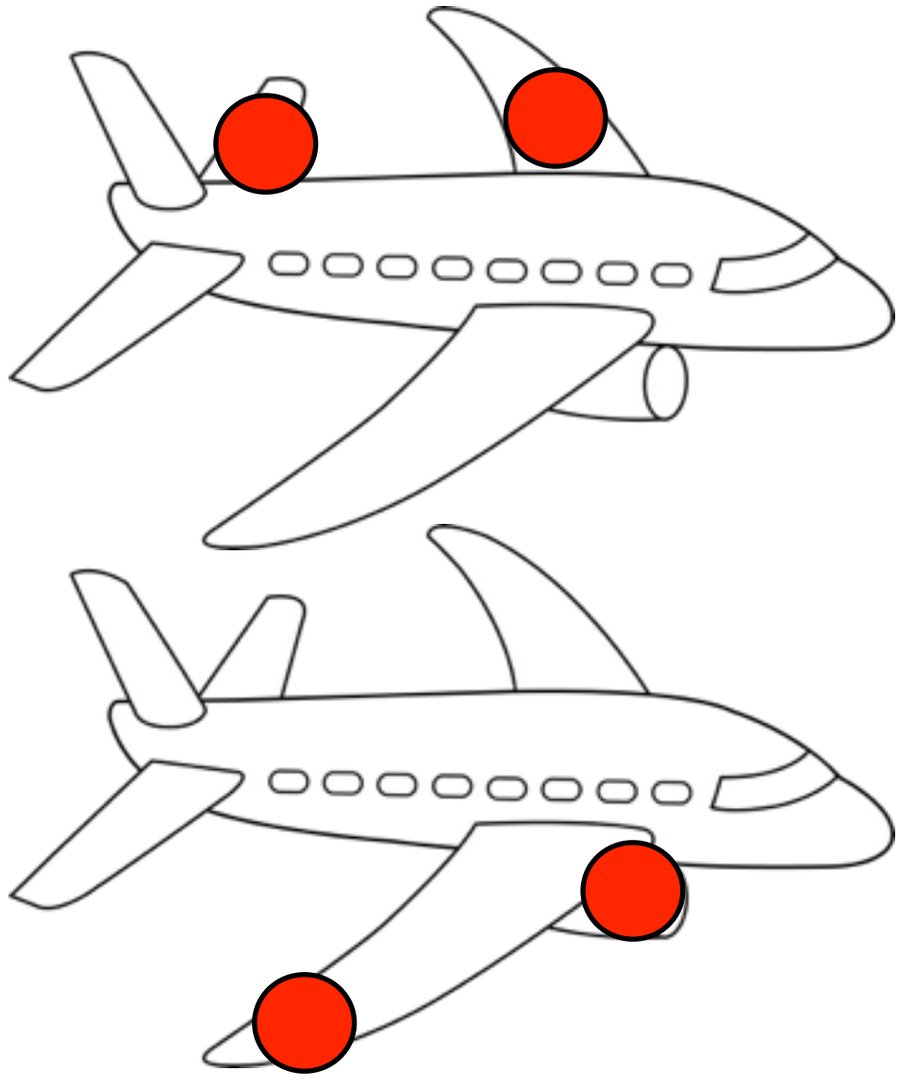
1936, most people don't own cars, mostly rich ppl.

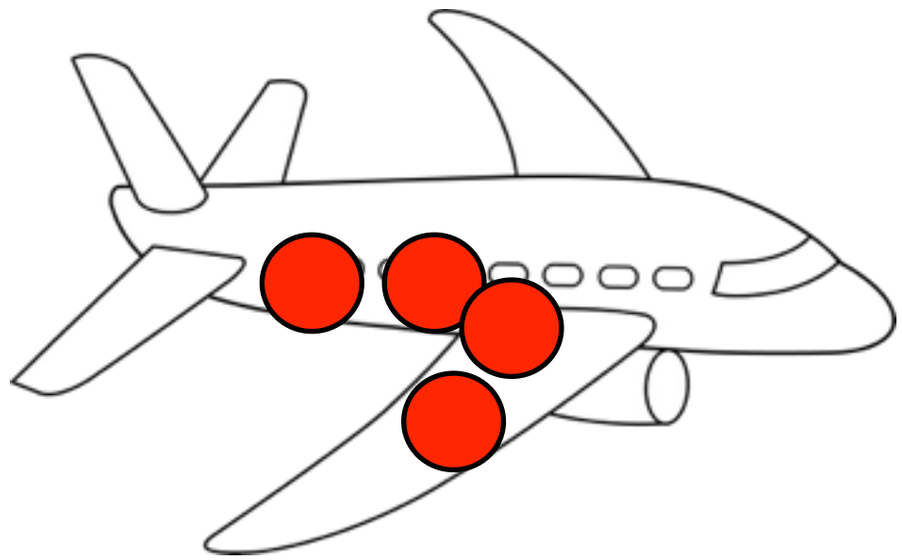
huge non-response bias
selection bias: literary digest readers are disproportionally conservative.

Wald and the Bullet Holes

On airplanes, where to put on armor?







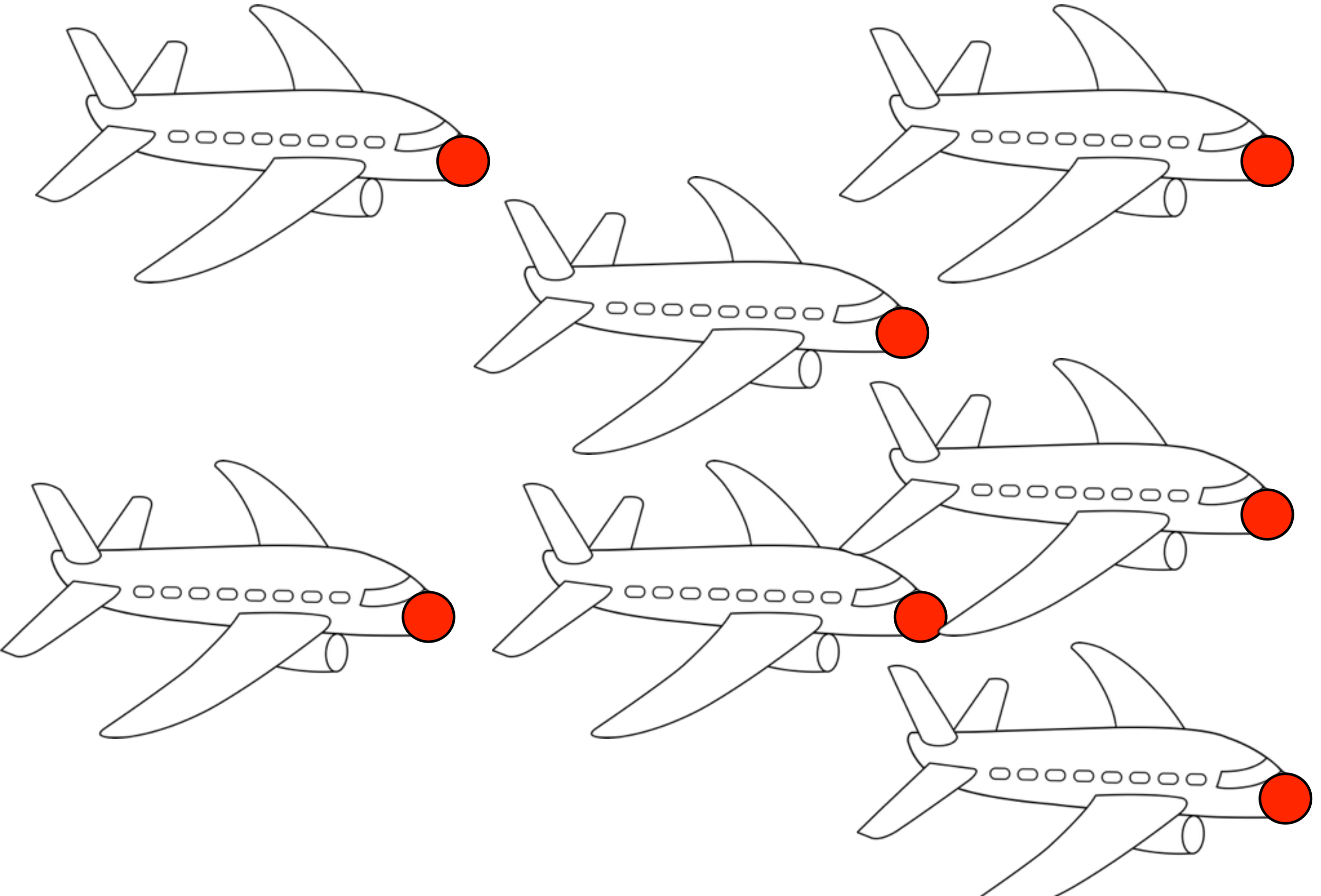
What about the *unobserved* planes? Missing data!

Planes survived with this damage!

Reinforce areas not hit - they didn't survive hits to those areas

- this is unobserved data.

What about the *unobserved* planes? Missing data!



Longevity Study from Lombard (1835)

Profession	Average Longevity
chocolate maker	73.6
professors	66.6
clocksmiths	55.3
locksmiths	47.2
students	20.2

Student is a transient profession: ends when you're young (but you don't die)
but this applies to other jobs: retirement differentials? turnover?

Sources: Lombard (1835), Wainer (1999), Stigler (2002)

Class Size Paradox

Why do so many schools boast small
average class size but then so many students
end up in huge classes?

Simple example: each student takes one course;
suppose there is one course with 100 students,
fifty courses with 2 students.

Dean calculates: (100+50*2)/51 = 3.92

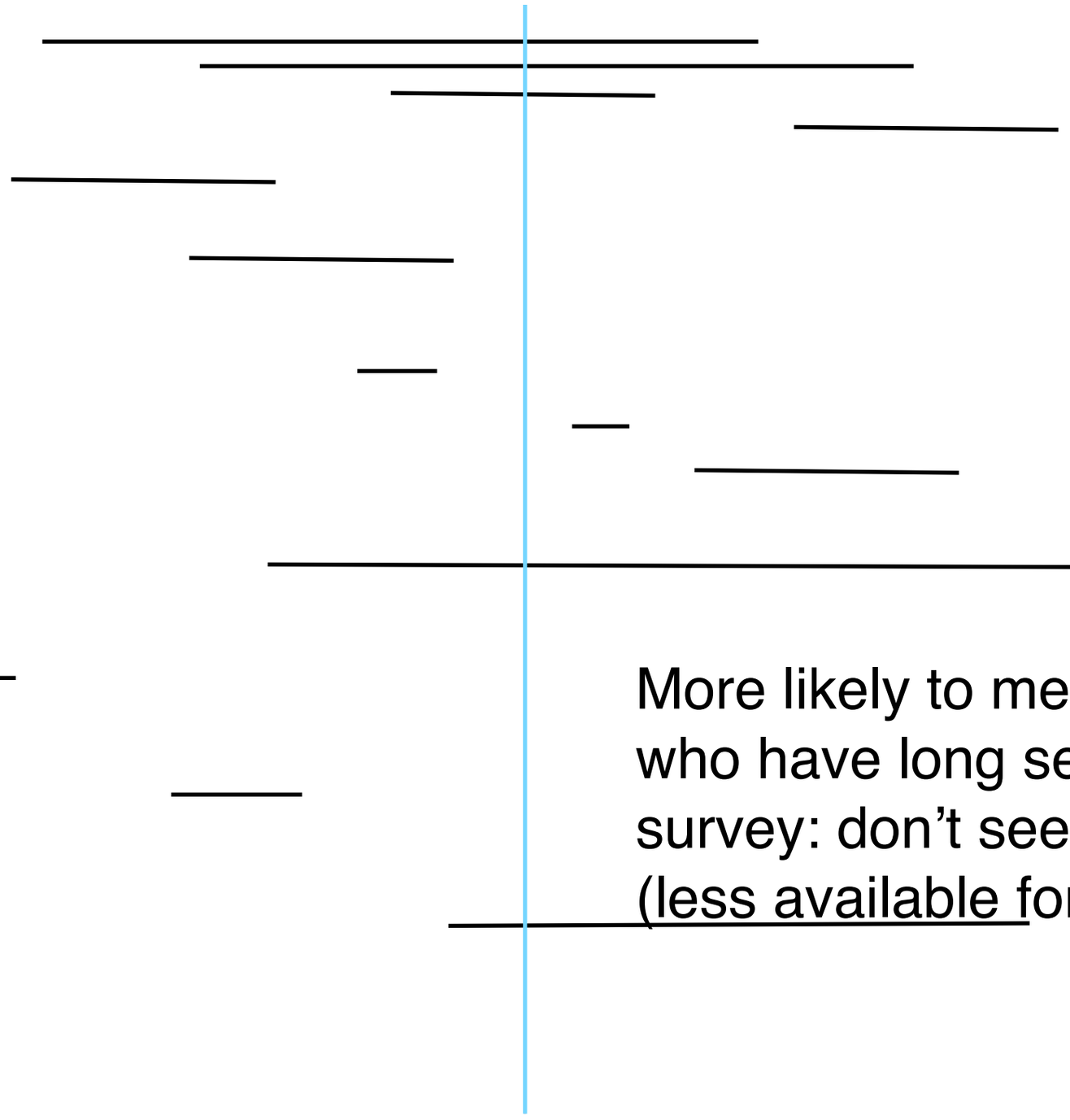
Students calculate: (100*100+100*2)/200 = 51
much higher

“About 10 percent of the 1.6 million inmates in America’s prisons are serving life sentences; another 11 percent are serving over 20 years.”

source: http://www.nytimes.com/2012/02/26/health/dealing-with-dementia-among-aging-criminals.html?pagewanted=all

Length-Biasing Paradox

How would you measure the average prison sentence?



each line is a prisoners sentence.

More likely to measure the ones who have long sentences through survey: don't see short sentences (less available for survey)

Bias of an Estimator

The *bias* of an estimator is how far off it is on average:

$$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

avg estimator

true value

So why not just subtract off the bias?

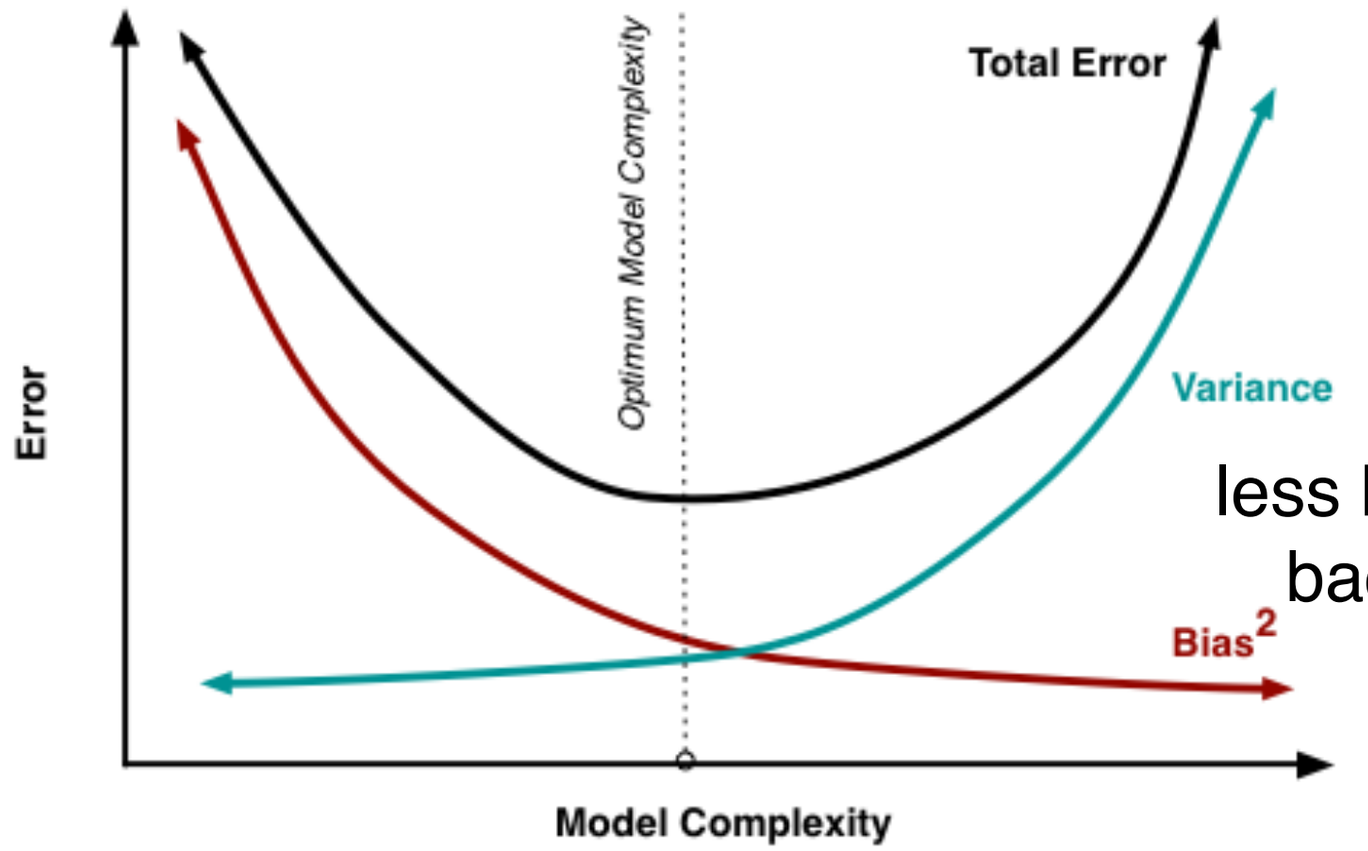
You don't know the true value …

Bias-Variance Tradeoff

one form: $$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{bias}^2(\hat{\theta})$$

often a little bit of bias can make it possible to have much lower MSE

more bias, more precession



less bias -> high var
bad …

Text

http://scott.fortmann-roe.com/docs/BiasVariance.html

Unbiased Estimation: Poisson Example

$$X \sim \text{Pois}(\lambda)$$ counting

$$\text{Goal: estimate } e^{-2\lambda}$$

$(-1)^X$ is the best (and only) unbiased estimator of $e^{-2\lambda}$

sensible?

ridiculous: Pois is restricted to (+) value, but (-1)^x can achieve negative values

Fisher Weighting

How should we combine independent, *unbiased* estimators for a parameter into one estimator?

$$\hat{\theta} = \sum_{i=1}^{k} w_i \hat{\theta}_i$$

The *weights* should sum to one, but how should they be chosen?

more weight on estimators that are reliable: measure by variance
more reliable: less variance

$$w_i \propto \frac{1}{\mathrm{Var}(\hat{\theta}_i)}$$

(Inversely proportional to variance; why not SD?)

Nate Silver Weighting Method

- Exponential decay based on recency of poll
- Sample size of poll
- Pollster rating

http://fivethirtyeight.com/features/how-the-fivethirtyeight-senate-forecast-model-works/

Nate Silver good at thinking of where data come from, how reliable are polls, and how to combine the data.
—> combine by weighting: ie weigh recent polls more heavily: uses an exp decay
bigger sample: weigh that poll more
weigh by bias of polls: pollster rating
ie, how do they find ppl (selection bias - phone, online, etc.)
do they lean in a political direction?
used previous election cycles data where available.

Multiple Testing, Bonferroni

Just by chance: many variables -> significant correlations among at least some of them (ie at p< 0.05, you get significance randomly for 5% of coefficients, more coefficients, more significant ones.

Issue with larger datasets

How should we handle p-values when testing multiple hypotheses?

For example, what if we are looking at diet (with 10 kinds of food) and disease (with 10 diseases)?

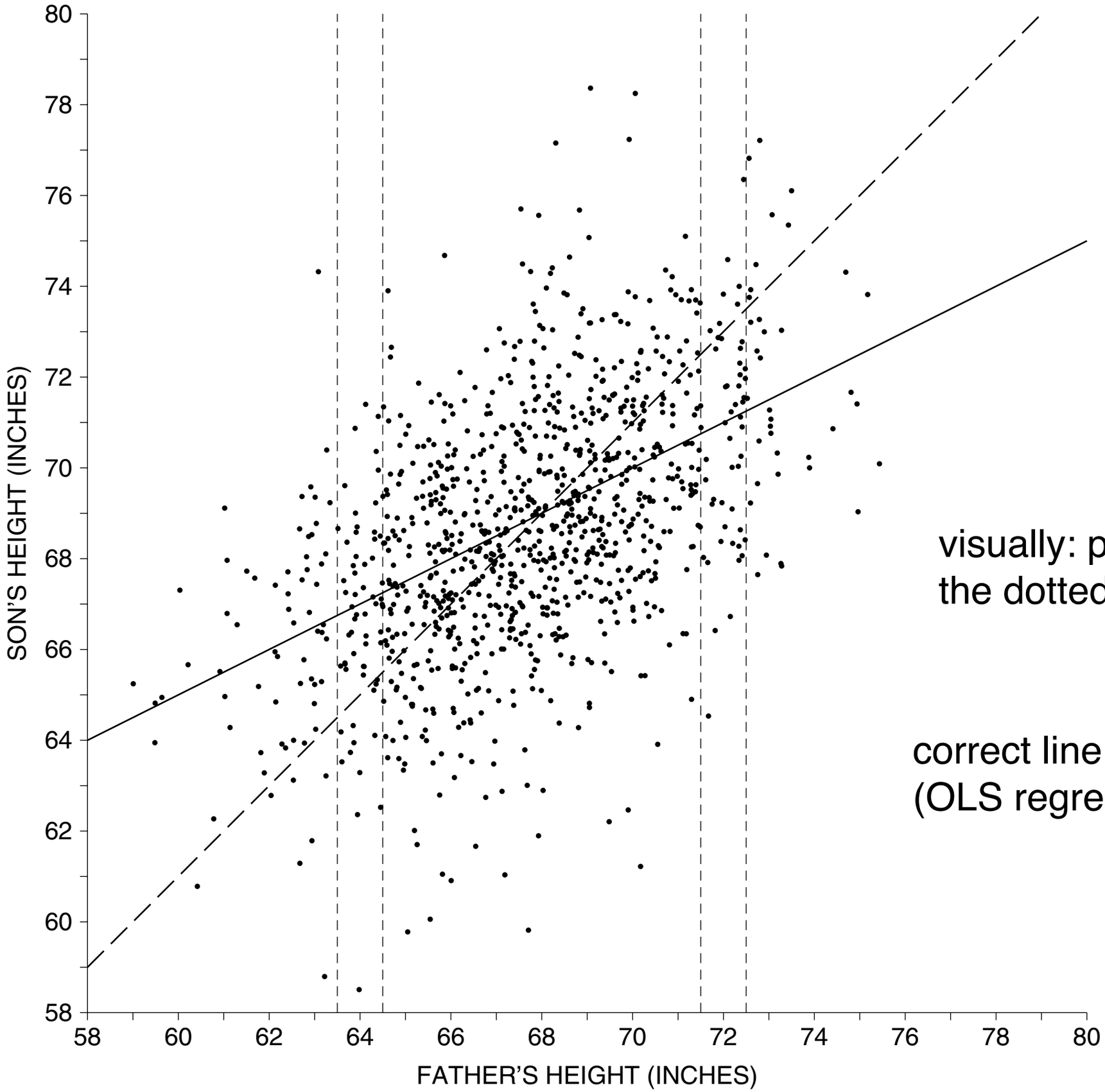
significance level: alpha = 0.05 divide by # of hypotheses

A simple, conservative approach is Bonferroni: divide significance level by number of hypotheses being tested.

$$FWER = Pr\left\{\bigcup_{I_o}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{I_o}\left\{Pr\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq m_0\frac{\alpha}{m} \leq m\frac{\alpha}{m} = \alpha$$

family-wise error rate: multiple hypotheses form a family

https://en.wikipedia.org/wiki/Bonferroni_correction



visually: ppl will draw the dotted lines

correct line: solid line (OLS regression)

plot from Freedman, data from Pearson-Lee

Regression Toward the Mean (RTTM)

Examples are everywhere...

Test scores
Sports
Inherited characteristics, e.g., heights
Traffic accidents at various sites

ie, child of tall father more likely to be taller than average, but is also likely
to be closer to average -> regression towards mean

intuition: everything combination of luck and skill: luck fluctuates and averages out
(assuming luck is mean 0)-> regress towards mean
(skill advantage determines deviation from mean).

Daniel Kahneman Quote on RTTM

regression towards mean is where word regression came into statistics

> I had the most satisfying Eureka experience of my career while attempting to teach flight instructors that praise is more effective than punishment for promoting skill-learning....
>
> [A flight instructor objected:] "On many occasions I have praised flight cadets for clean execution of some aerobatic maneuver, and in general when they try it again, they do worse. On the other hand, I have often screamed at cadets for bad execution, and in general they do better the next time. So please don't tell us that reinforcement works and punishment does not..."
>
> This was a joyous moment, in which I understood an important truth about the world: because we tend to reward others when they do well and punish them when they do badly, and because there is regression to the mean, it is part of the human condition that we are statistically punished for rewarding others and rewarded for punishing them.

Regression Paradox

y: child's height (standardized)
x: parent's height (standardized)

Regression line: predict y = rx;
think of this as a weighted average of
the parent's height and the mean

Now, what about predicting the parent's height from the child's height? Use x = y/r?

Regression line is x = ry, the r stays the same!

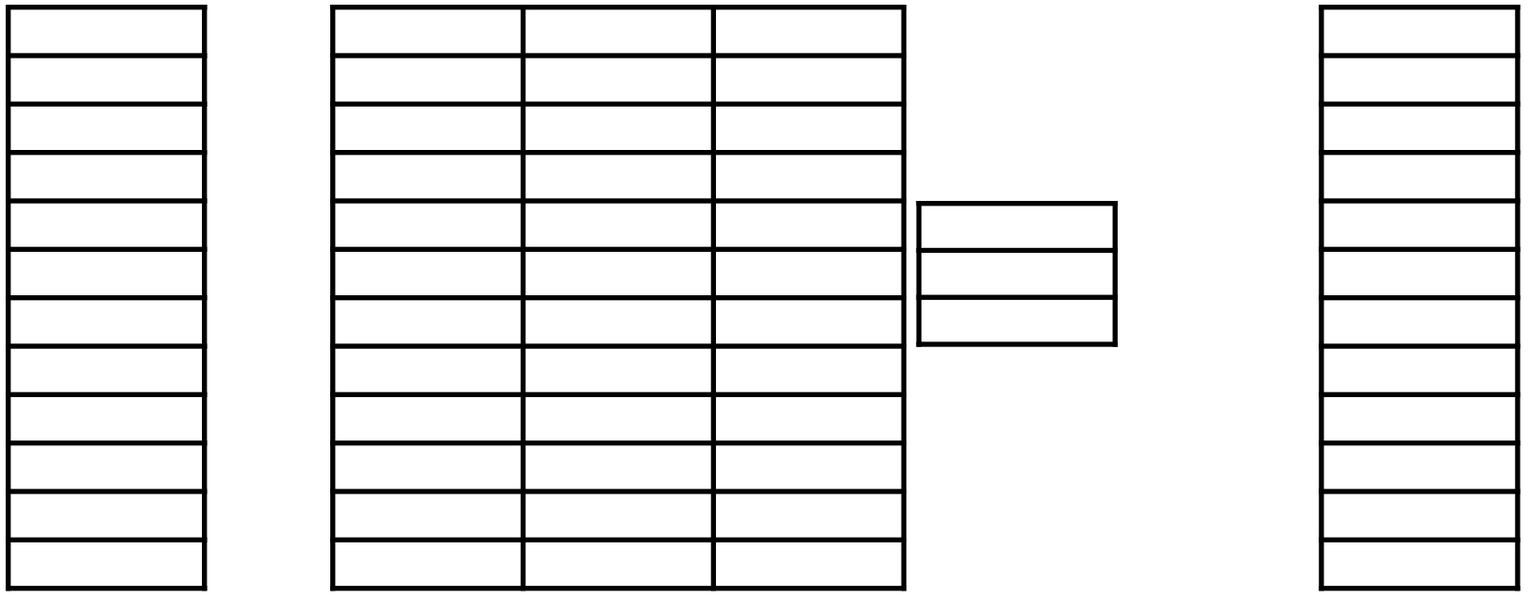
r is correlation: same regardless of direction

Linear Model

often called "OLS" (ordinary least squares), but that puts the focus on the procedure rather than the model.

$$\underbrace{y}_{n\times 1} = \underbrace{X}_{n\times k}\underbrace{\beta}_{k\times 1} + \underbrace{\epsilon}_{n\times 1}$$

error term:
assume i.i.d mean 0



What's linear about it?

$$\underbrace{y}_{n\times 1} = \underbrace{X}_{n\times k}\underbrace{\beta}_{k\times 1} + \underbrace{\epsilon}_{n\times 1}$$

Linear refers to the fact that we're taking linear combinations of the predictors. Still linear if, e.g., use both x and its square and its cube as predictors.

x^2 or x^3 data would still go in as a column -> apply **linear** algebra still get linear combinations and a linear regression.

want more predictors than you have data?

Sample Quantities vs. Population Quantities

sample version (think of x and y as data vectors)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

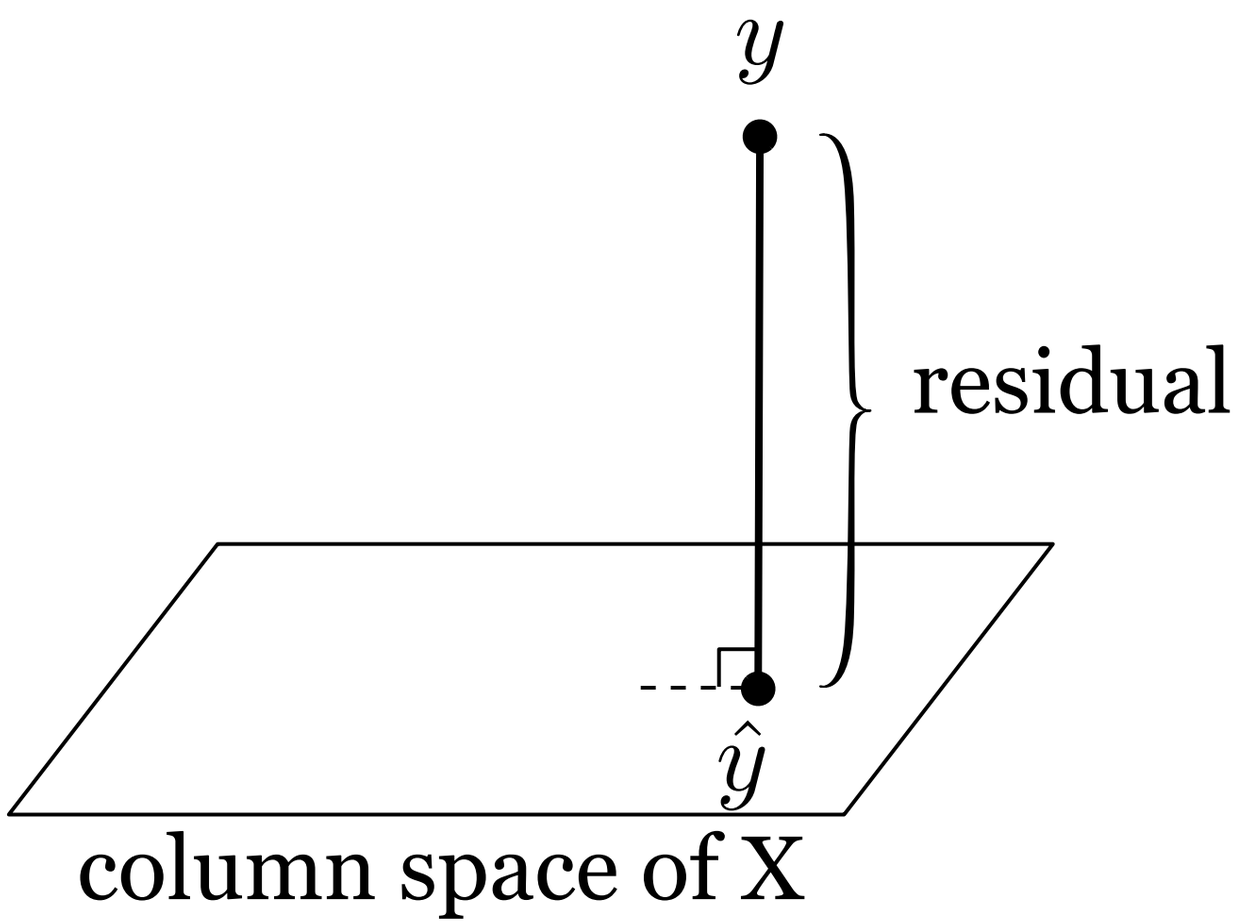
population version (think of x and y as r.v.s)

$$y = \beta_0 + \beta_1 x + \epsilon$$

$$E(y) = \beta_0 + \beta_1 E(x)$$

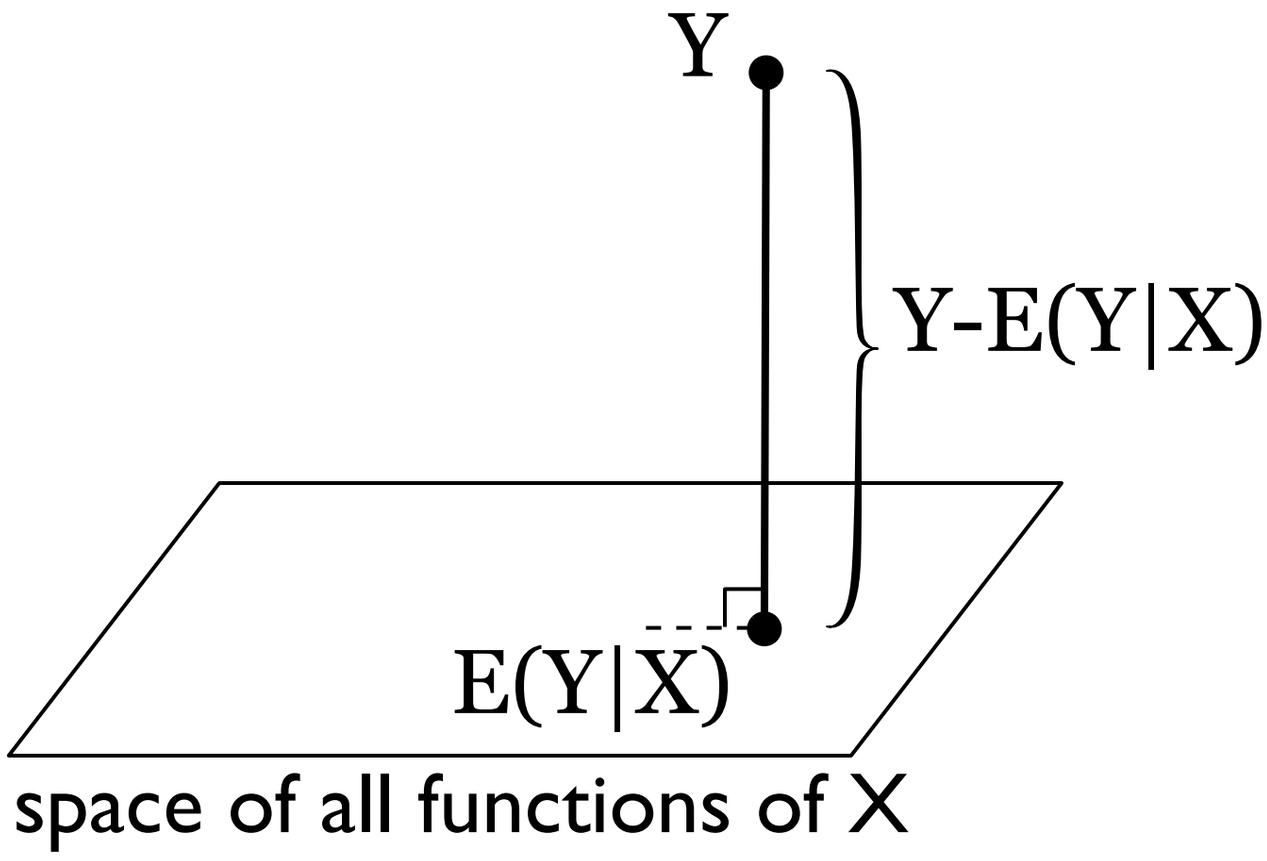
$$\text{cov}(y, x) = \beta_1 \text{cov}(x, x)$$

visualize regression as a *projection*



X is data matrix
column space is all vectors resulting of linear combos of columns of X (oh, duh)
what combination of X gets you closest to y (minimizes size of residuals)

or as a *conditional expectation*



Normal distributions (error terms i.i.d. Normal) -> E[YIXI is linear (minimizes MSE)

Gauss-Markov Theorem

Consider a linear model

$$y = X\beta + \epsilon$$

where y is n by 1, $\mathbf{X}$ is an n by k matrix of covariates, β is a k by 1 vector of parameters, and the errors ϵ_j are uncorrelated with equal variance, $\epsilon_j \sim [0, \sigma^2]$. The errors do not need to be assumed to be Normally distributed.

Then it follows that...

$$\hat{\beta} \equiv (X'X)^{-1}X'y$$

is **BLUE** (the **B**est **L**inear **U**nbiased **E**stimator).

For Normal errors, this is also the MLE.

Residuals

$$y = X\hat{\beta} + e$$

mirrors

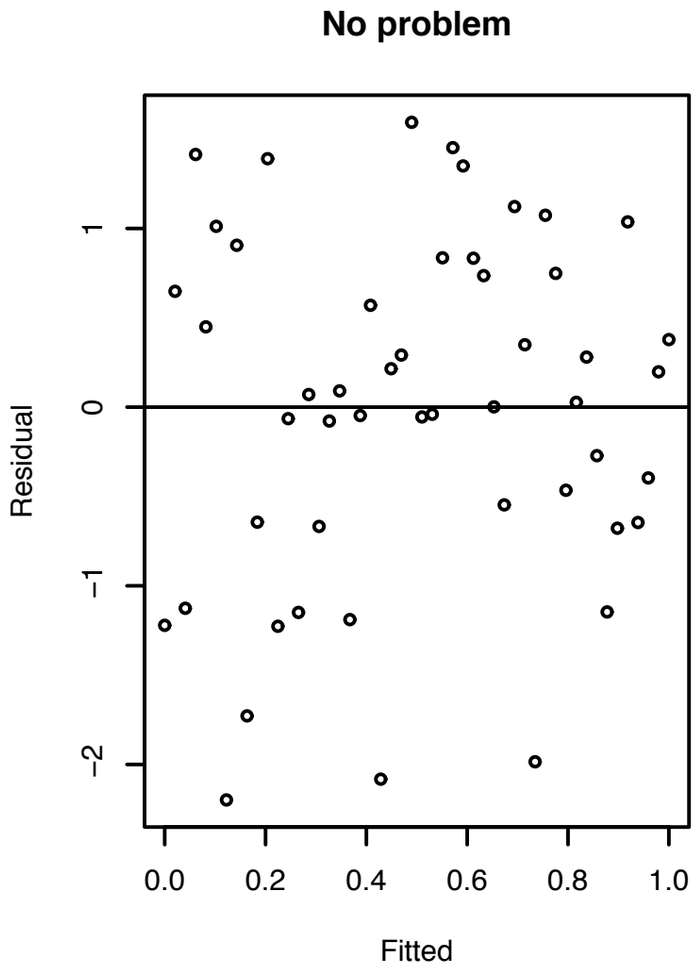
$$y = X\beta + \epsilon$$

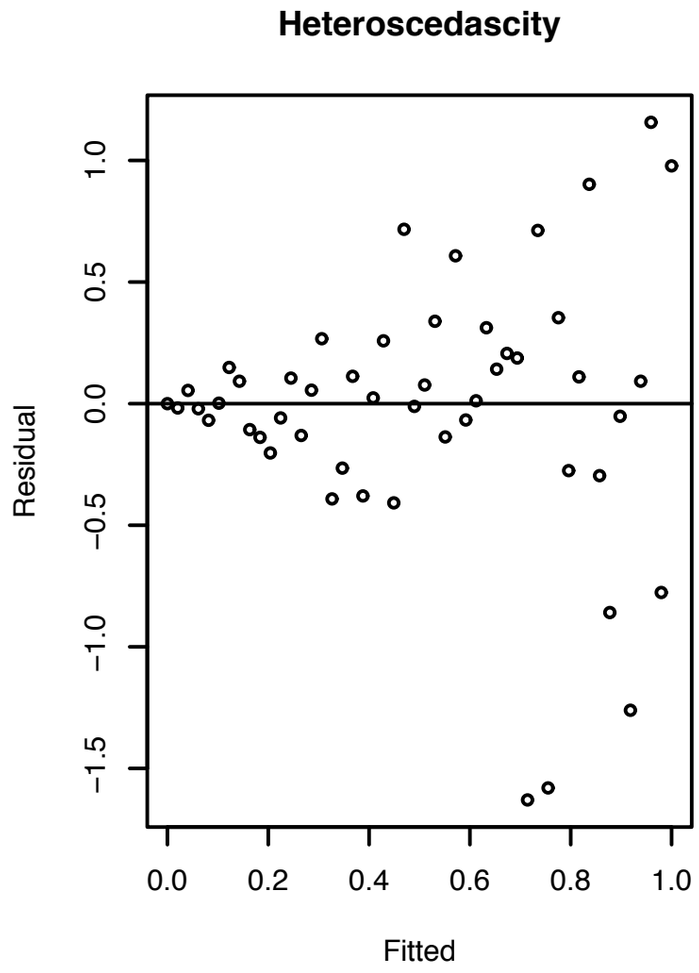
The residual vector e is *orthogonal* to all the columns of X.

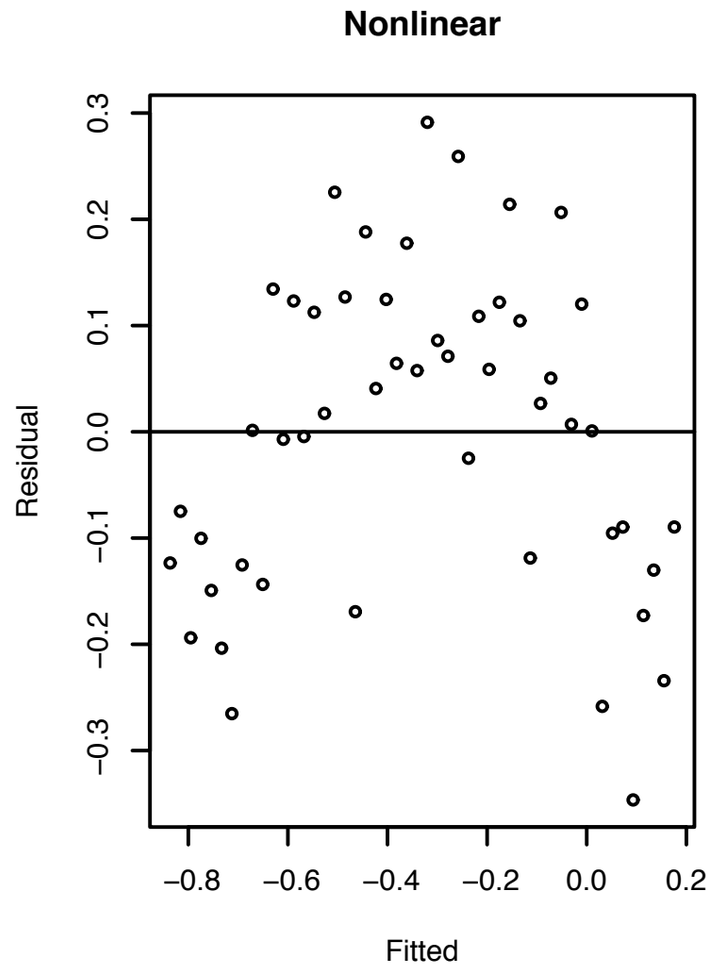
Never going to know the true epsilon - unobservable
We know X and y, but only have B-hat estimate for coefficient (by OLS)
and then solve for epsilon -> estimate of residuals

Residual Plots

Always plot the residuals! (Plot residuals vs. fitted values, and residuals vs. each predictor variable)







Faraway, http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf

"Explained" Variance

R^2 = how much of variance captured by model vs total amount of variation

$$\mathrm{var}(y) = \mathrm{var}(X\hat{\beta}) + \mathrm{var}(e)$$

$$R^2 = \frac{\mathrm{var}(X\hat{\beta})}{\mathrm{var}(y)} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

R^2 measures goodness of fit, but
it does *not* validate the model.

Adding more predictors can only increase R^2.

but, there is an adjusted R squared that removes the benefit due to adding more regressors

considered a crude fix: prefer cross validation