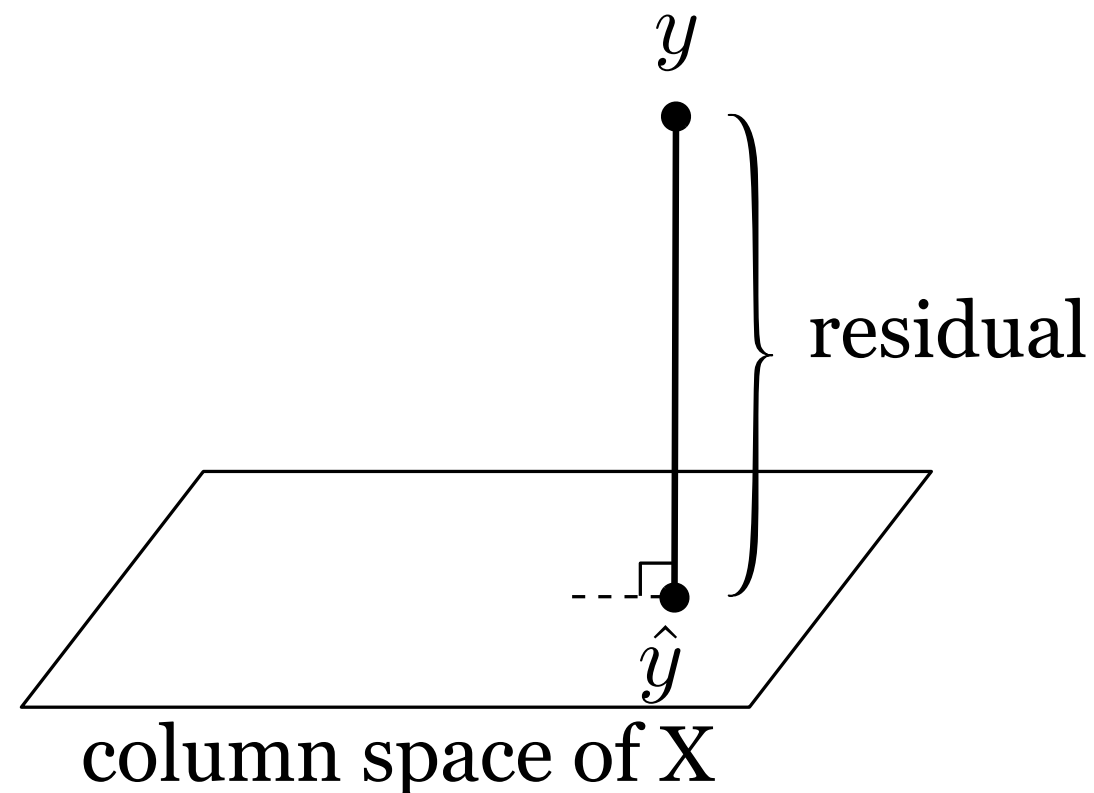


CS109/Stat121/AC209/E-109

# Data Science

## Bias and Regression

Hanspeter Pfister, Joe Blitzstein, and Verena Kaynig



# This Week

- HW1 due tonight at 11:59 pm (Eastern Time)
- HW2 posted soon

# Census: Everybody's moving into their parents' basements

A



0

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:

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.



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)

This number does not include co-habiting 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



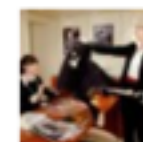
- 2 Honey isn't as healthy as we think



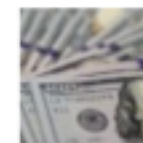
- 3 These are the hardest places for minimum wage workers to live



- 4 Why Americans dress so casually

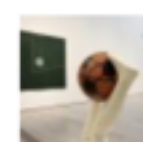


- 5 Is your adviser truly protecting your retirement?



## The Most Popular All Over

The Atlantic  
The Rise of Victimhood Culture





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

By ADAM DAVIDSON JUNE 20, 2014



SLIDE SHOW | 14 Photos

'Hi, Mom, I'm Home!'



# 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>

If you're not careful, you'd misconstrue the previous example and not see that in actuality there're more college students rather than more kids staying at home.

# Some Forms of Bias

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

# 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**

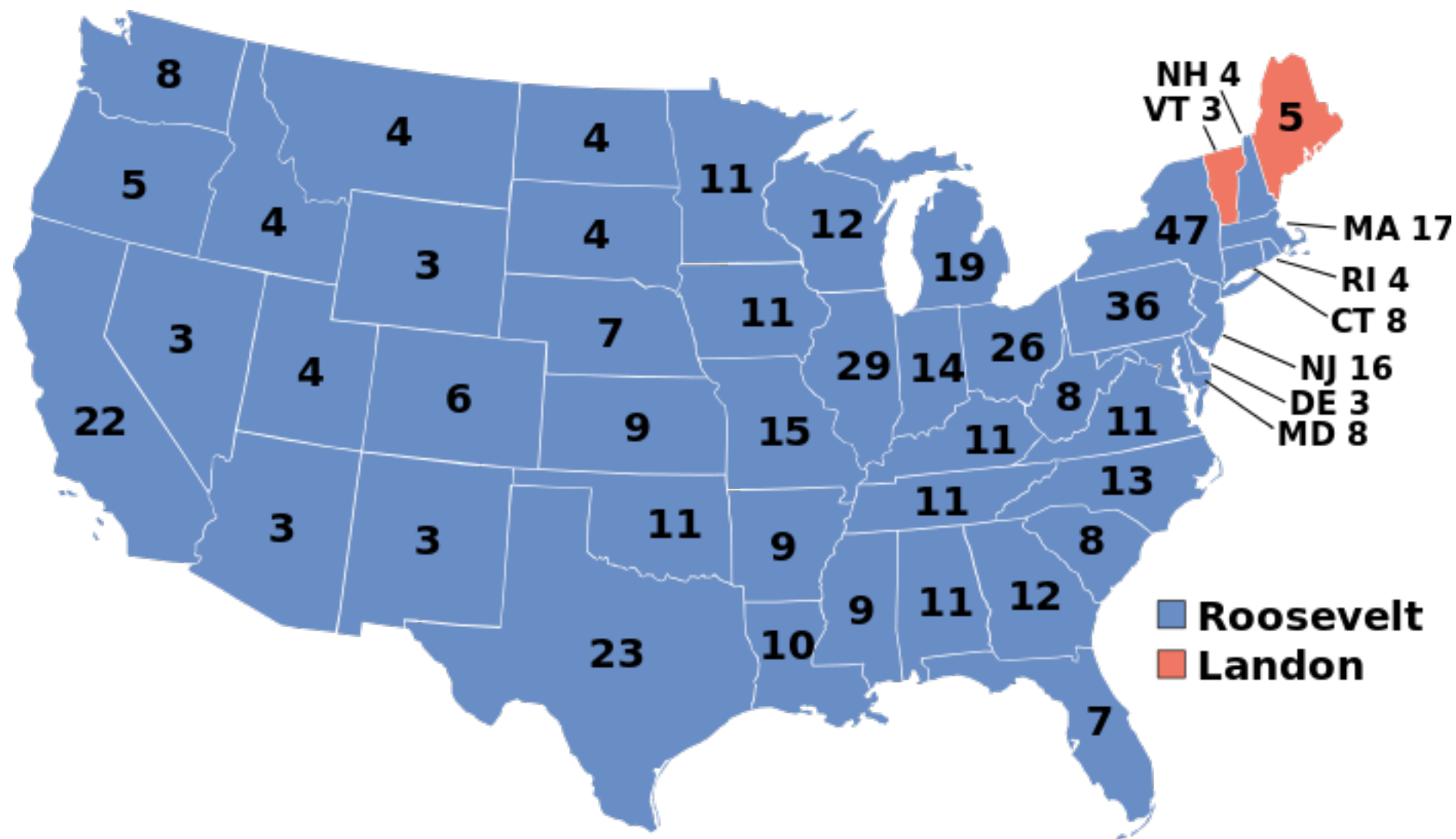
**Running mate**

**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](https://en.wikipedia.org/wiki/United_States_presidential_election,_1936)



# 1936 Presidential Election, Landon vs. FDR

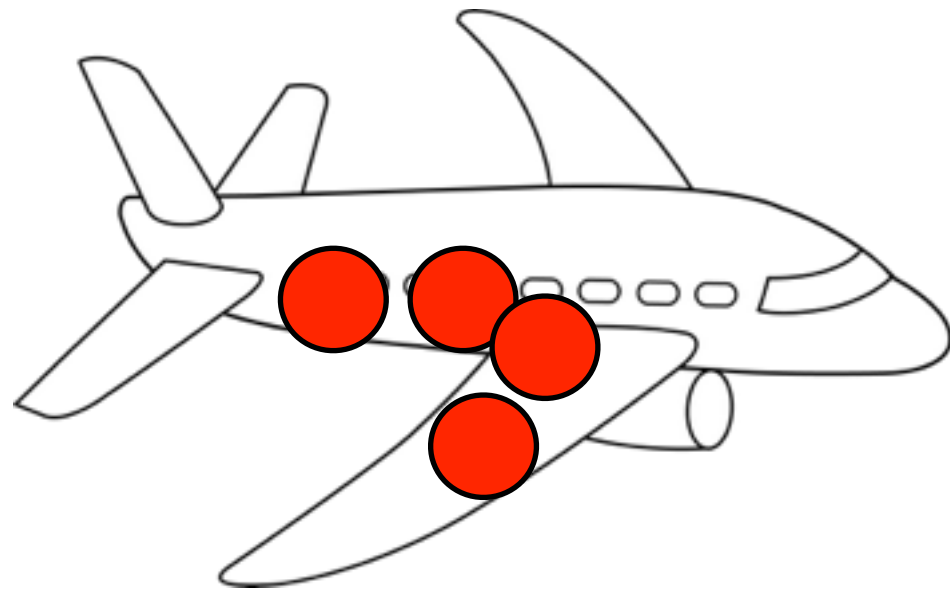
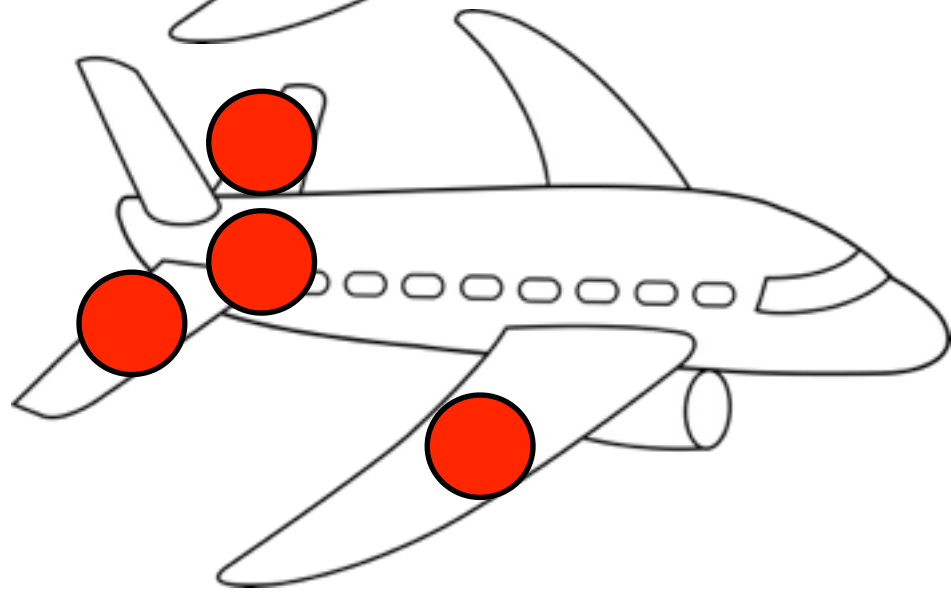
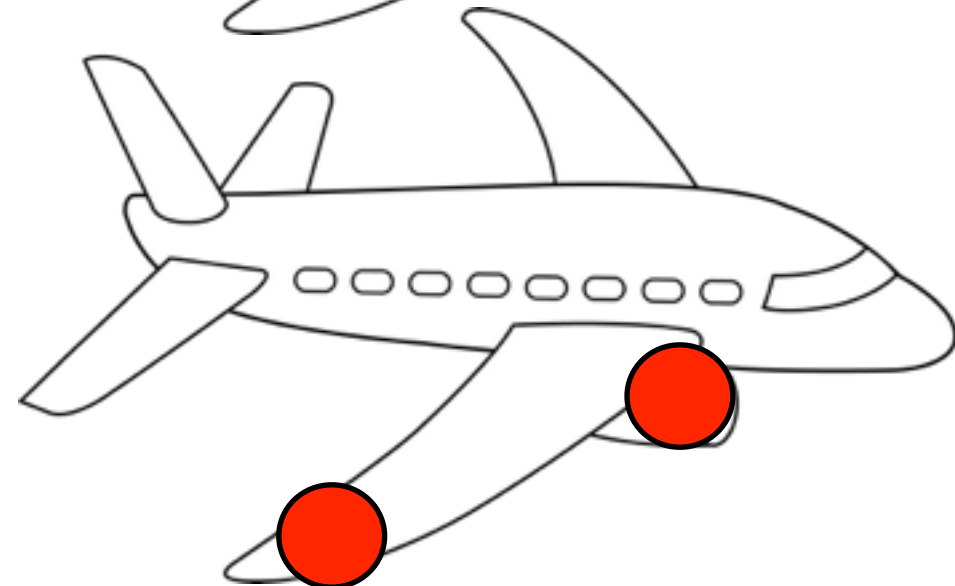
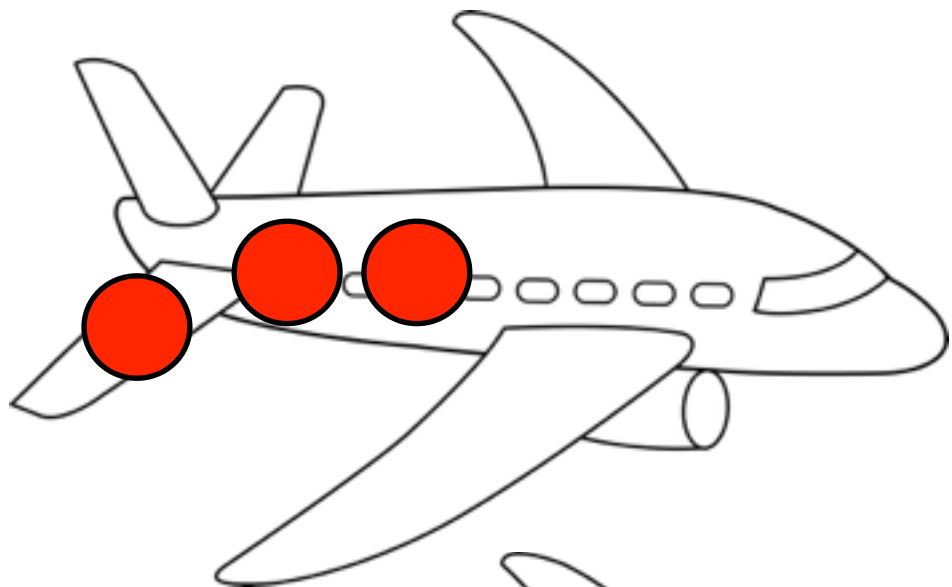
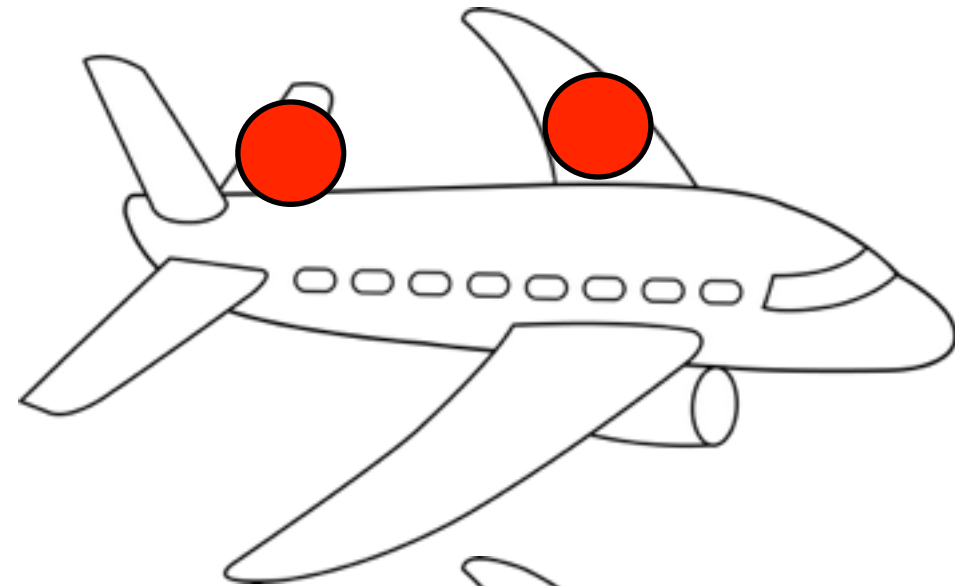
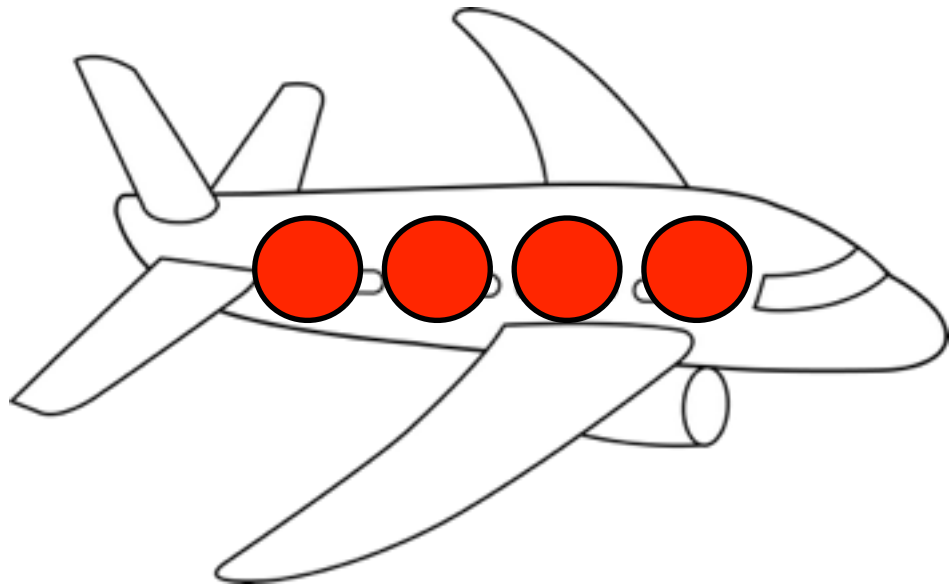
Literary Digest got responses from 2.3 million out of 10 million people surveyed.

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

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



# Wald and the Bullet Holes

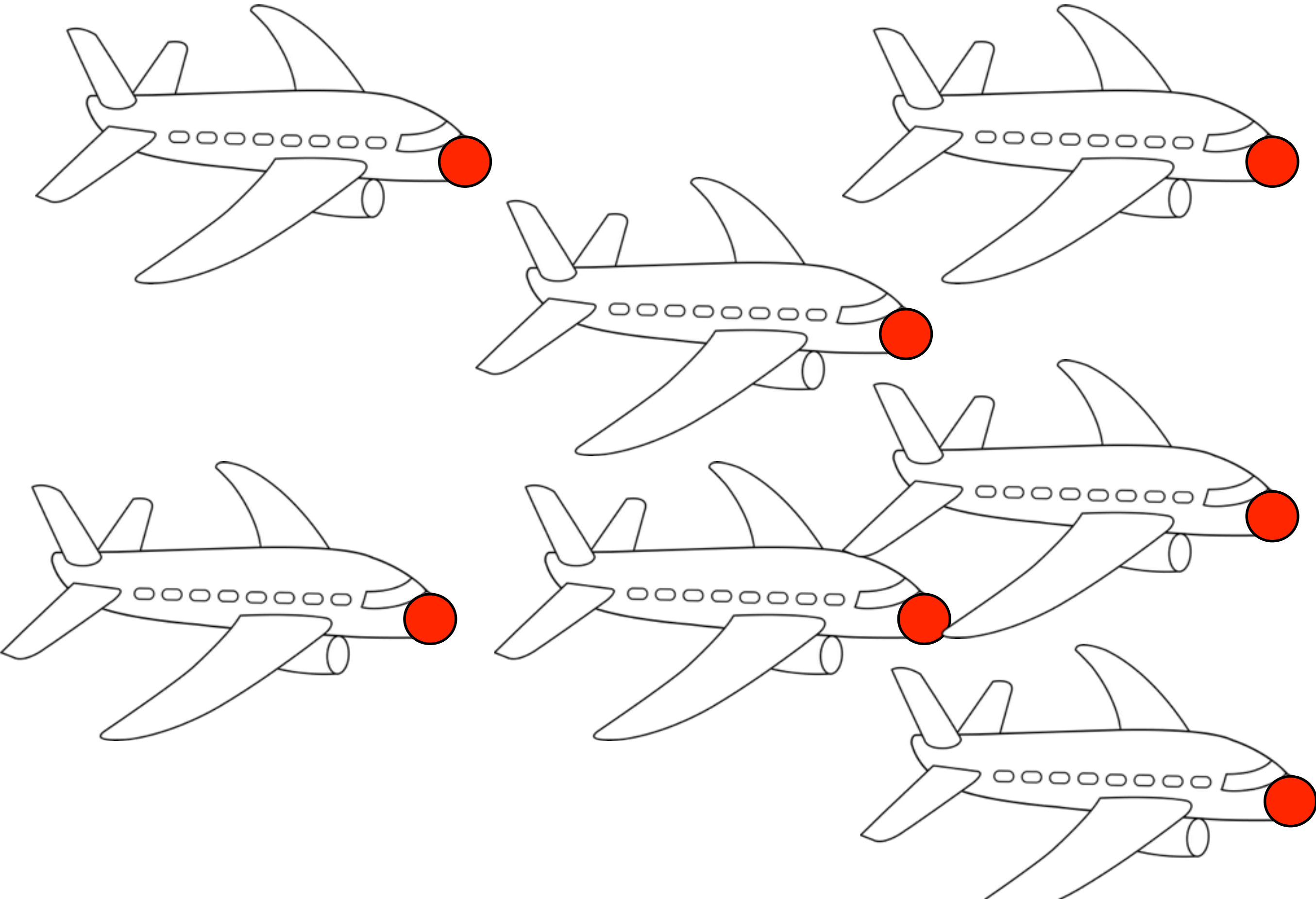


What about the *unobserved* planes? Missing data!





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

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:  $(100 + 50 \cdot 2) / 51 = 3.92$

Students calculate:  $(100 \cdot 100 + 100 \cdot 2) / 200 = 51$

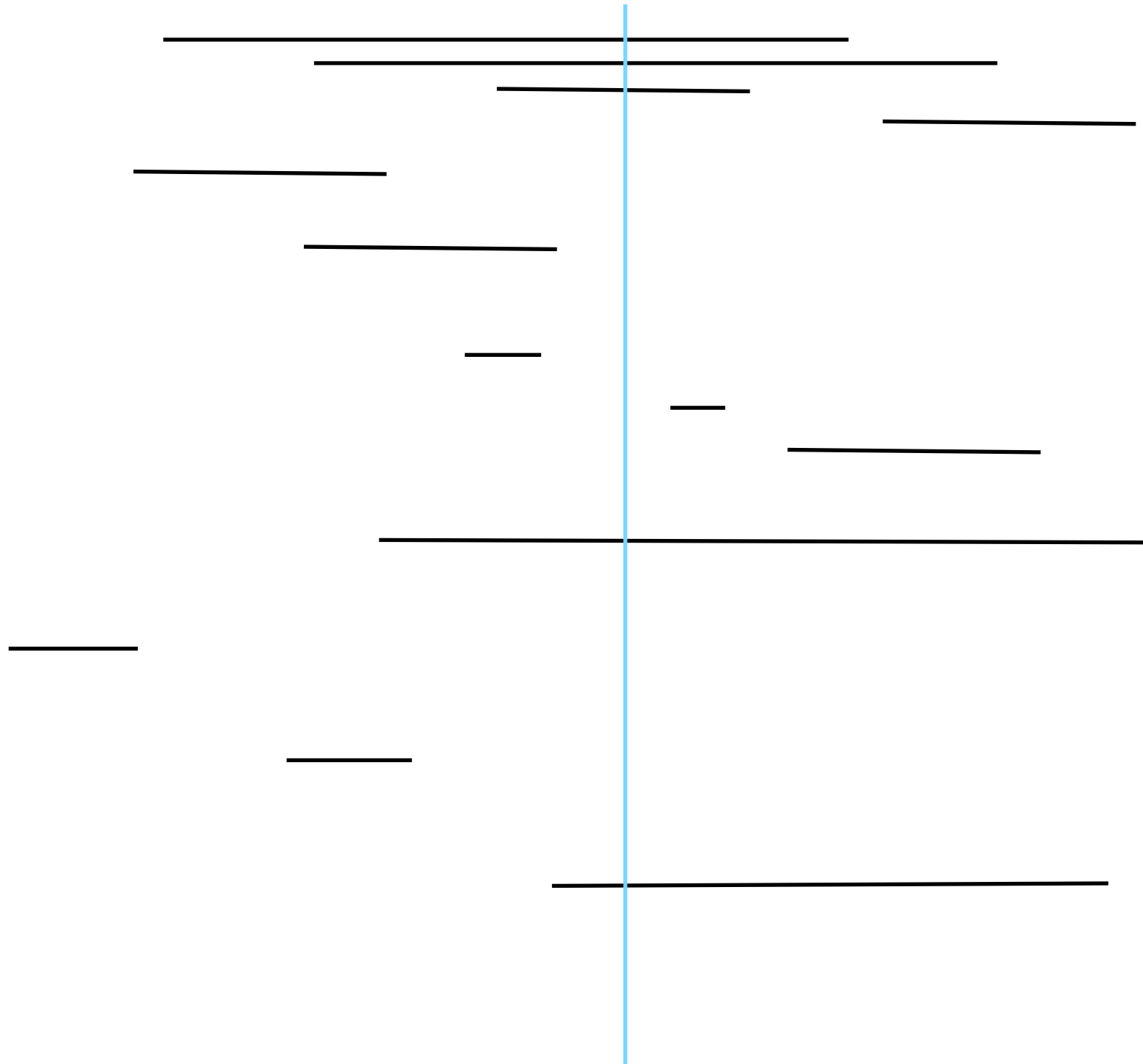


“About 10 percent of the 1.6 million inmates in America’s prisons are serving life sentences; another 11 percent are serving over 20 years.”

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?



# Bias of an Estimator

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

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

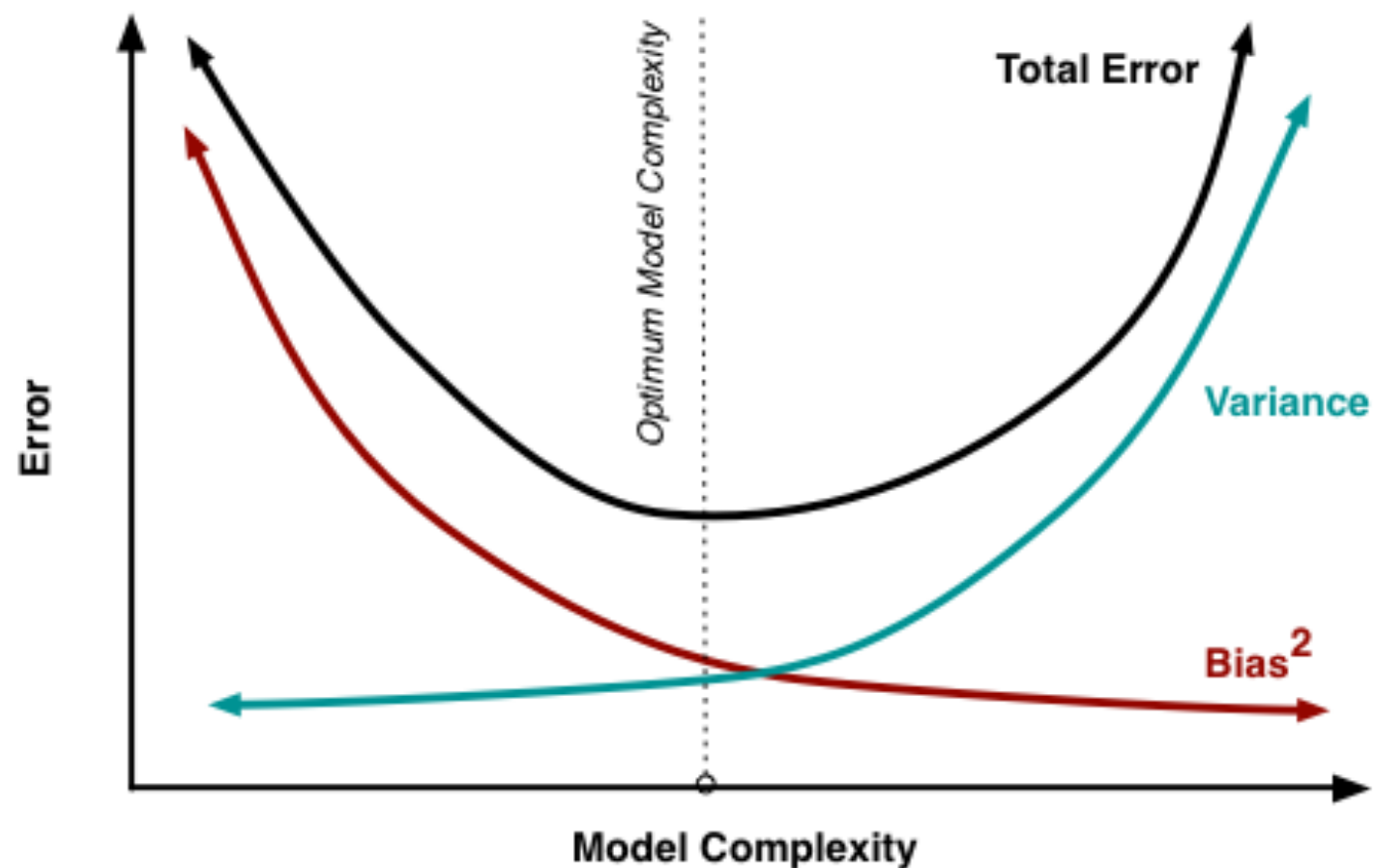
So why not just subtract off the bias?



# Bias-Variance Tradeoff

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

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



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

# Unbiased Estimation: Poisson Example

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

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

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

**sensible?**

# 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?

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

(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/>

# Multiple Testing, Bonferroni

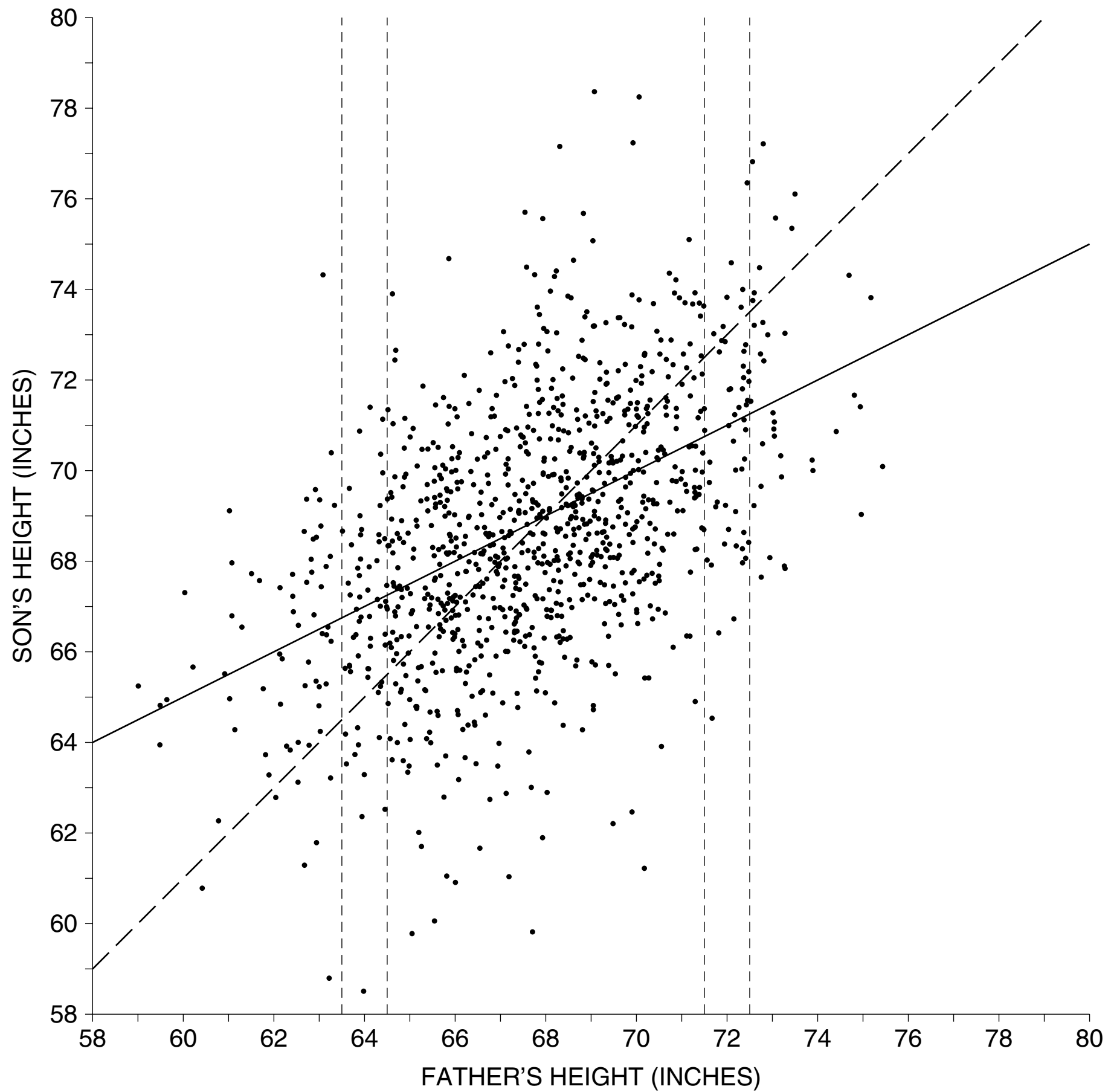
How should we handle p-values  
when testing multiple hypotheses?

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

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

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





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

# Daniel Kahneman Quote on RTTM

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

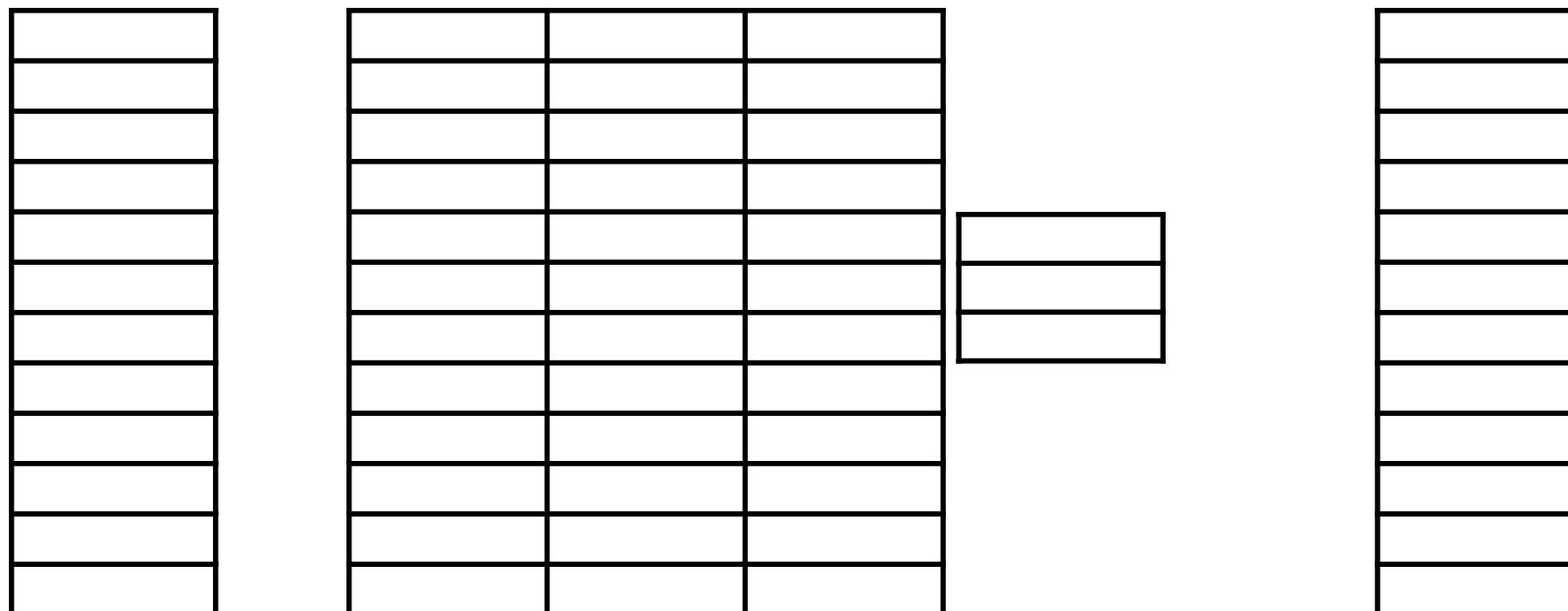
Now, what about predicting the parent's height from  
the child's height? Use  $x = y/r$ ?

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

# Linear Model

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

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





# 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.

# Sample Quantities vs. Population Quantities

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

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

sample version

(think of  $x$  and  $y$  as  
data vectors)

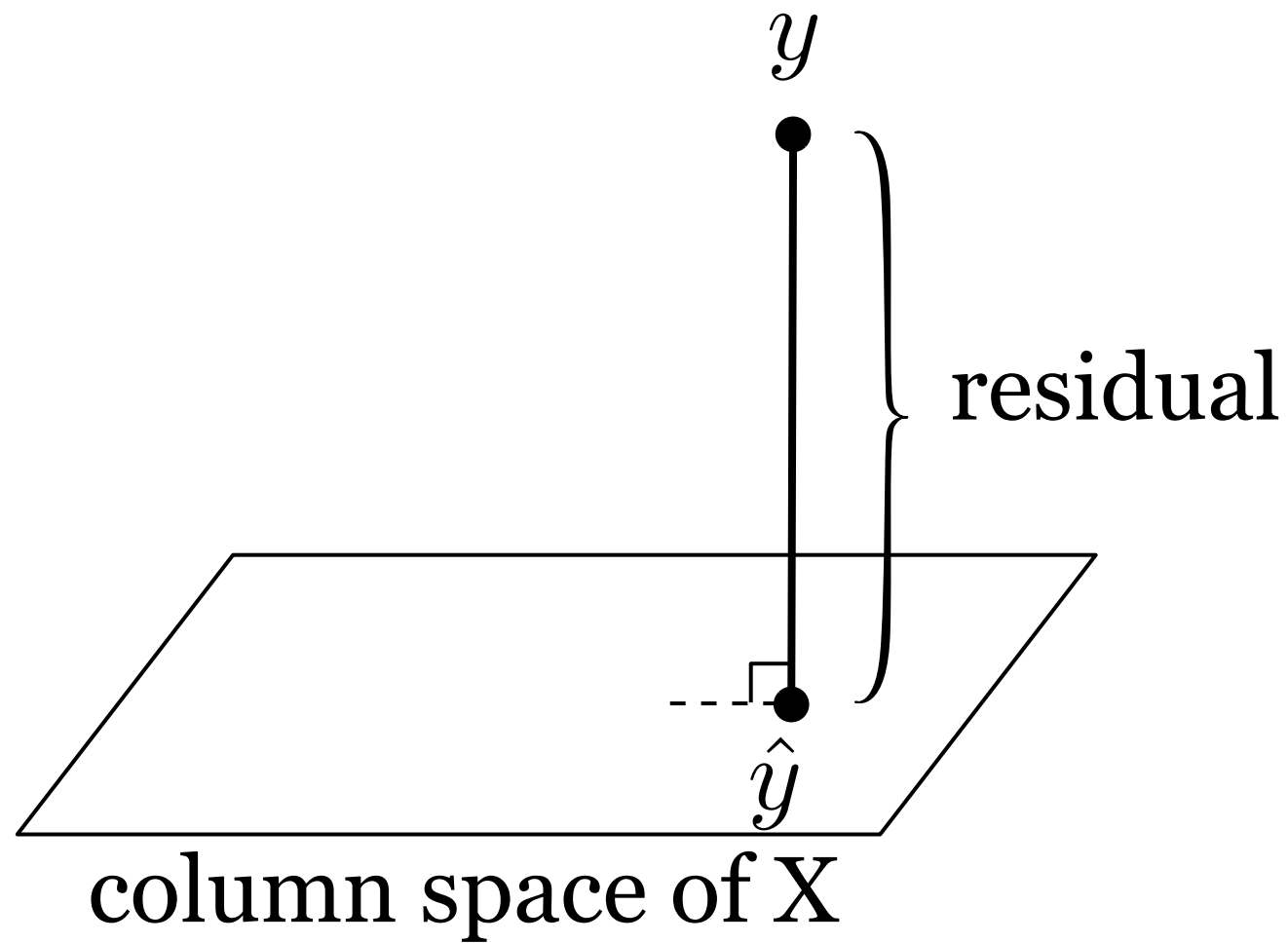
$$y = \beta_0 + \beta_1 x + \epsilon$$

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

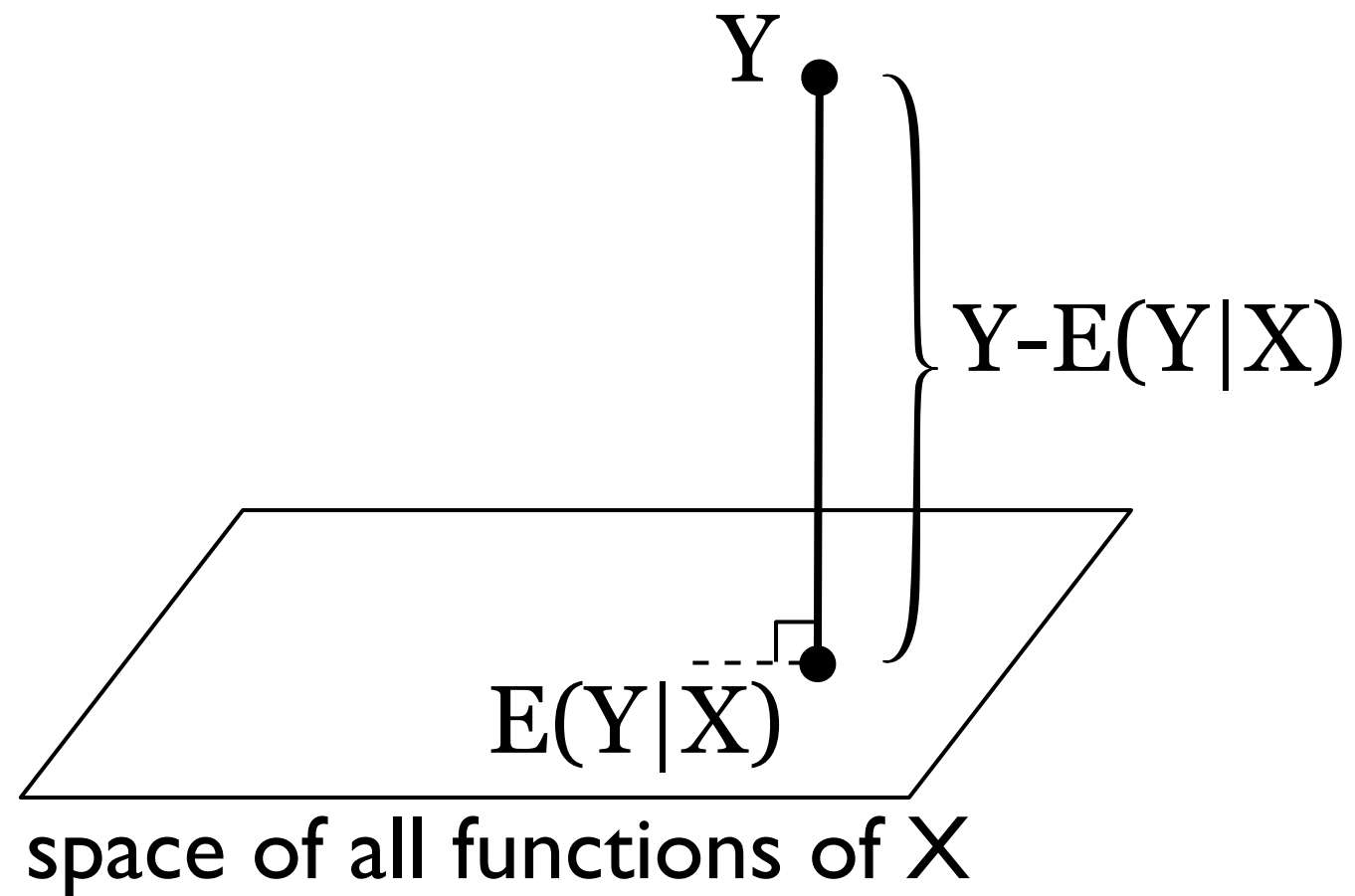
$$\text{cov}(y, x) = \beta_1 \text{cov}(x, x)$$

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

visualize regression as a *projection*



or as a *conditional expectation*



# 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,  $\beta$  is a  $k$  by 1 vector of parameters, and the errors  $\epsilon_j$  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

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

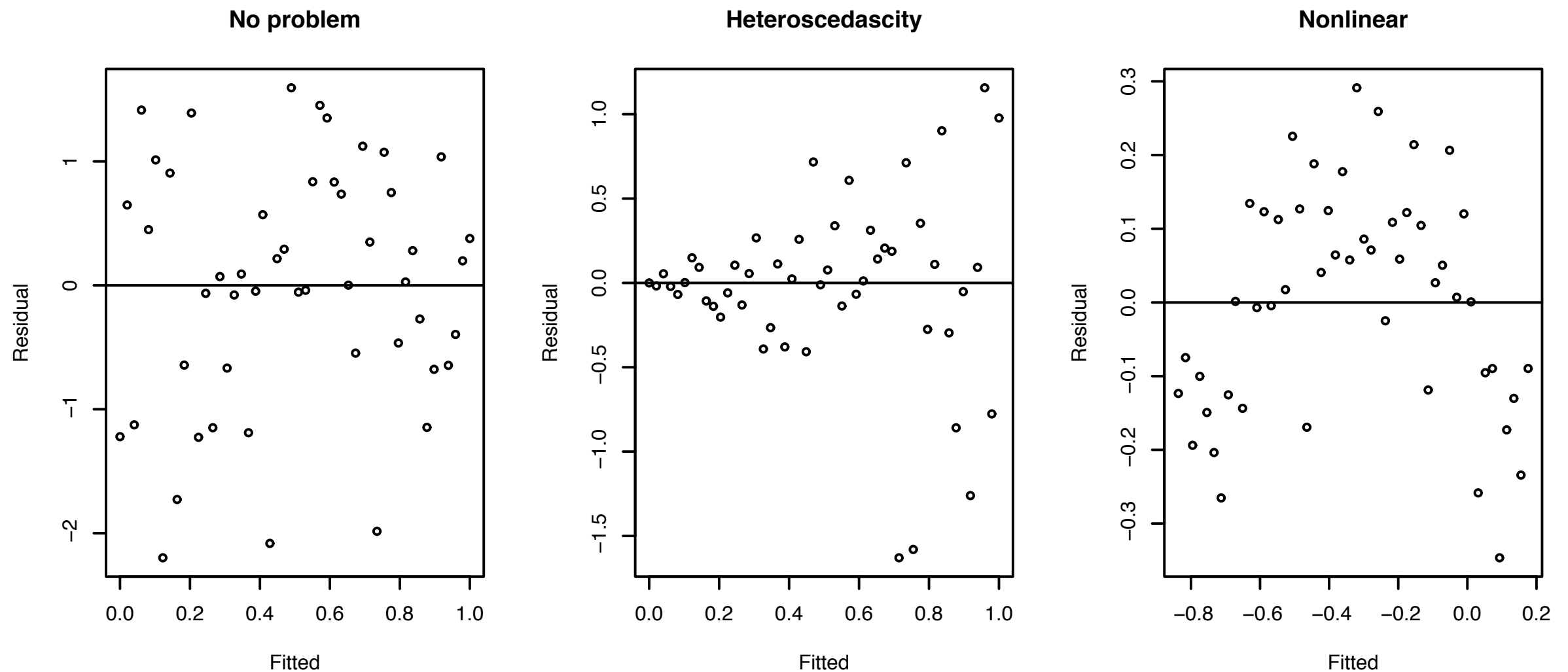
mirrors

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

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

# 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

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

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

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

Adding more predictors can only increase  $R^2$ .