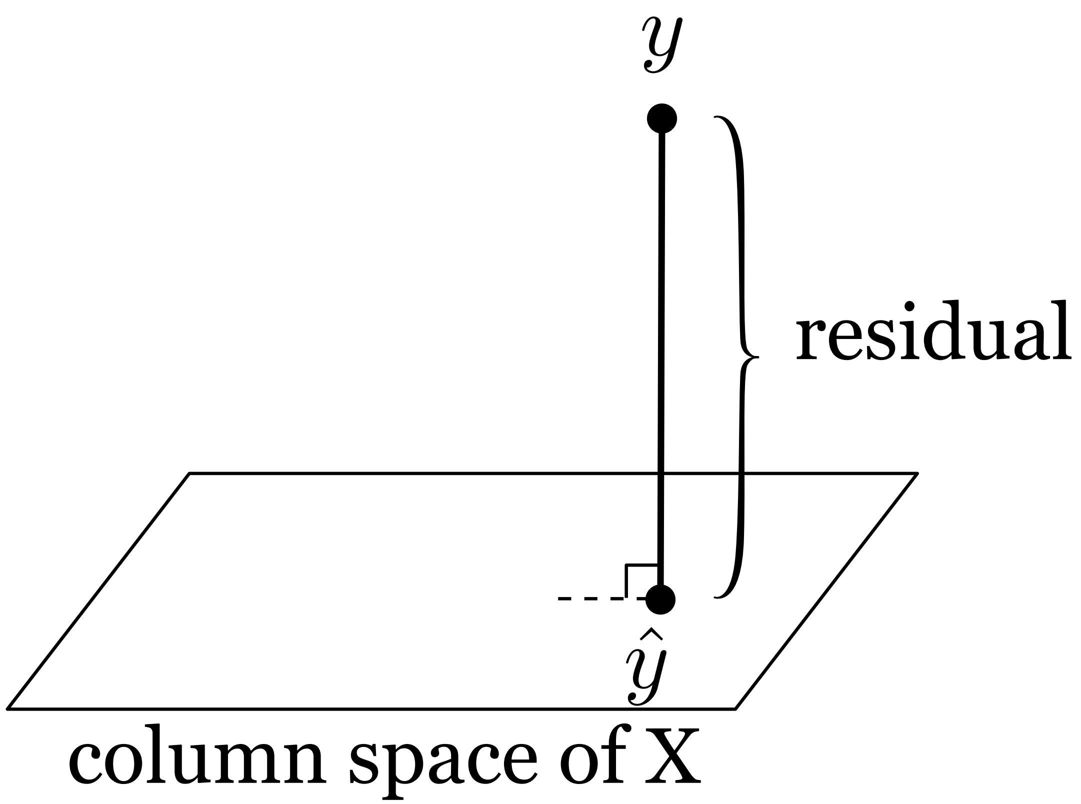
CS109/Stat121/AC209/E-109 Data Science

Regression

Hanspeter Pfister & Joe Blitzstein
pfister@seas.harvard.edu / blitzstein@stat.harvard.edu



This Week

- HW2 due 10/3 at 11:59 pm – start now!
- Friday lab **10-11:30 am** in MD G115

What’s the probability of a tie in an election? What’s the chance that your vote will be decisive?

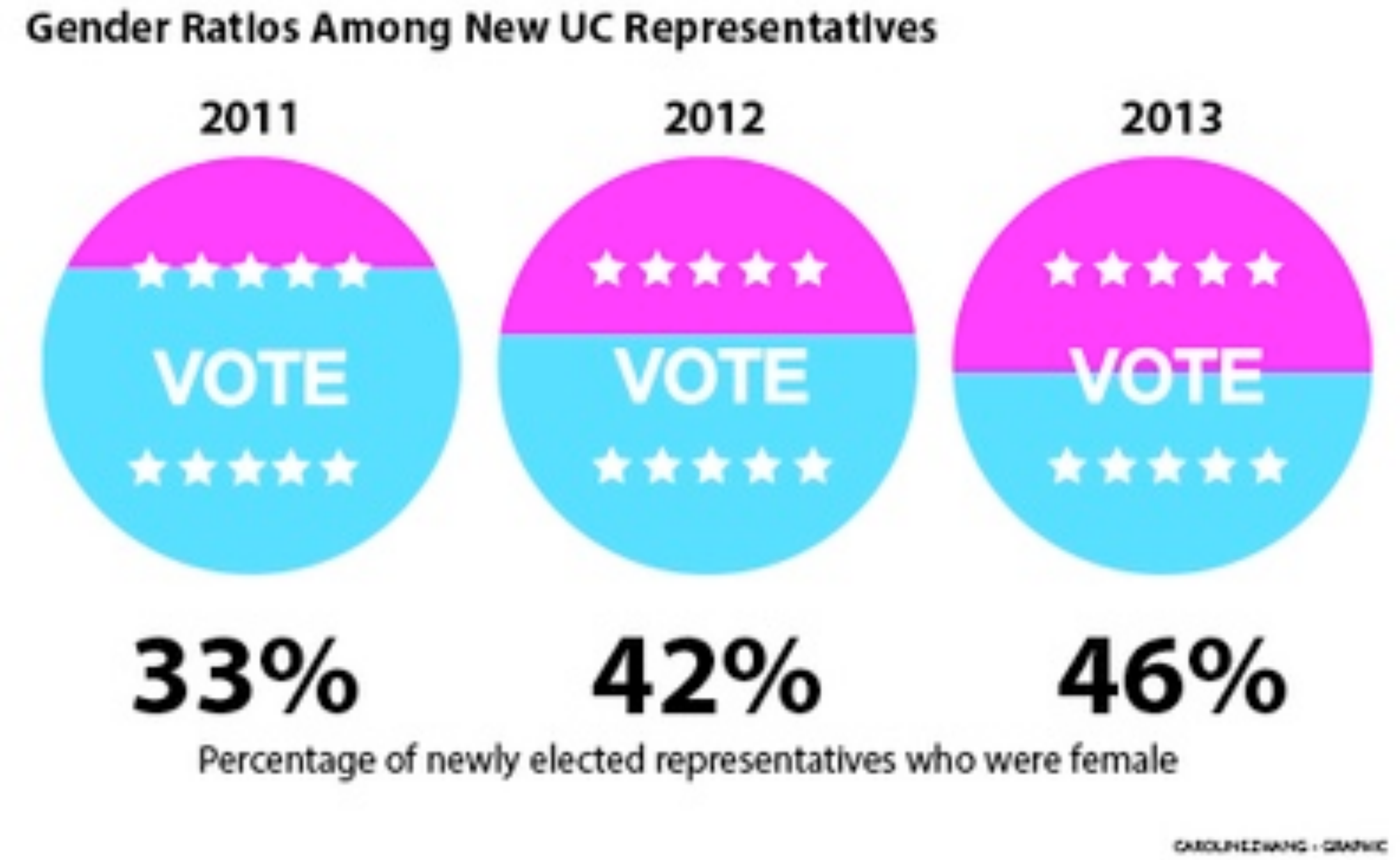
Simplest version: n voters, 2 candidates,
probability 1/2 for each candidate

n must be even...

Scales inversely to sqrt(n). Use Stirling or CLT to show.

$$P(\text{election is tied}) = P(X = n/2) = \binom{n}{n/2}\frac{1}{2^n} \approx \frac{\sqrt{2\pi n}(\frac{n}{e})^n}{2\pi(n/2)(\frac{n}{2e})^n} \cdot \frac{1}{2^n} = \frac{1}{\sqrt{\pi n/2}}.$$

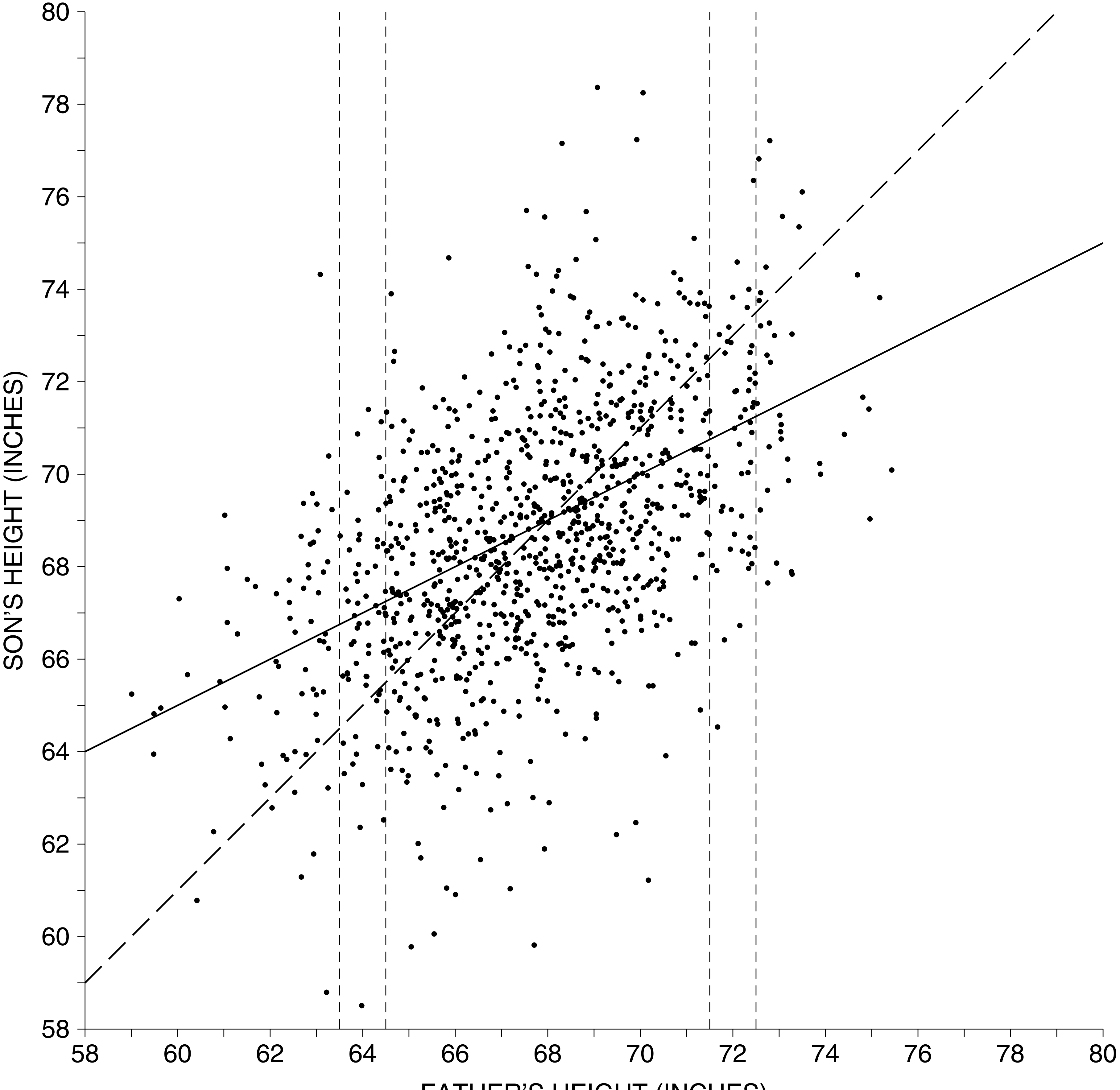
Crimson article on Undergraduate Council election



http://www.thecrimson.com/article/2013/9/23/uc-election-results-announced/

two tied elections; "unprecedented tie in
the UC's voting history"

"[A] and [B] of Eliot House tied with two votes each, while
[C] and [D] of Cabot House each received three votes."



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

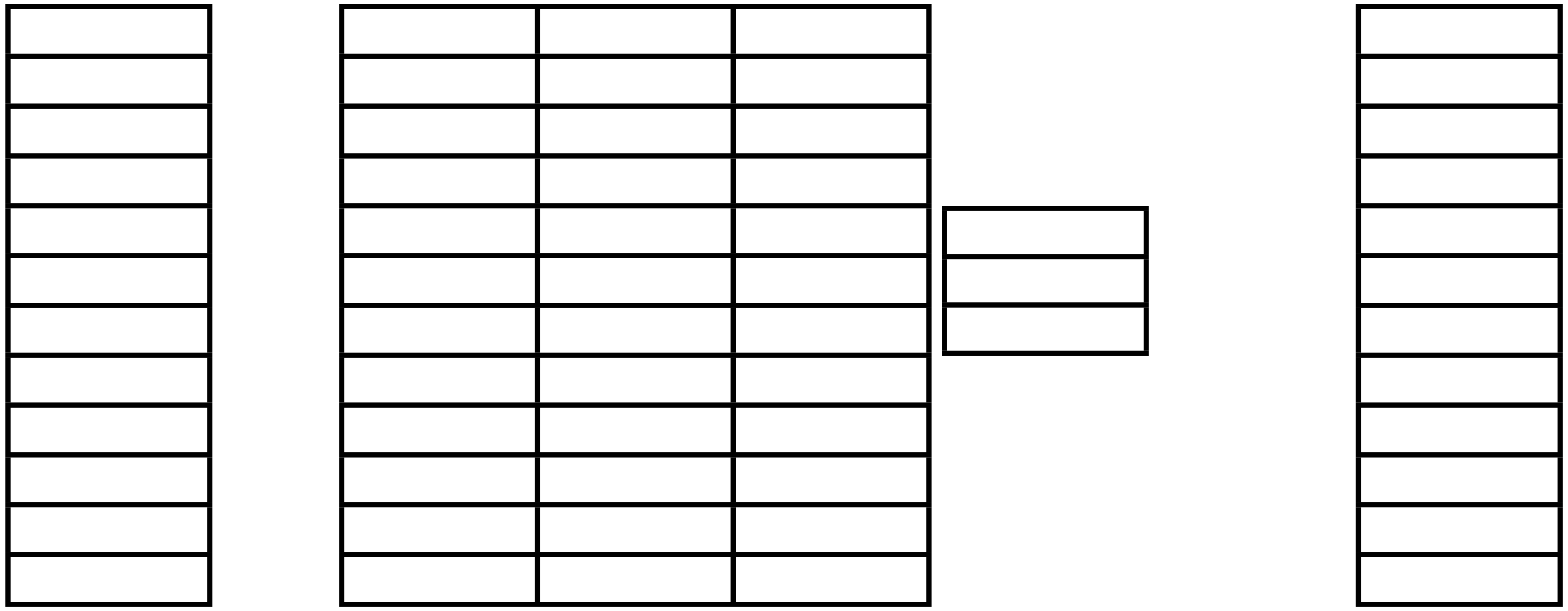
Now, what about predicting the parent's height from
the child's height? Use $x = y/r$?

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

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

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

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

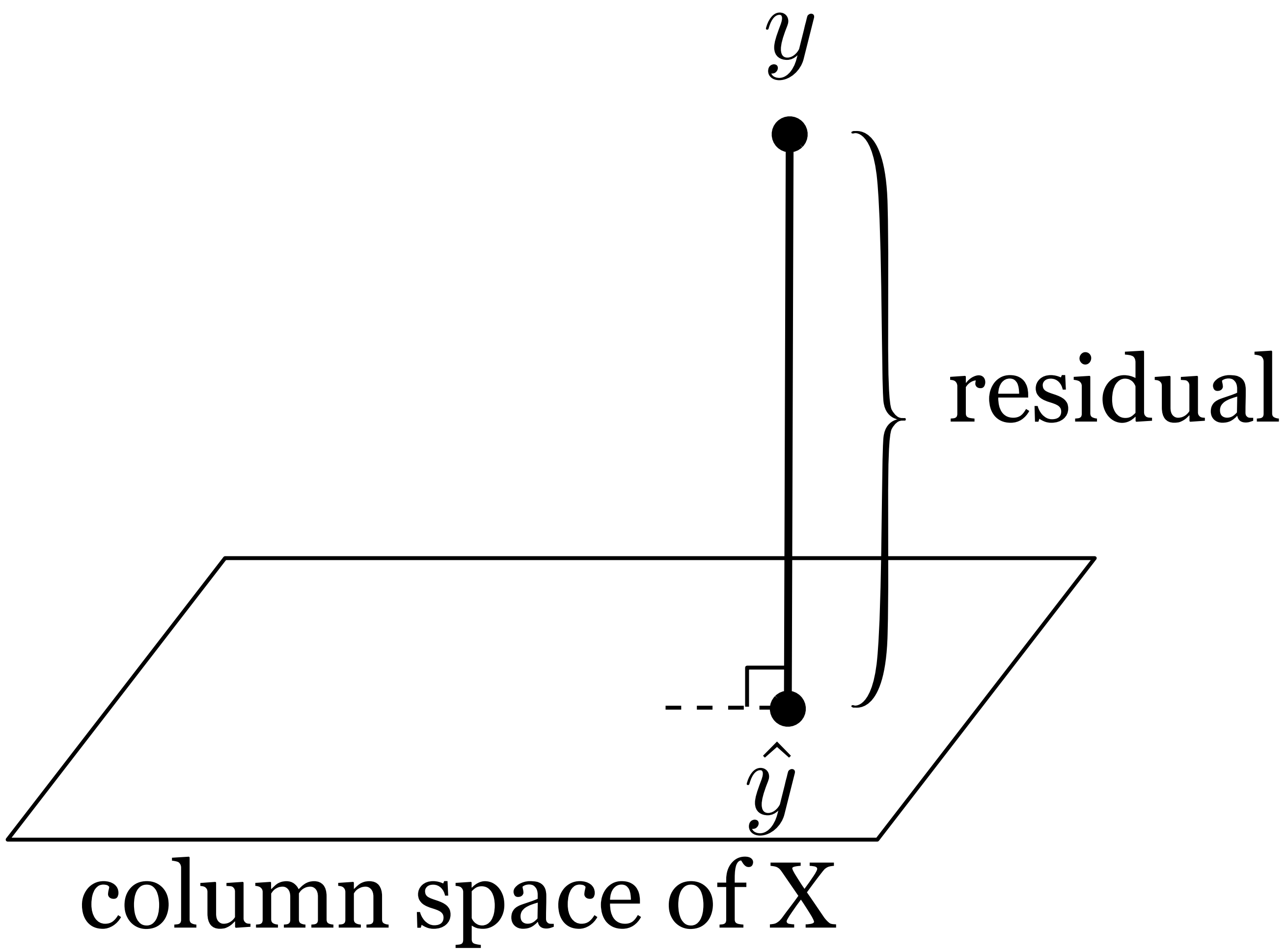
population version
(think of x and y
as r.v.s)

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

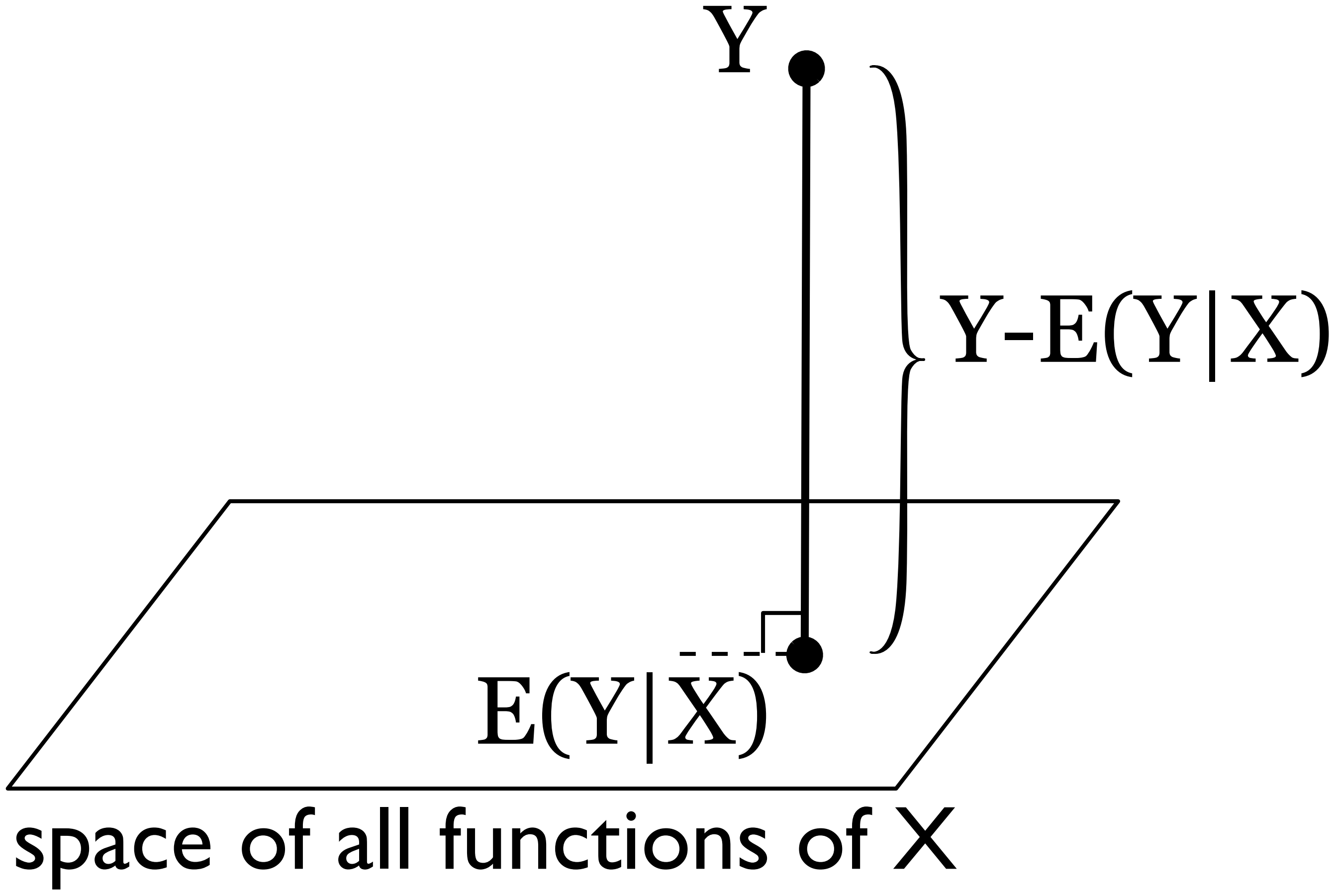
$$E(y) = \beta_0 + \beta_1 E(x)$$

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

visualize regression as a *projection*



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, 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.

Hat Matrix

$$H = X(X'X)^{-1}X'$$

this matrix represents *projection* into the column space of X

$$Hy = X\hat{\beta} = \hat{y}$$

H puts the hat on y....

Residuals

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

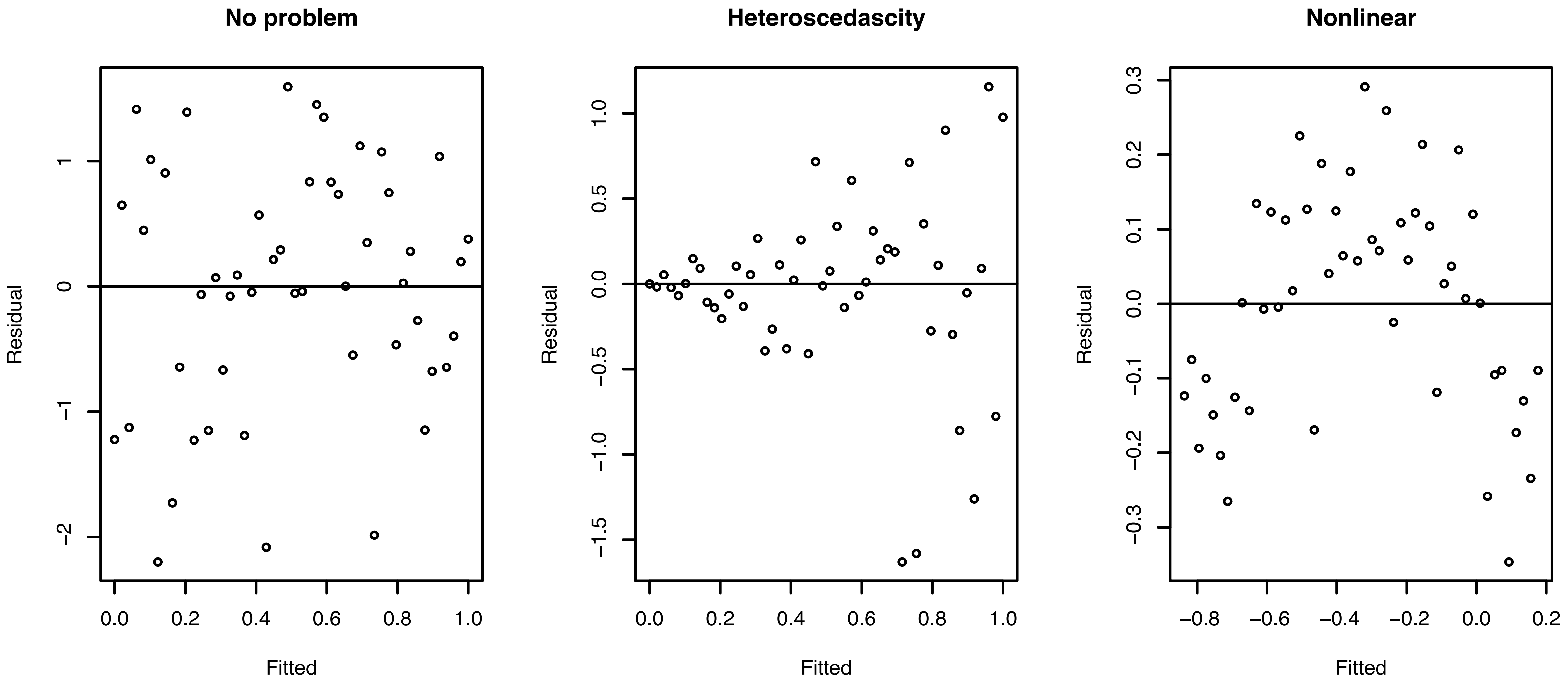
mirrors

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

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

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

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

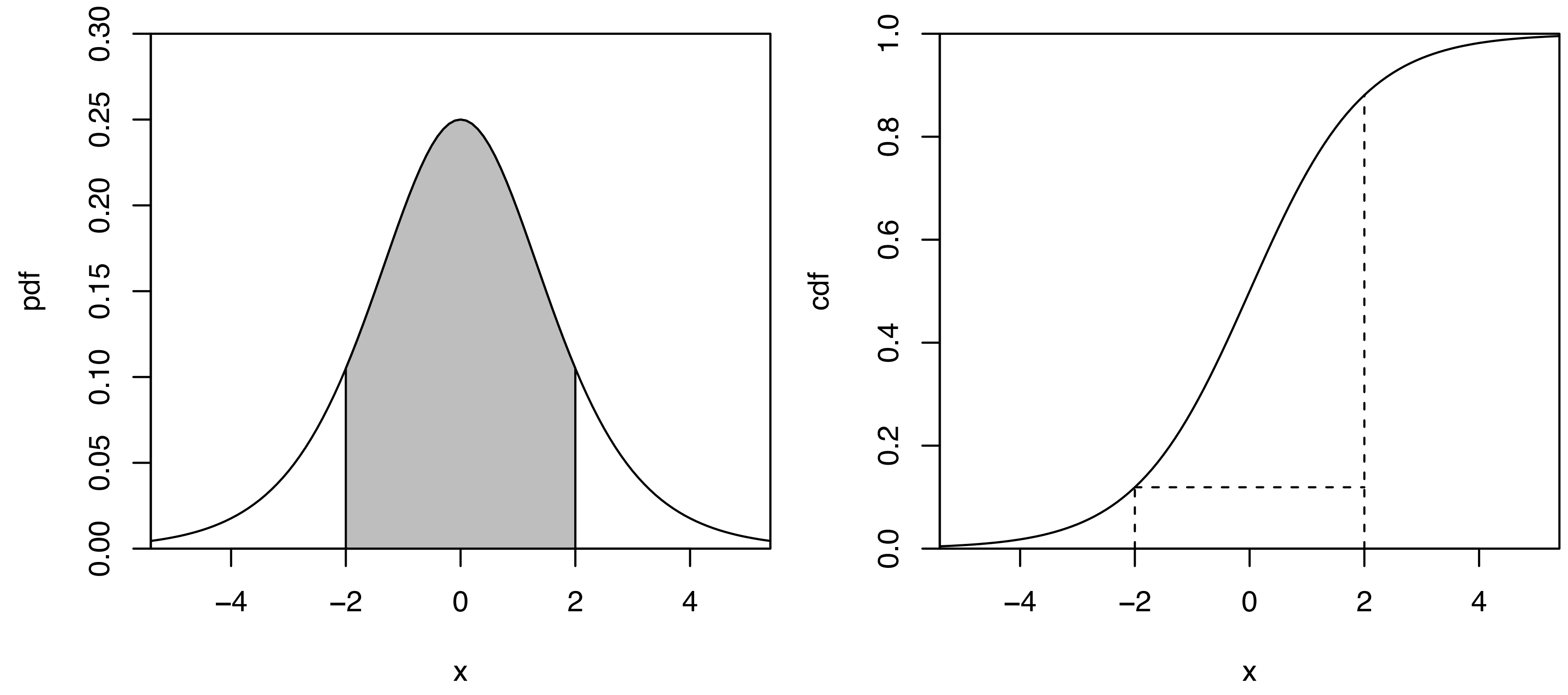
R^2 measures goodness of fit, but
it does *not* validate the model.
Adding more predictors can only increase R^2.

Logistic Regression

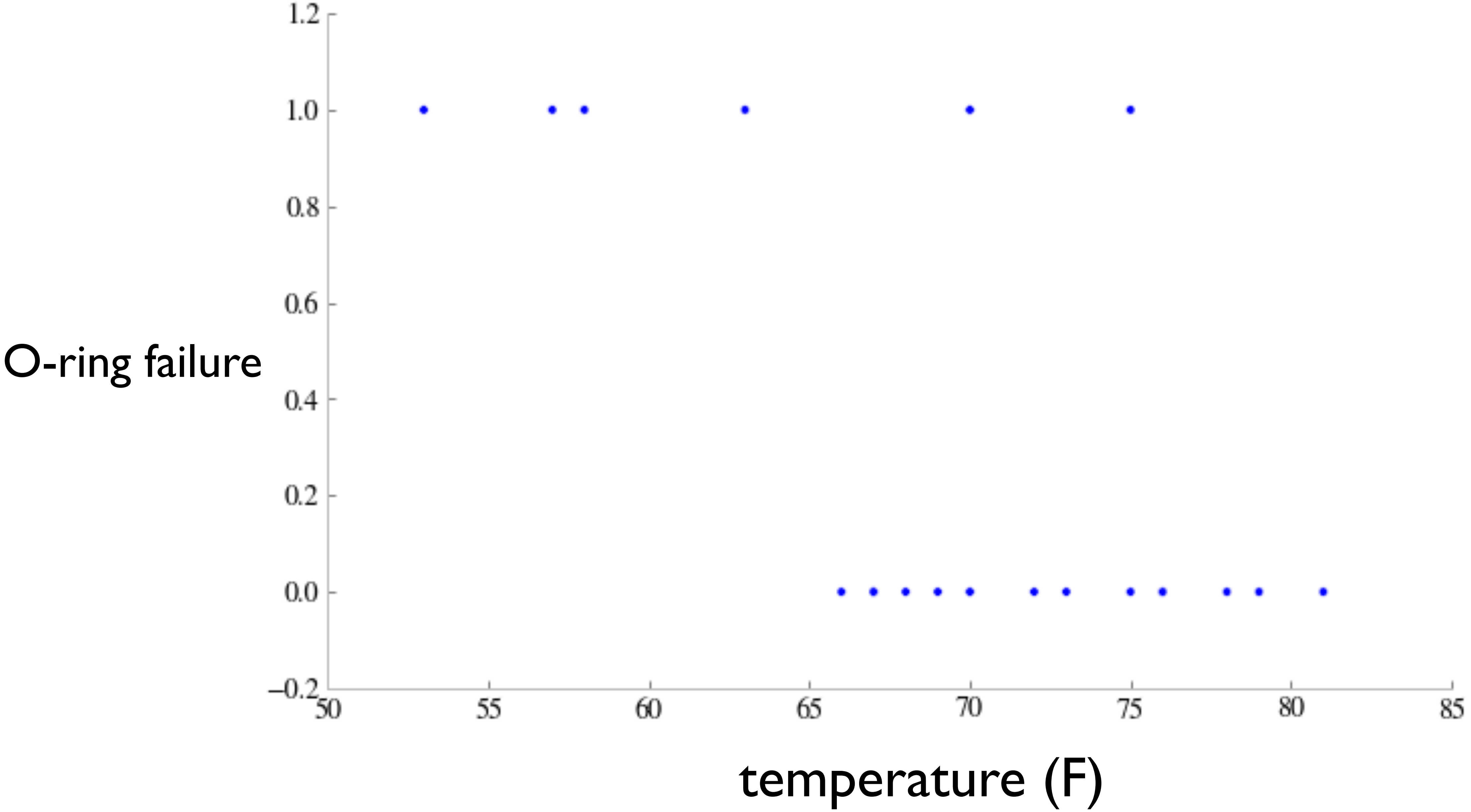
log odds:

$$\operatorname{logit}(p) \equiv \log\left(\frac{p}{1-p}\right)$$

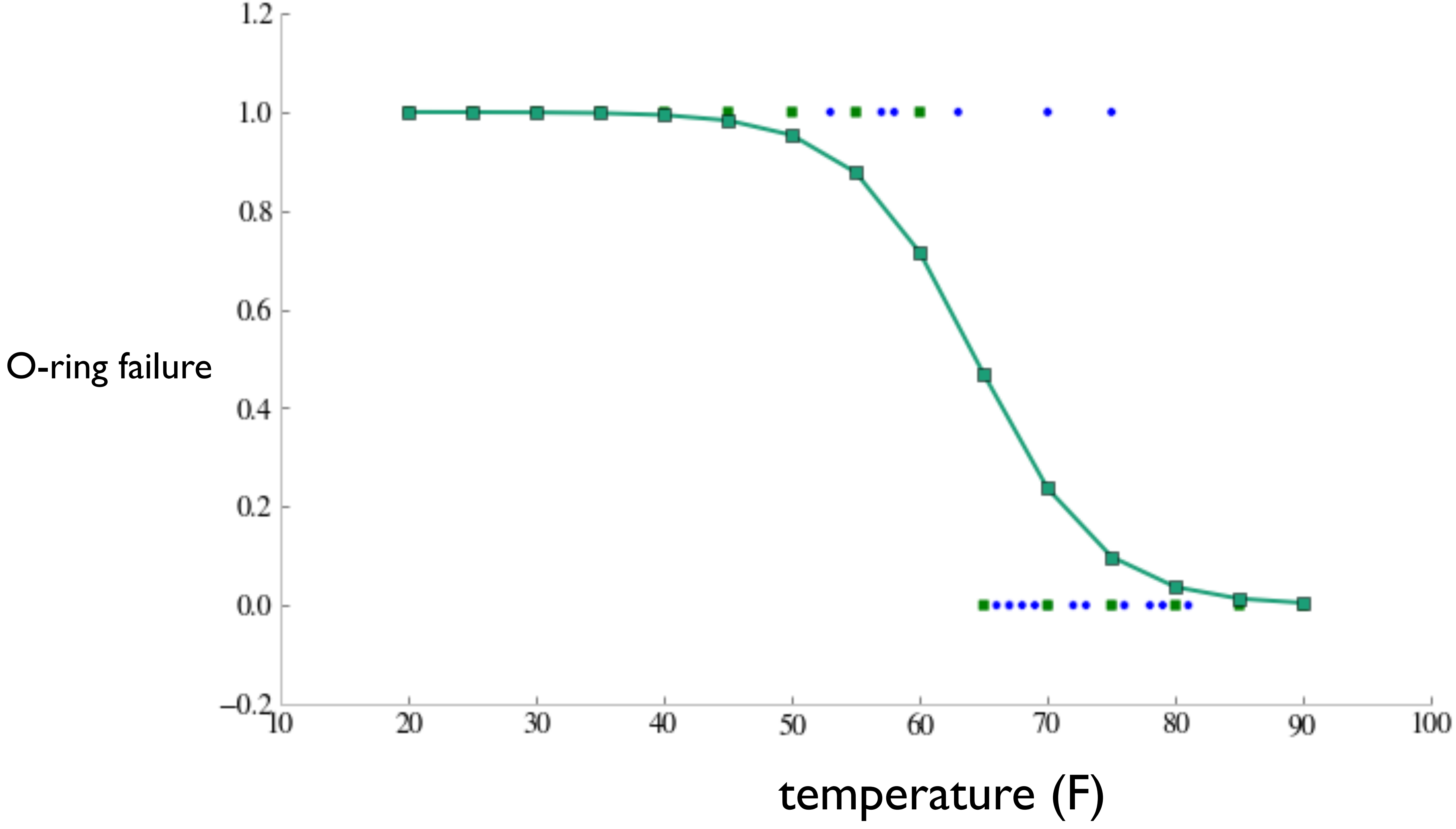
$$\operatorname{logit}^{-1}(x) = \frac{e^x}{1+e^x}.$$



Preview: Logistic Regression for Challenger Disaster



Preview: Logistic Regression for Challenger Disaster



Latent Variable Interpretation

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1}$$

Think of binary data y as having come from thresholding a continuous, unobserved variable following a linear model (1 if positive, 0 if negative).

Similarly, can think of *probit regression*

$$p = \Phi(\beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1})$$

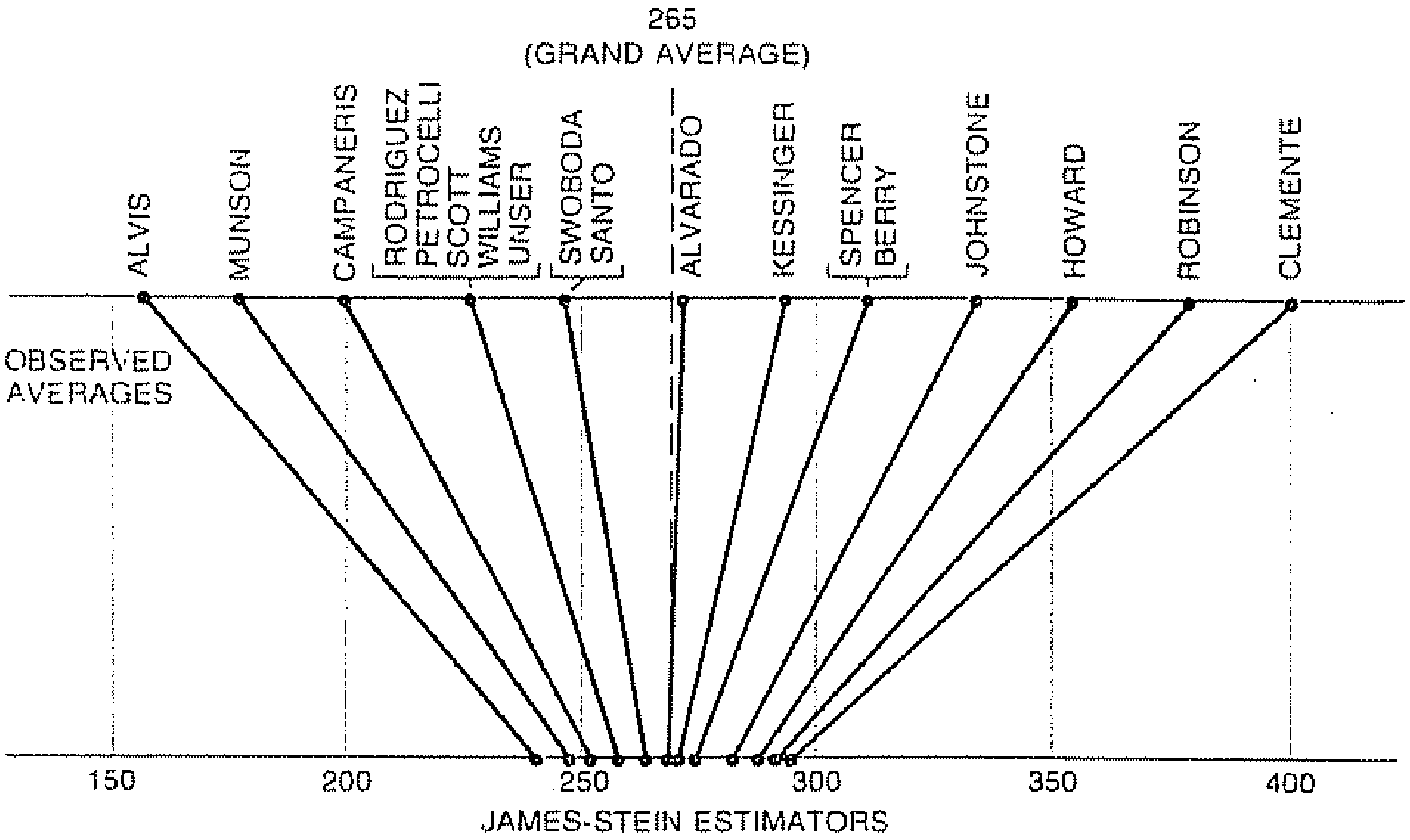
in terms of a latent variable with Normal errors

Stein's Paradox and Shrinkage Estimation

Let $y_1 \sim \mathcal{N}(\theta_1, 1), y_2 \sim \mathcal{N}(\theta_2, 1), \ldots, y_k \sim \mathcal{N}(\theta_k, 1)$ with $k \geq 3$. How should we estimate the vector θ, under sum of squared error loss?

Stein: the vector y is *inadmissible*; uniformly beaten by the James-Stein estimator

$$\hat{\theta}_j = \left(1 - \frac{k-2}{\sum_i y_i^2}\right) y_j.$$



JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value. Thus the theorem on which Stein's method is based asserts that the true batting abilities are more tightly clustered than the preliminary batting averages would seem to suggest they are.

Source: Efron-Morris, Scientific American 1977