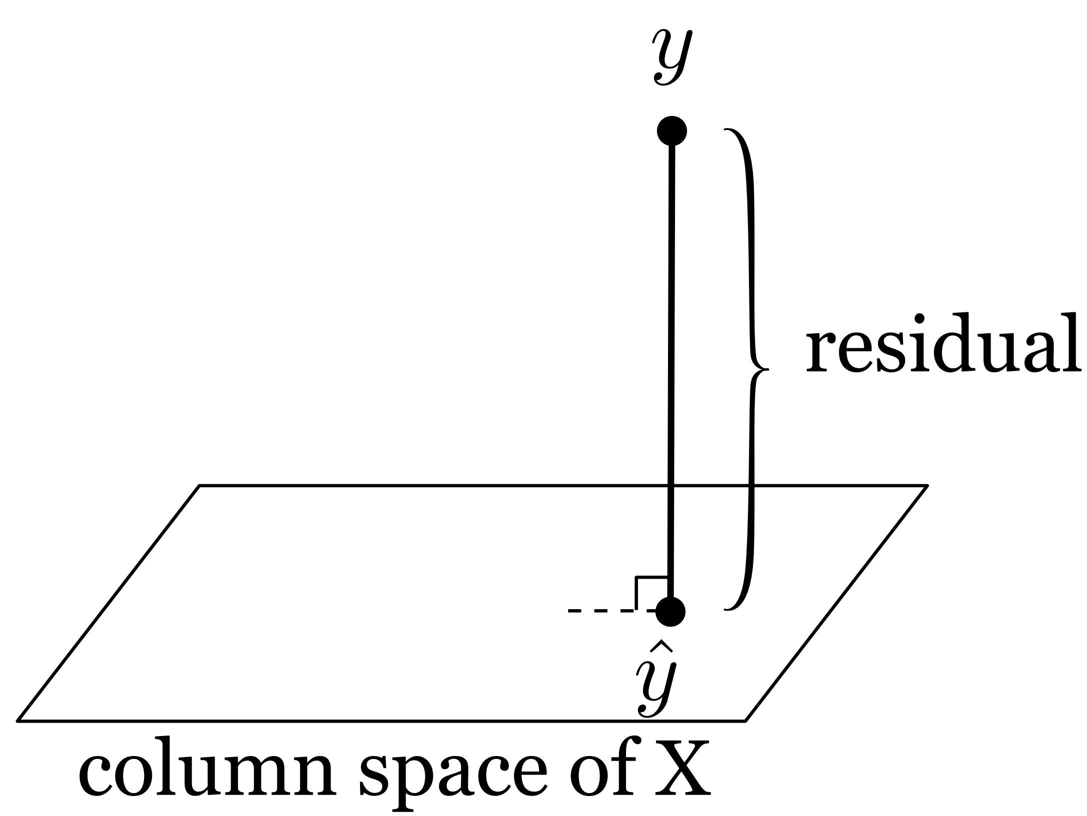
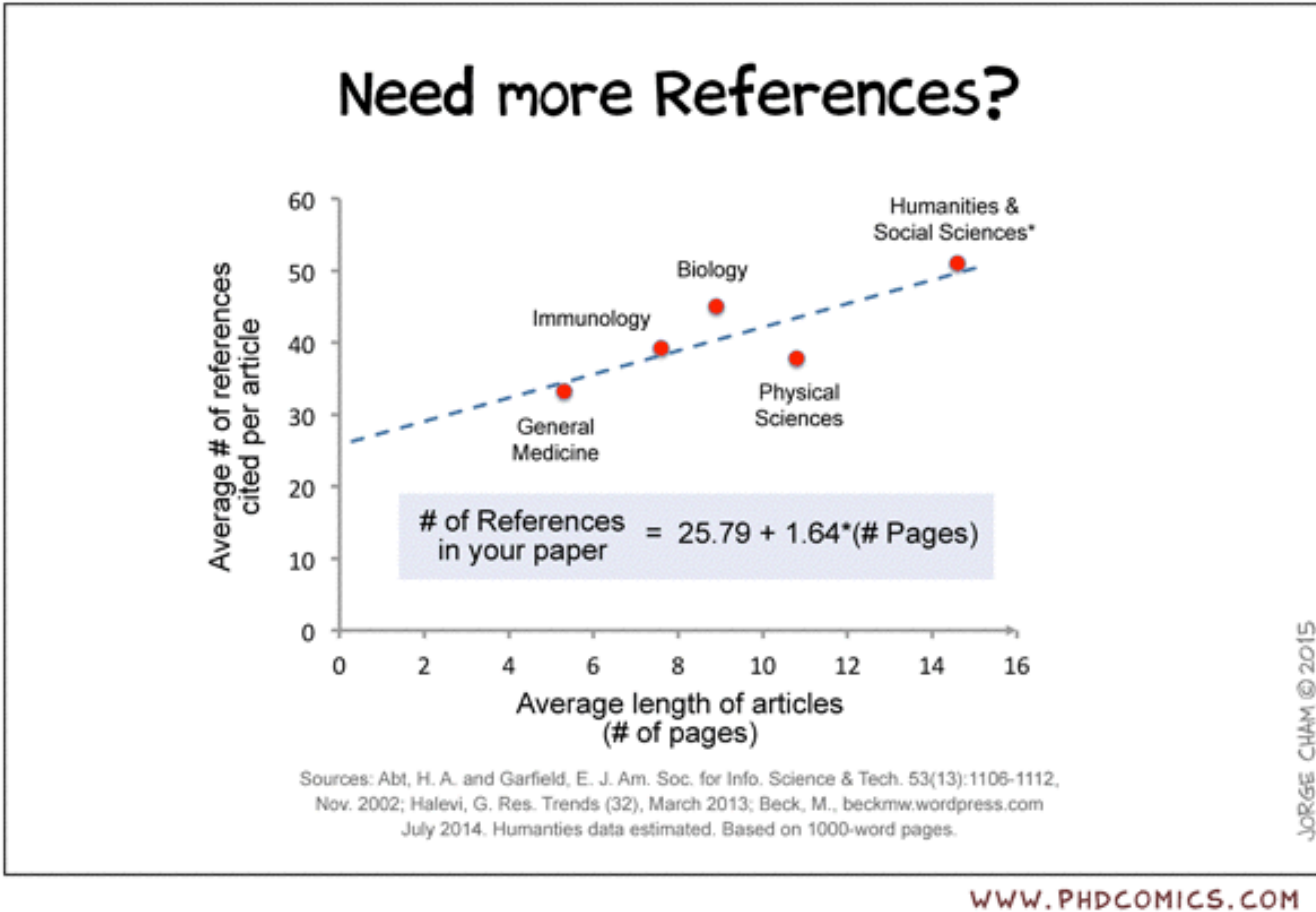
CS109/Stat121/AC209/E-109
Data Science
Regression Continued

Hanspeter Pfister, Joe Blitzstein, and Verena Kaynig



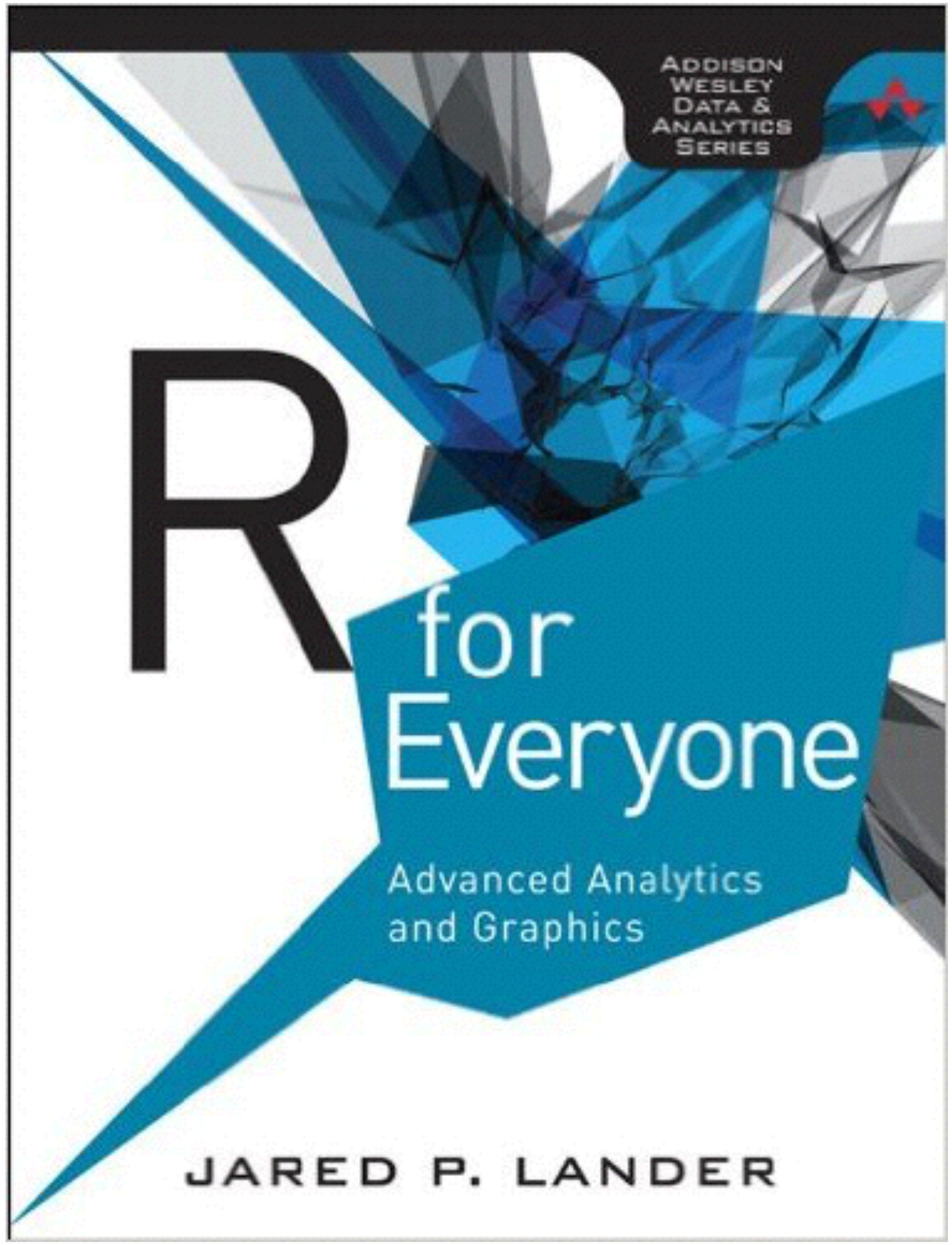
This Week

- HW2 due next Thursday (Oct 8) at 11:59 pm (Eastern Time)
- See updated Piazza posting guidelines (pinned note) and follow the format described there

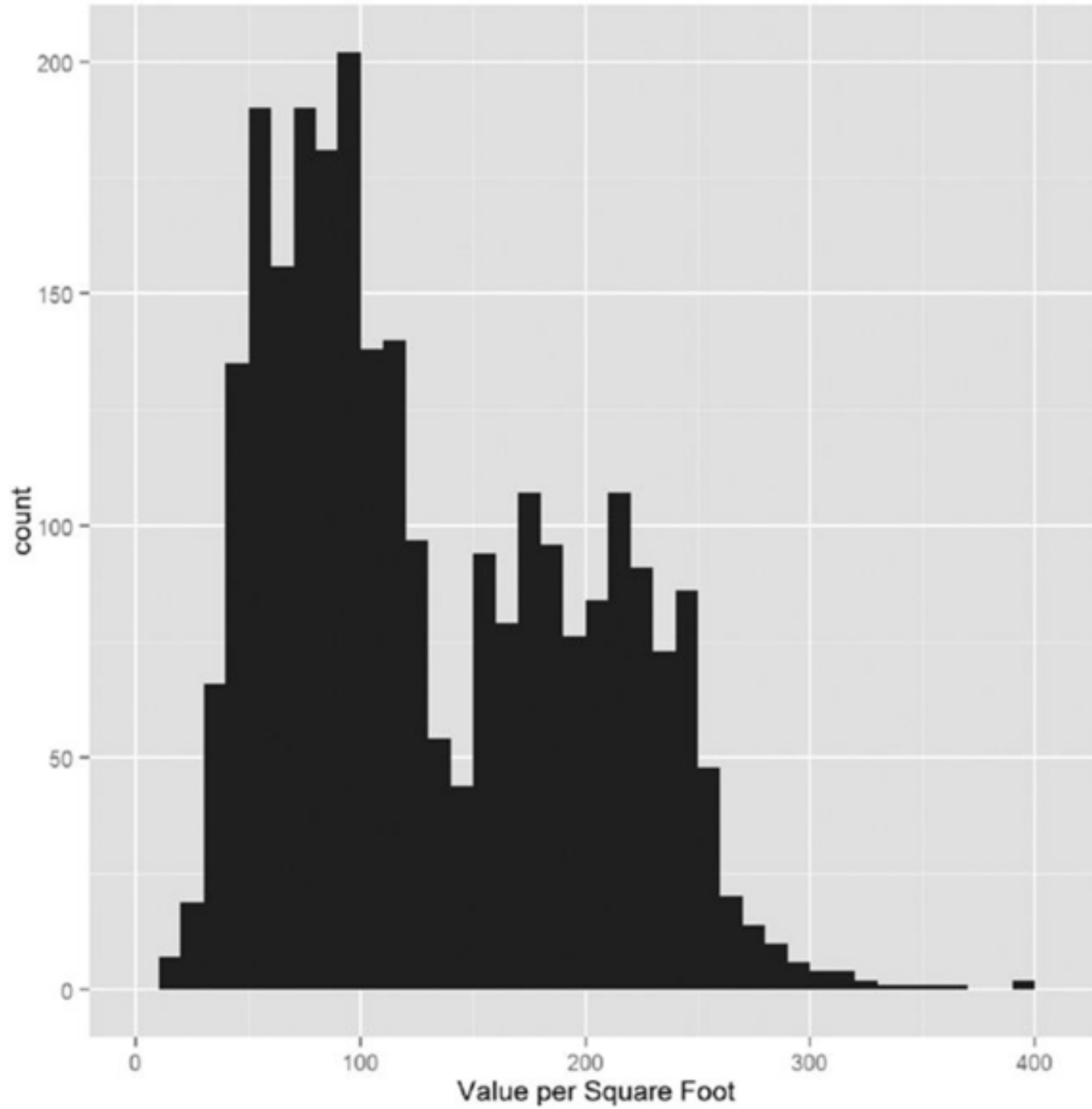


http://www.phdcomics.com/comics.php?f=1823

NYC Housing Example

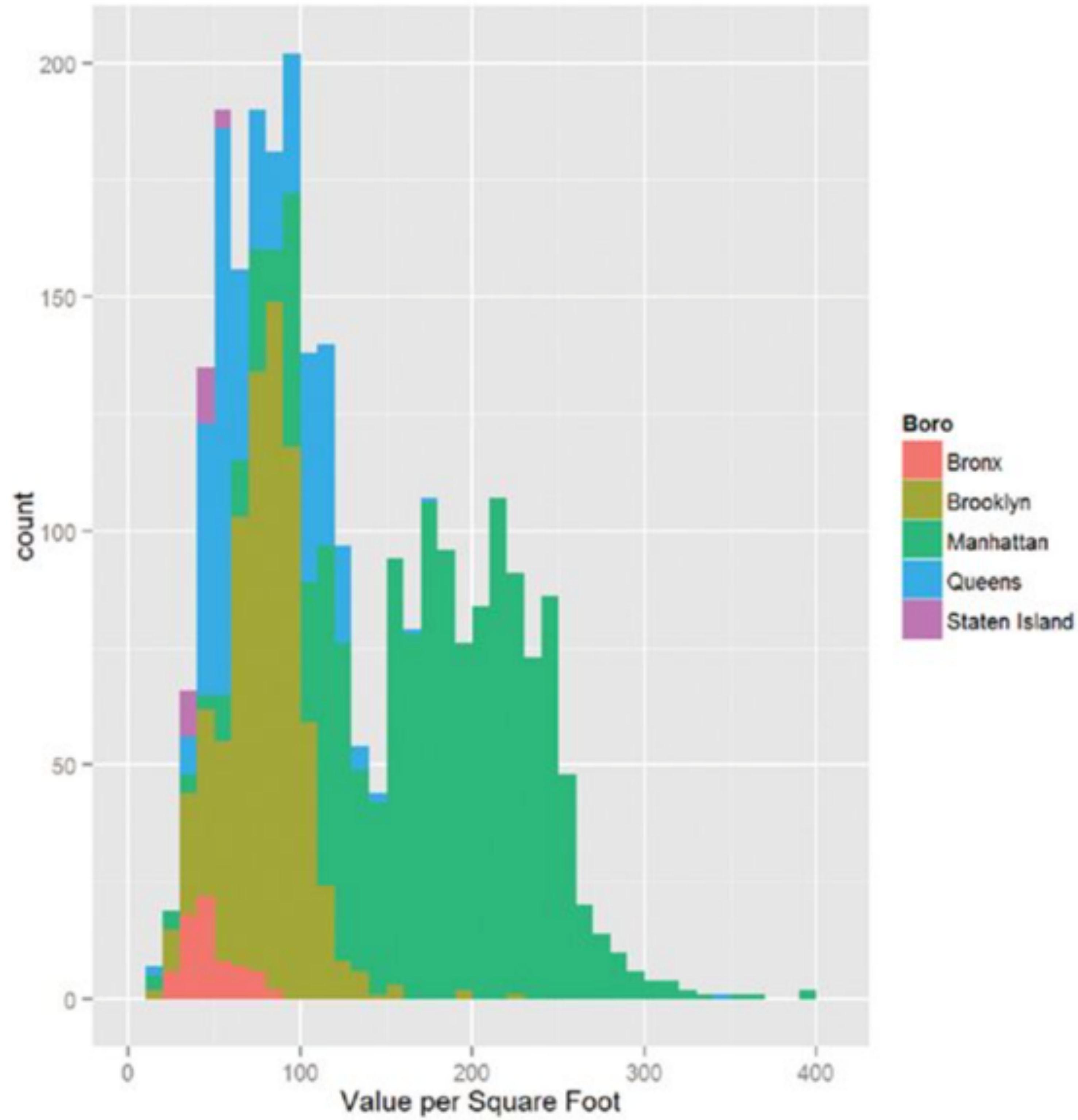


NYC Housing Example



Lander, R for Everyone; NYC Open Data

NYC Housing Example



Lander, R for Everyone; NYC Open Data

NYC Housing Example

- Response variable (y): price per square foot
- Predictor variables (x’s): number of units in complex, number of square feet, borough indicators
- Try linear regression; it may help to take logs of some of the continuous variables first

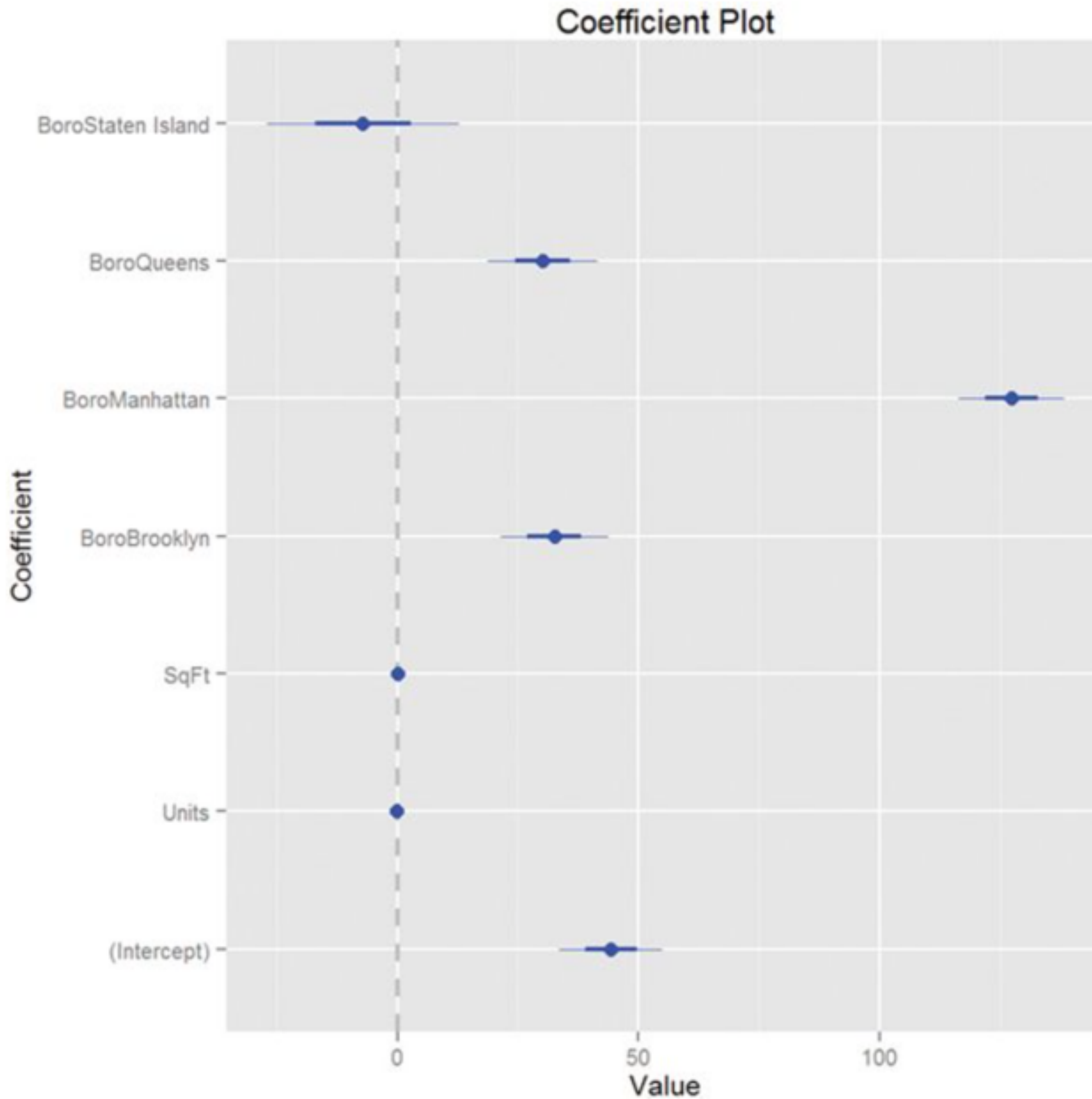
Lander, R for Everyone; NYC Open Data

NYC Housing Example

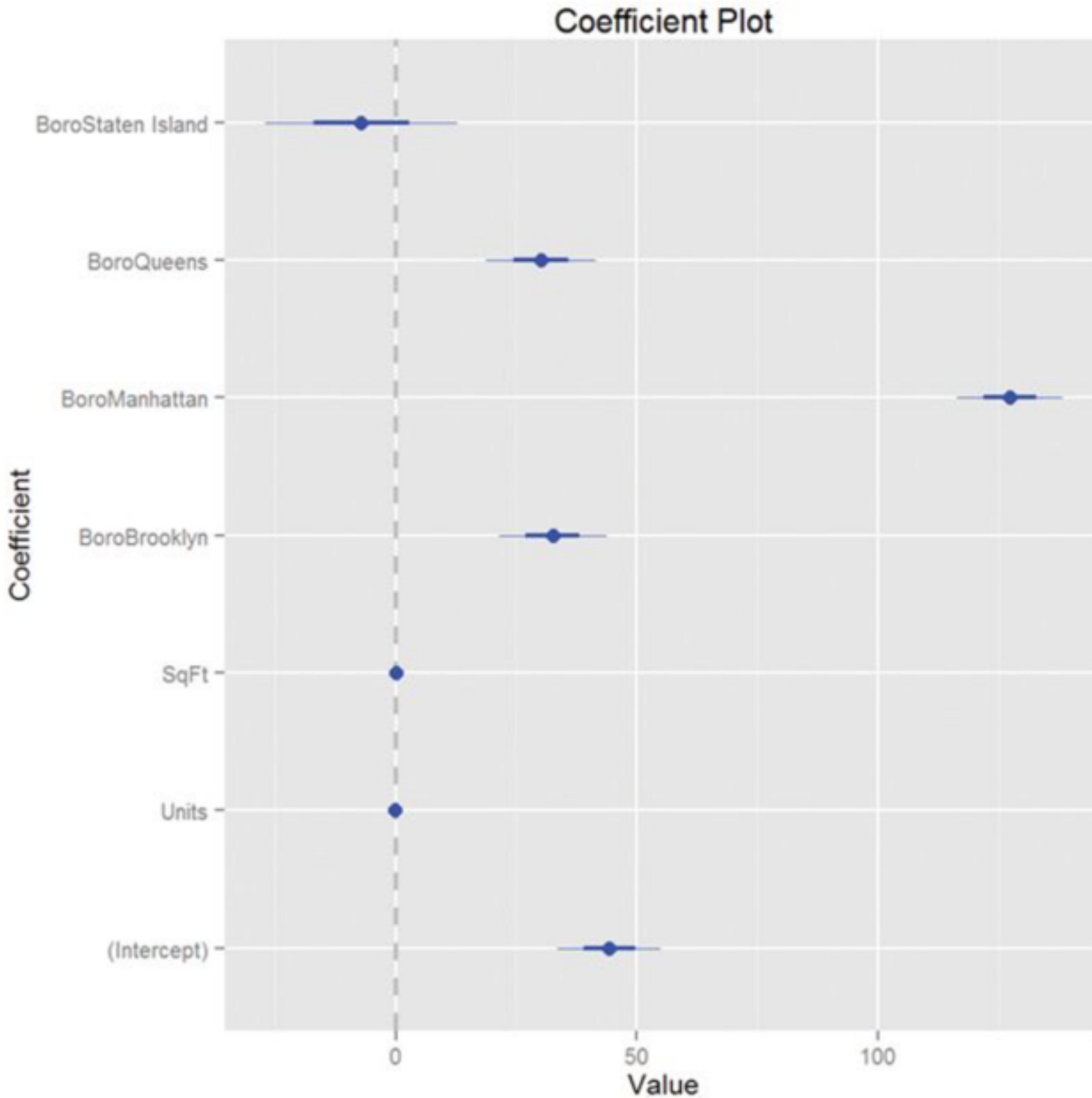
```
Coefficients:
                   Estimate   Std. Error t value  Pr(>|t|)
(Intercept)         4.430e+01     5.342e+00    8.293     <2e-16
***
Units              -1.532e-01     2.421e-02   -6.330   2.88e-10
***
SqFt                2.070e-04     2.129e-05    9.723    < 2e-16
***
BoroBrooklyn        3.258e+01     5.561e+00    5.858   5.28e-09
***
BoroManhattan       1.274e+02     5.459e+00   23.343    < 2e-16
***
BoroQueens          3.011e+01     5.711e+00    5.272   1.46e-07
***
BoroStaten Island -7.114e+00    1.001e+01  -0.711     0.477
---
```

Lander, R for Everyone; NYC Open Data

NYC Housing Example

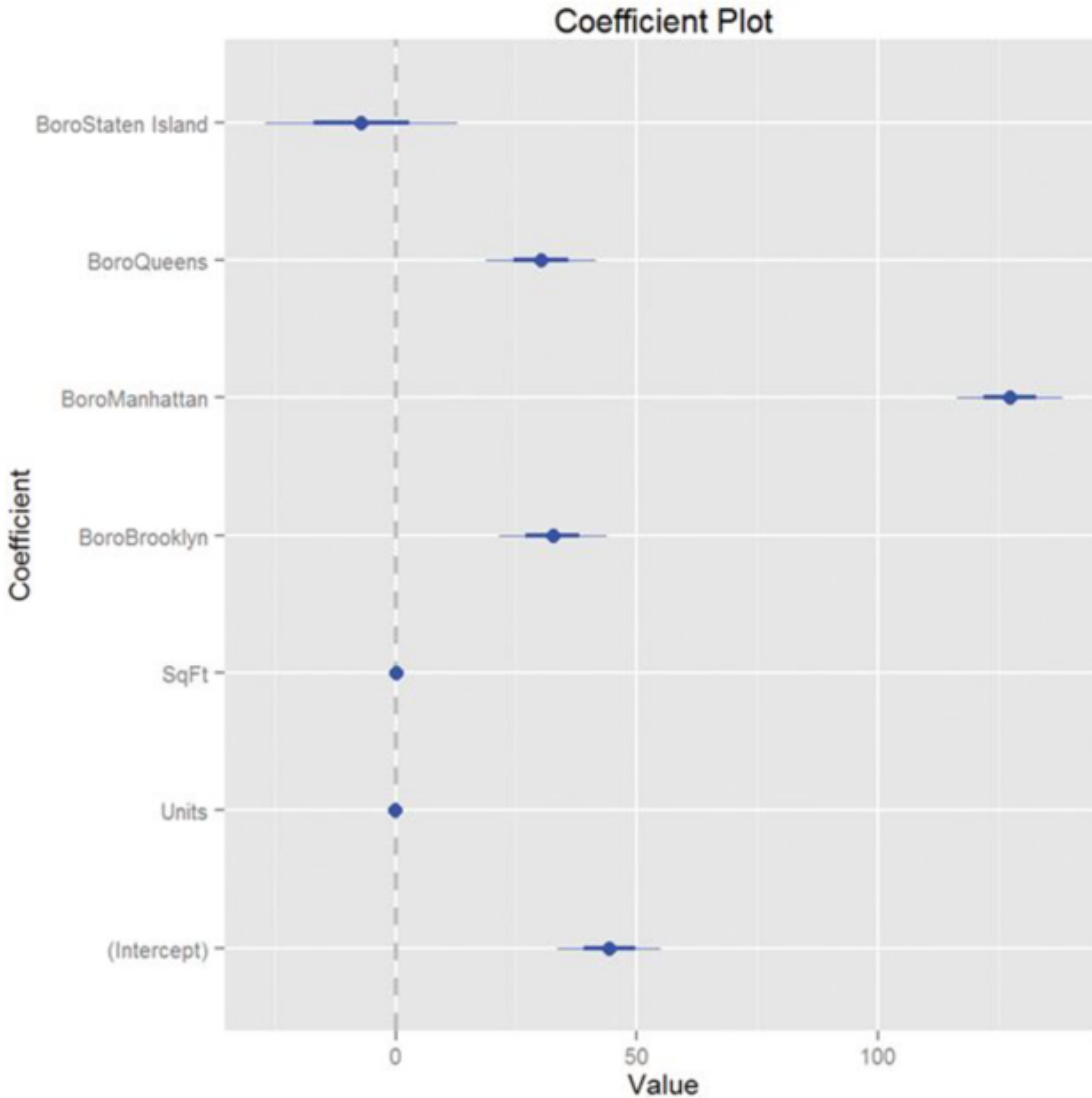


Lander, R for Everyone; NYC Open Data



Where did the Bronx go?
Why are SqFT and Units so close to 0?

Lander, R for Everyone; NYC Open Data



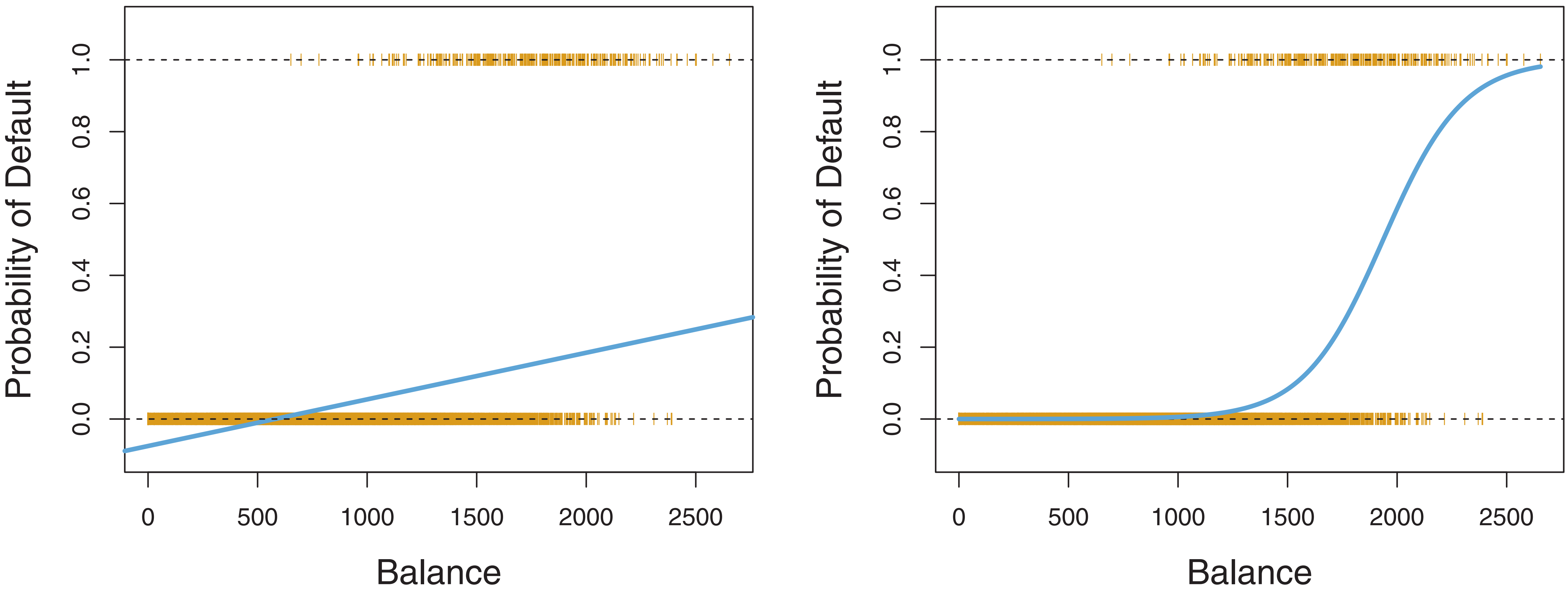
Why are SqFT and Units so close to 0?
Where did the Bronx go?

Lander, R for Everyone; NYC Open Data

Collinearity

- Should avoid having predictor variables that are highly correlated with each other (collinearity results in instability, high variances in estimates, and worse interpretability)
- An extreme case of collinearity would be also including a Bronx indicator in the NYC Housing example. Instead, use one borough as a baseline.

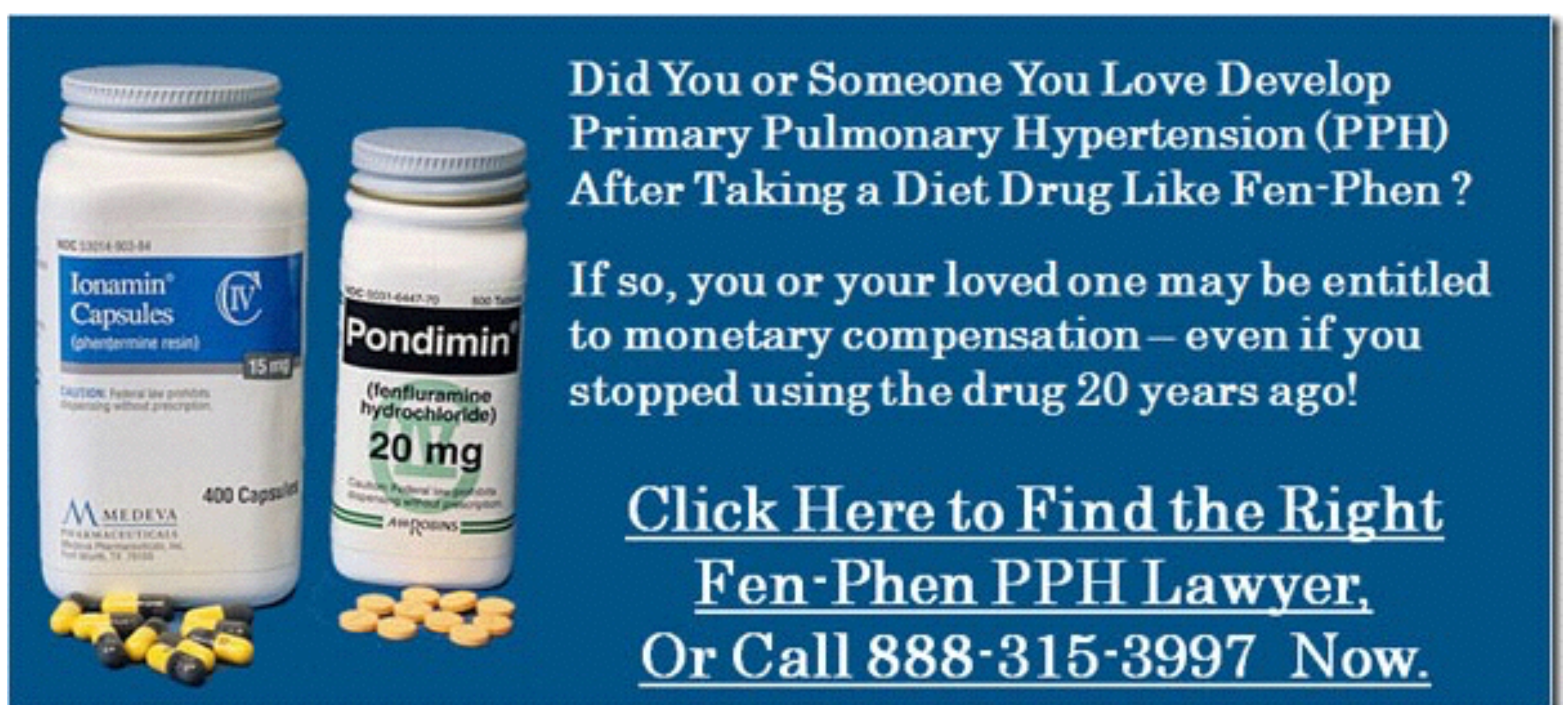
Predicting a Binary Response



source: Introduction to Statistical Learning, James, Witten, Hastie, Tibshirani, http://www-bcf.usc.edu/~gareth/ISL/

Fen-Phen Case Study

On July 8, 1997 Mayo Clinic investigators described 24 cases of valvular heart disease in patients taking the recently released appetite suppressant combination fen/phen (fenfluramine plus phentermine). The FDA issued an advisory to encourage reporting of similar cases.



Example from Stat S105 with Xiao-Li Meng, Yves Chretien

Strong Association?

Recall we obtained the following sample of patients in a follow-up study:

	Heart disease	No heart disease	Total
Fen/phen	53	180	233
Control	3	230	233

- So do you think there is strong association between heart disease and fen/phen usage?

- How would you defend your assertion scientifically? Can you just say "Well, 53/3=17.7 is very large to me"?

How about this one then?

Now suppose that instead of heart disease, you wanted to test whether fen/phen increased the risk of a rare type of cancer. Using the same patients, you observe that:

	cancer	No cancer	Total
Fen/phen	1	232	233
Control	0	233	233

Is that the strongest evidence of association one can ever get, since 1/0 is infinite?

Measures of Association: Odds Ratio

If someone's probability of experiencing an outcome is p , then that person's *odds* of the outcome are $p/(1-p)$

The *odds ratio* is the ratio of two different people's odds of some outcome. If people in group A have probability p_A of disease, and people in group B have probability p_B, then the odds ratio of group A vs. group B is

$$\text{Odds Ratio} = \frac{p_A}{1-p_A} \Big/ \frac{p_B}{1-p_B} = \frac{(1-p_B)p_A}{(1-p_A)p_B}$$

Crude Odds Ratio Estimate

The data in one study were as follows:

		Fen/phen		
		+	-	
Aortic Regurgitation	+	6	162	168
	-	13	2343	2356
		19	2505	2524

A crude estimate of the odds ratio is

$$\frac{6 \times 2343}{13 \times 162} = 6.7$$

Palmieri V, Arnett DK, Roman MJ, Liu JE, Bella JN, Oberman A, Kitzman DW, Hopkins PN, Morgan D, de Simone G, Devereux RB. Appetite suppressants and valvular heart disease in a population-based sample: the HyperGEN study. Am J Med. 2002 Jun 15;112(9):710-5.

What about confounding factors?

But what if there are confounding factors? For example, what if fen/phen users are more likely to be obese, and obesity increases the risk of heart disease?

We can set up a *logistic regression model* to predict a person's odds of heart disease, given the predictor variables.

We can also use this to *compare* fen/phen users vs. non-fen/phen users, controlling for the other predictors.

Then we can use the data to estimate the parameters, using Maximum Likelihood Estimation (MLE).

Variables in the model

$$Y = \begin{cases} 1, & \text{if cardiac valve abnormality} \\ 0, & \text{if not} \end{cases}$$

$$X_{fen} = \begin{cases} 1, & \text{if taking fen/phen} \\ 0, & \text{if not} \end{cases}$$

$$X_{age} = \text{subject's age}$$

$$X_{sex} = \begin{cases} 1, & \text{male} \\ 0, & \text{if female} \end{cases}$$

(plus other X-variables...)

$$p = P\left(Y = 1 \mid X_{fen}, X_{age}, X_{sex}, \ldots, X_k\right)$$

So, how is p related to all of these factors?

A logistic regression model

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{fen}X_{fen} + \beta_{age}X_{age} + \beta_{sex}X_{sex} + \cdots + \beta_k X_k$$

The parameters of the model (the β's) are unknown, and are estimate from the data using MLE.

This gave 1.84 as an estimate for the fen/phen parameter. How can that be interpreted?

Two patients, A and B, are the same age, same gender, and similarly identical on all other variables. Patient A has taken fen/phen and Patient B has not. The model predicts that

$$\text{logit}(p_A) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{age}X_{age} + \beta_{sex}X_{sex} + \cdots + \beta_k X_k + \beta_{fen}X_{fen}$$

$$\text{logit}(p_B) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{age}X_{age} + \beta_{sex}X_{sex} + \cdots + \beta_k X_k$$

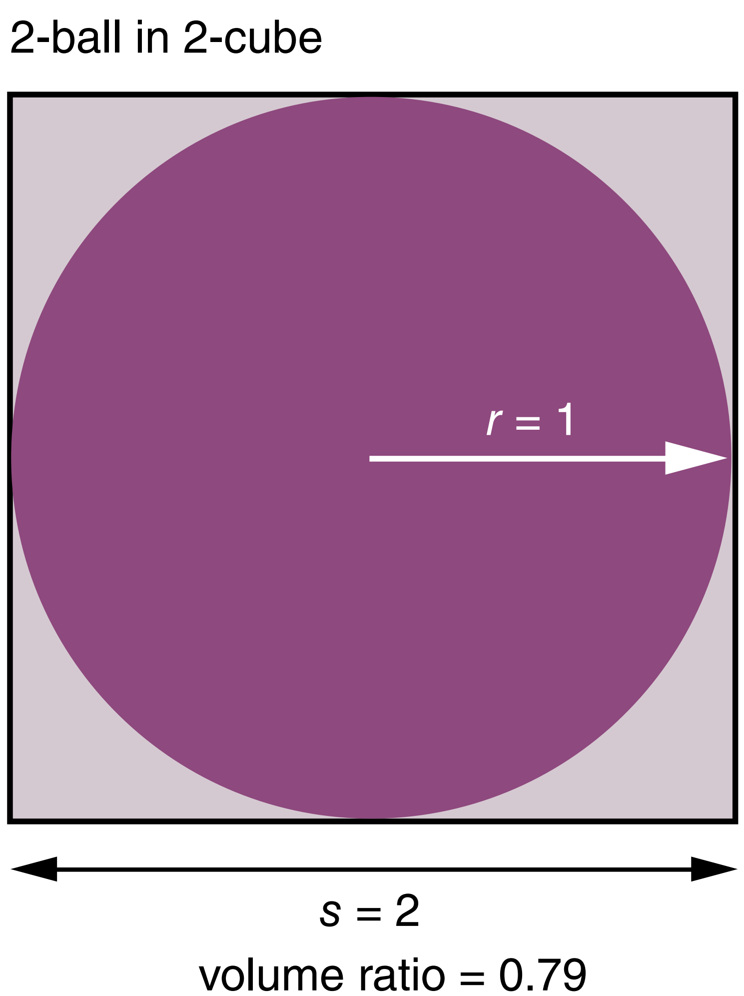
$$\beta_{fen} = \text{logit}(p_A) - \text{logit}(p_B) = \ln\left(\frac{\frac{p_A}{1-p_A}}{\frac{p_B}{1-p_B}}\right)$$

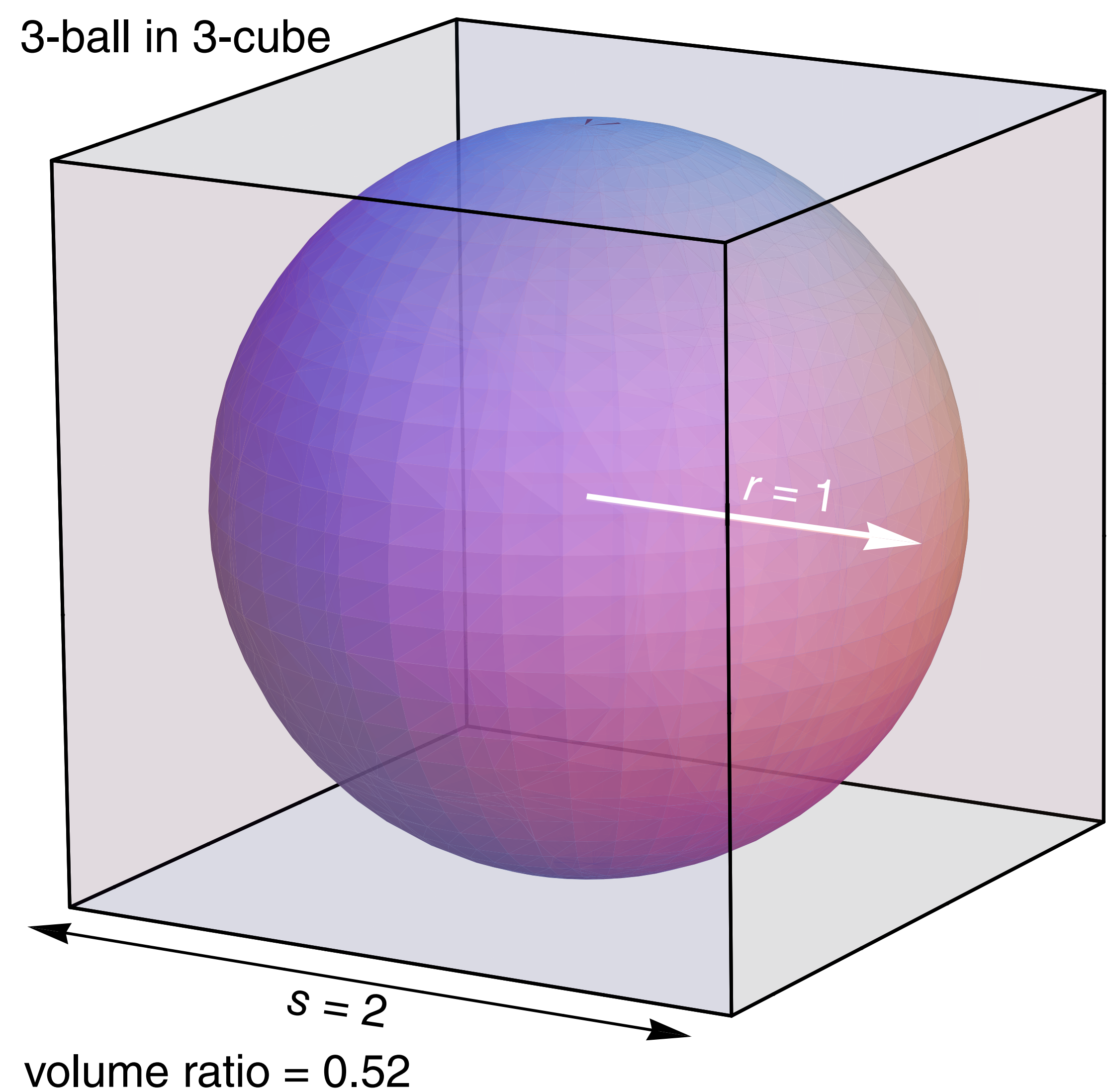
Using this model we can estimate an “adjusted” odds ratio that’s the odds ratio for two people with all other known factors held constant:

$$e^{\hat{\beta}_{fen}} = e^{1.84} \approx 6.3$$

Curse of Dimensionality

For a uniformly random point in a box with side length 2, what is the probability that the point is in the unit ball?





source: An Adventure in the nth Dimension, Brian Hayes, American Scientist 2011

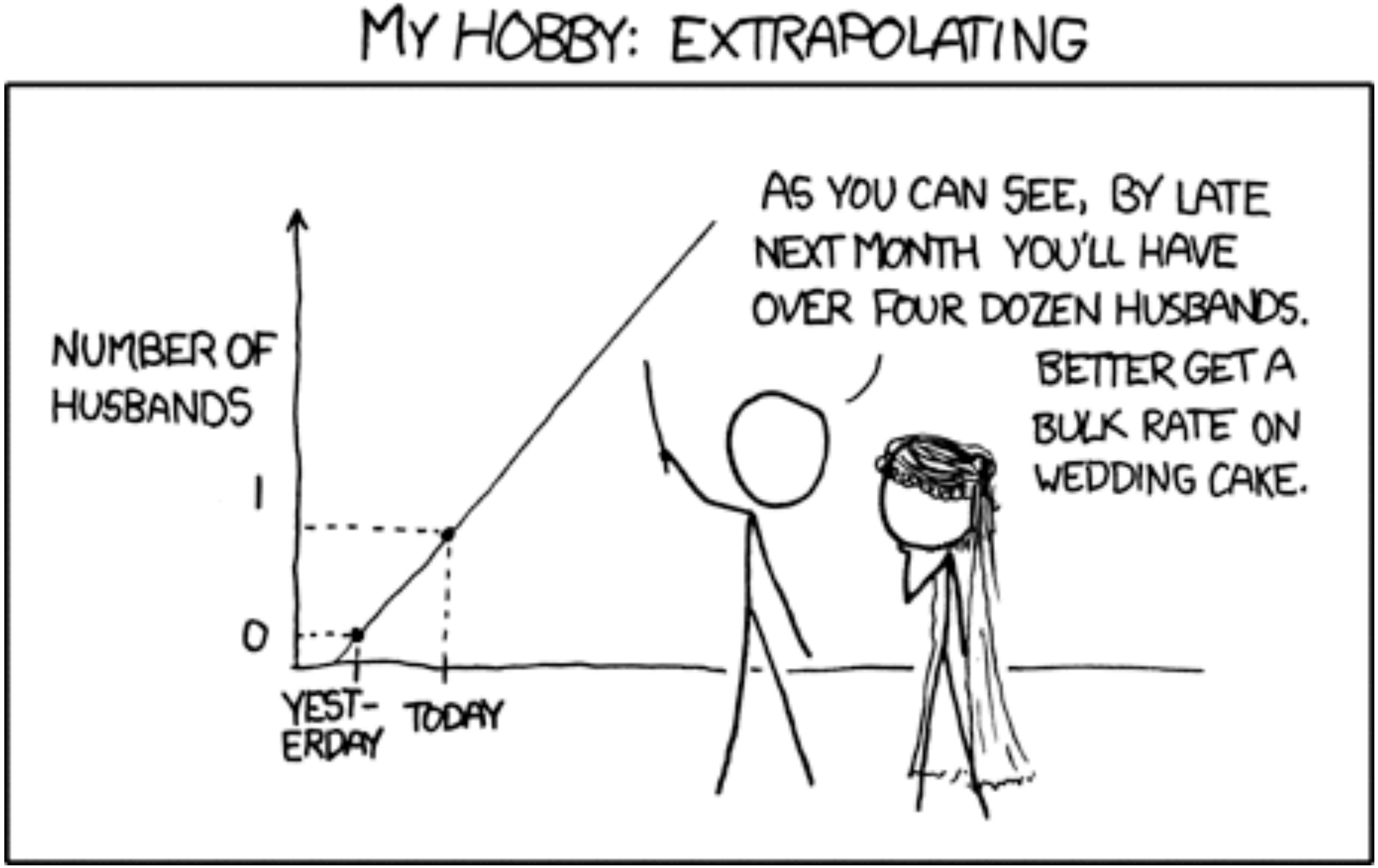
Curse of Dimensionality

For a uniformly random point in the box in d dimensions with length 2 in each dimension, what is the probability that the random vector is in the unit ball in d dimensions?

d	probability
2	0.79
3	0.52
6	0.08
10	0.002
15	0.00001
100	$1.87 \cdot 10^{-70}$

In many high-dimensional settings, the vast majority of data will be near the boundaries, not in the center.

Interpolation vs. Extrapolation



source: https://xkcd.com/605/

In high dimensions, nearest neighbor point tends to be very far away.
May be very hard to interpolate well, even with a lot of data points.

Blessing of Dimensionality

In statistics, “curse of dimensionality” is often used to refer to the difficulty of fitting a model when many possible predictors are available. But this expression bothers me, because more predictors is more data, and it should not be a “curse” to have more data....

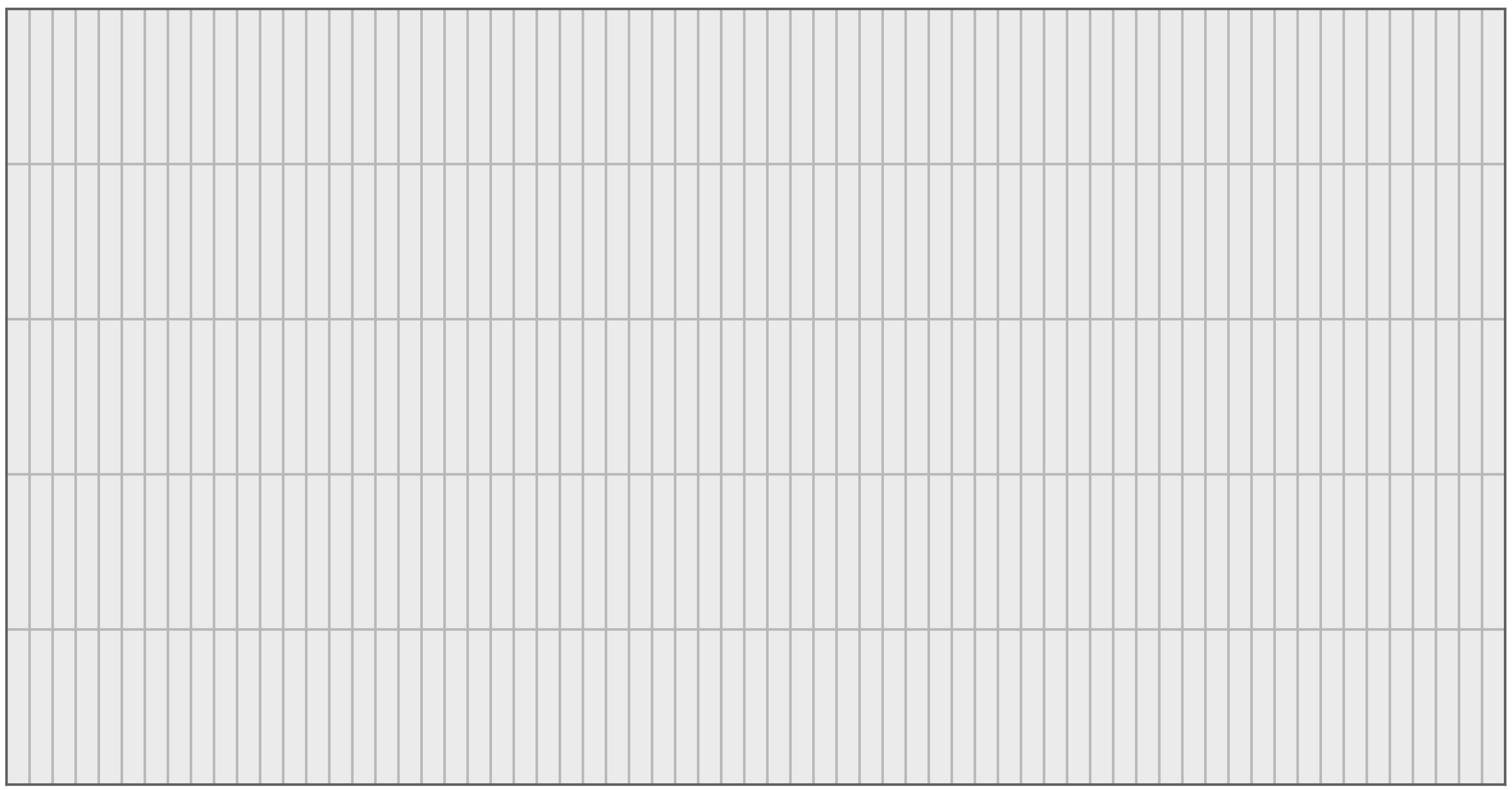
With multilevel modeling, there is no curse of dimensionality. When many measurements are taken on each observation, these measurements can themselves be grouped. Having more measurements in a group gives us more data to estimate group-level parameters (such as the standard deviation of the group effects and also coefficients for group-level predictors, if available).

In all the realistic “curse of dimensionality” problems I’ve seen, the dimensions–the predictors–have a structure. The data don’t sit in an abstract K-dimensional space; they are units with K measurements that have names, orderings, etc.

Andrew Gelman, http://andrewgelman.com/2004/10/27/the_blessing_of/

Tall data vs. wide data

vs.



n rows (individuals), p columns (variables) p measurements

n people

Wide data are increasingly common in applications, e.g., neuroimaging, microarrays, MOOC data. But many traditional statistical methods assume n greater than p.

Ridge Regression and Shrinkage

In a linear regression model, in place of minimizing the sum of squared residuals, ridge regression says to minimize

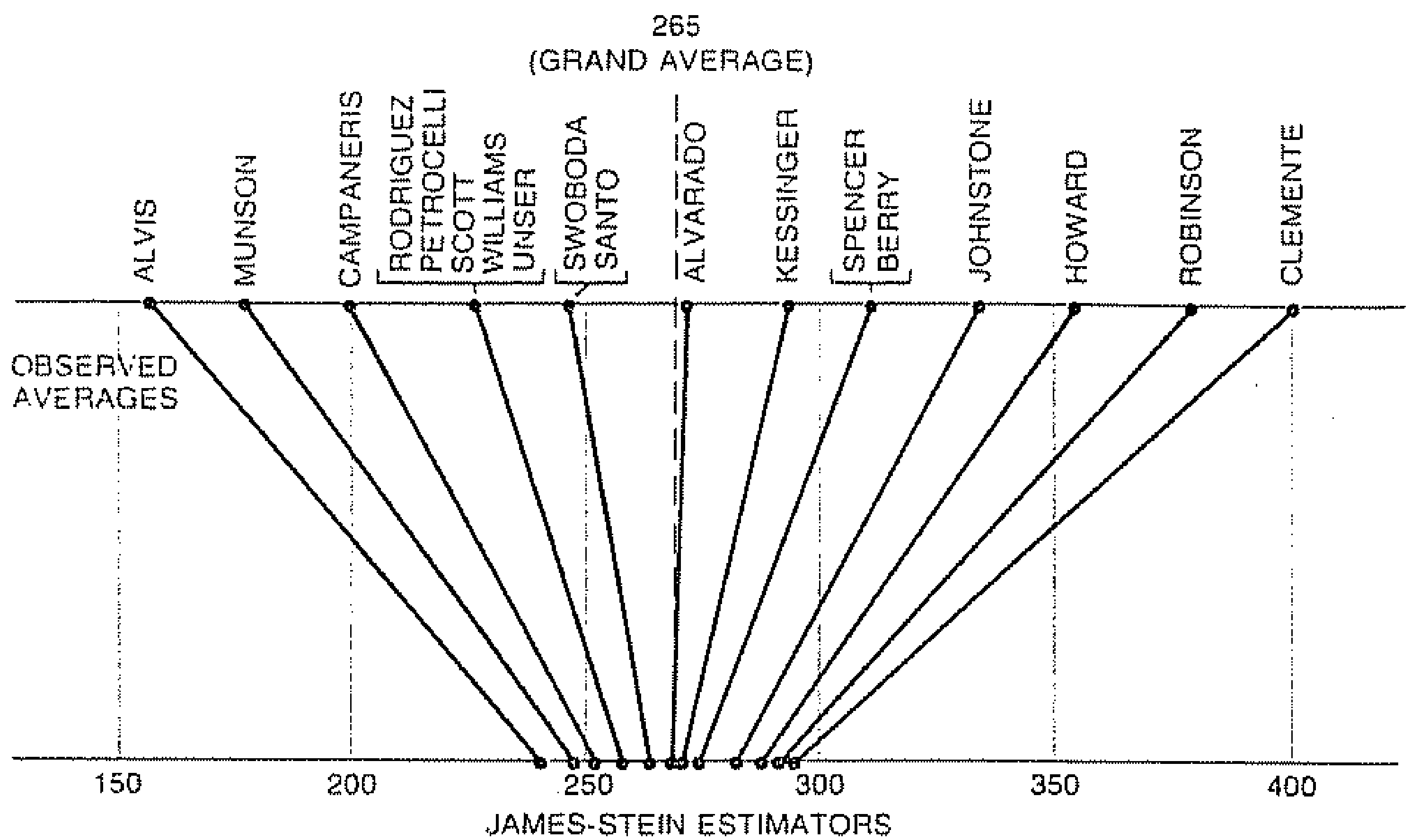
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2,$$

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

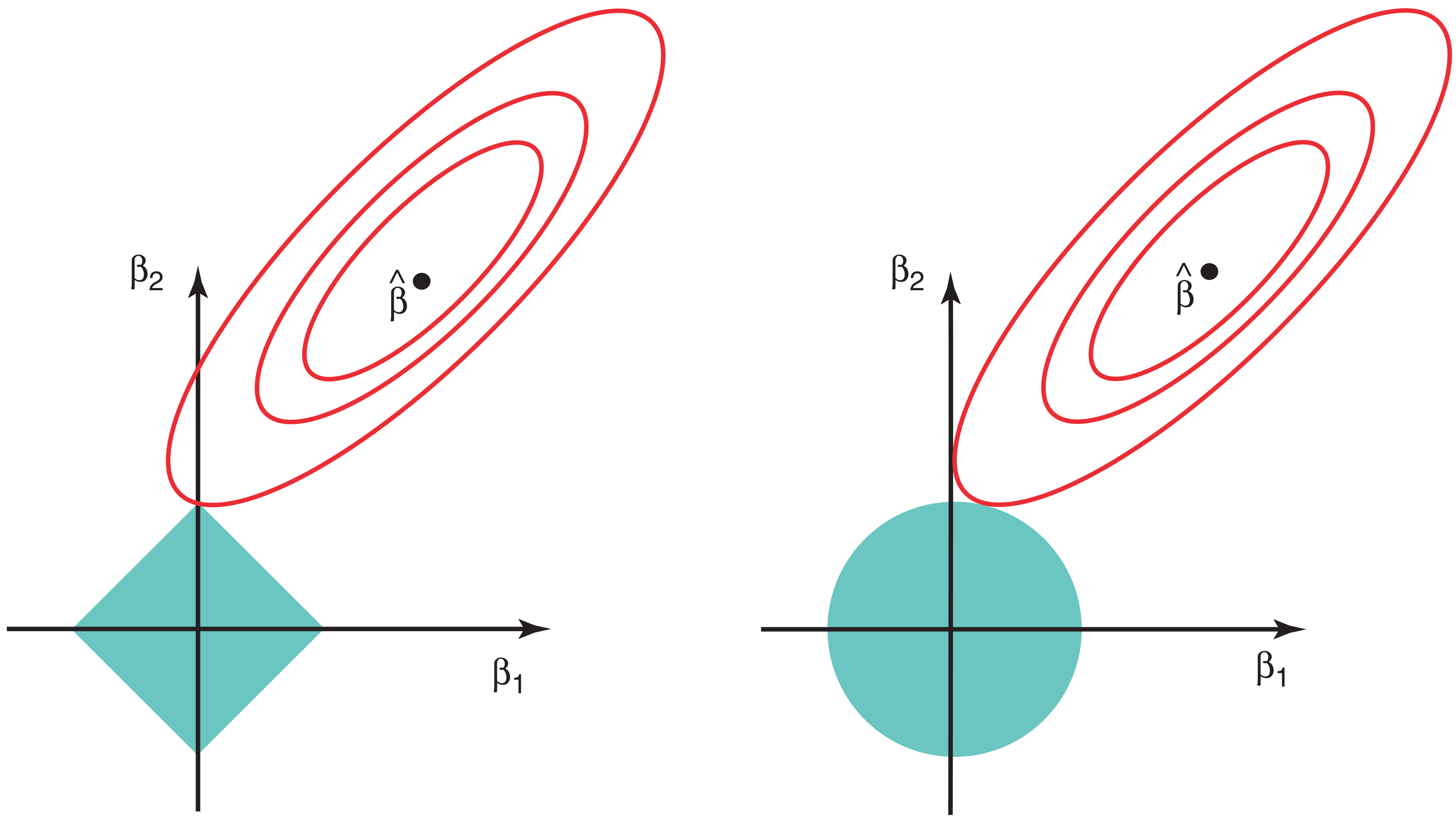
LASSO and Sparsity

In a linear regression model, in place of minimizing the sum of squared residuals, LASSO says to minimize

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

This helps induce *sparsity*, reducing the number of variables one has to deal with.

LASSO vs. Ridge Constraints



source: Introduction to Statistical Learning, James, Witten, Hastie, Tibshirani, http://www-bcf.usc.edu/~gareth/ISL/