This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike License. Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.

Copyright 2006, The Johns Hopkins University and Scott L. Zeger. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided “AS IS”; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.

How to Choose the Wrong Model

Scott L. Zeger
Department of Biostatistics
Johns Hopkins Bloomberg School

Questions

What is a model?

Which is the best (true, right) model?

How can you choose a useful model?

What are the 10 best ways to choose the wrong model?

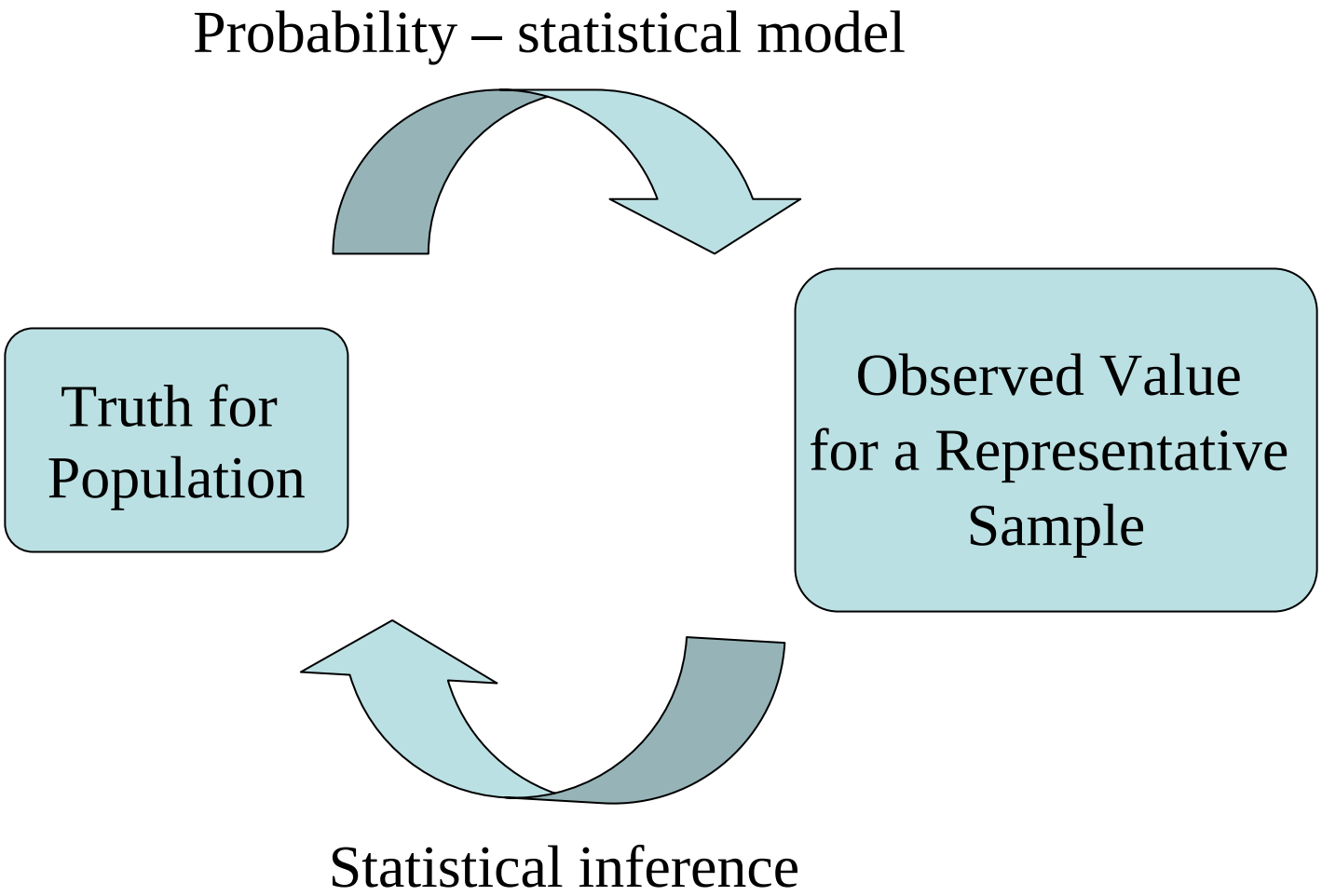
What is a statistical model?

Tool for those empirical sciences where signals come embedded in noise

Lens through which to view data to better understand the signal

Tool for quantifying the evidence in data about a particular truth we seek

Empirical science: search for “truth”



Data as Evidence

Schizophrenia “Relapse” within 400 days

from Figure 1 in Csernansky, et al. 2002

	Risp	Hal	Total
Relapse	41	67	108
No Relapse	136	121	257
Total	177	188	365
	41/177=0.23	67/188=0.35	

0.23/0.35 = 0.65

True Relative Rate of Relapse

Truth: population rate of relapse for haloperidol;
relative rate for risperidone versus haloperidol

Probability model: patients are independent;
"binomial" probability models for each group

Statistical inference: use data and the statistical model to make statements about true relative rate

Statistical Model: Probability of the Observed Data as a Function of Unknown True Parameters (rate for halo ; relative rate)

p = rate for halo

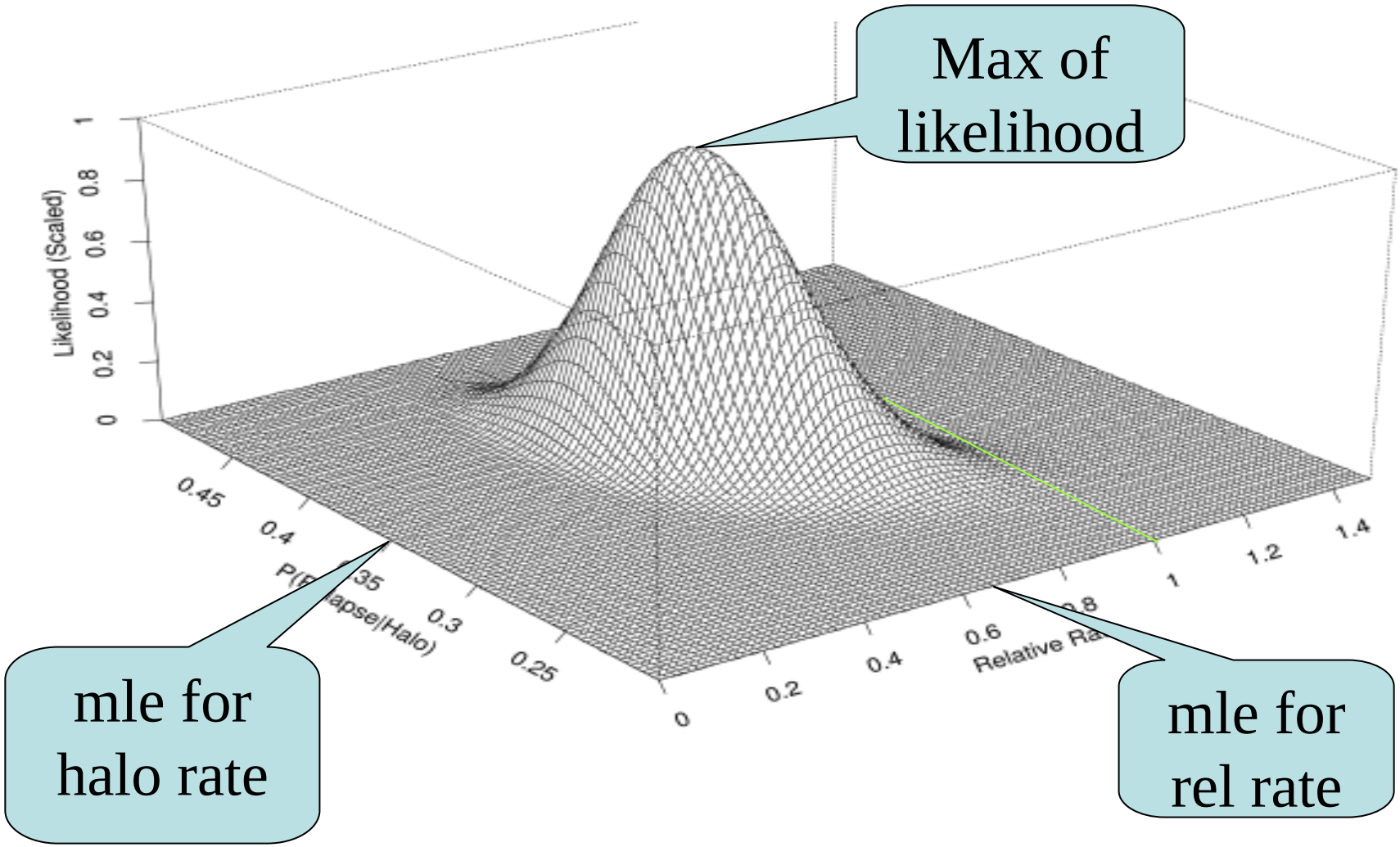
rr = relative rate – risp/halo

the likelihood function of the data is:

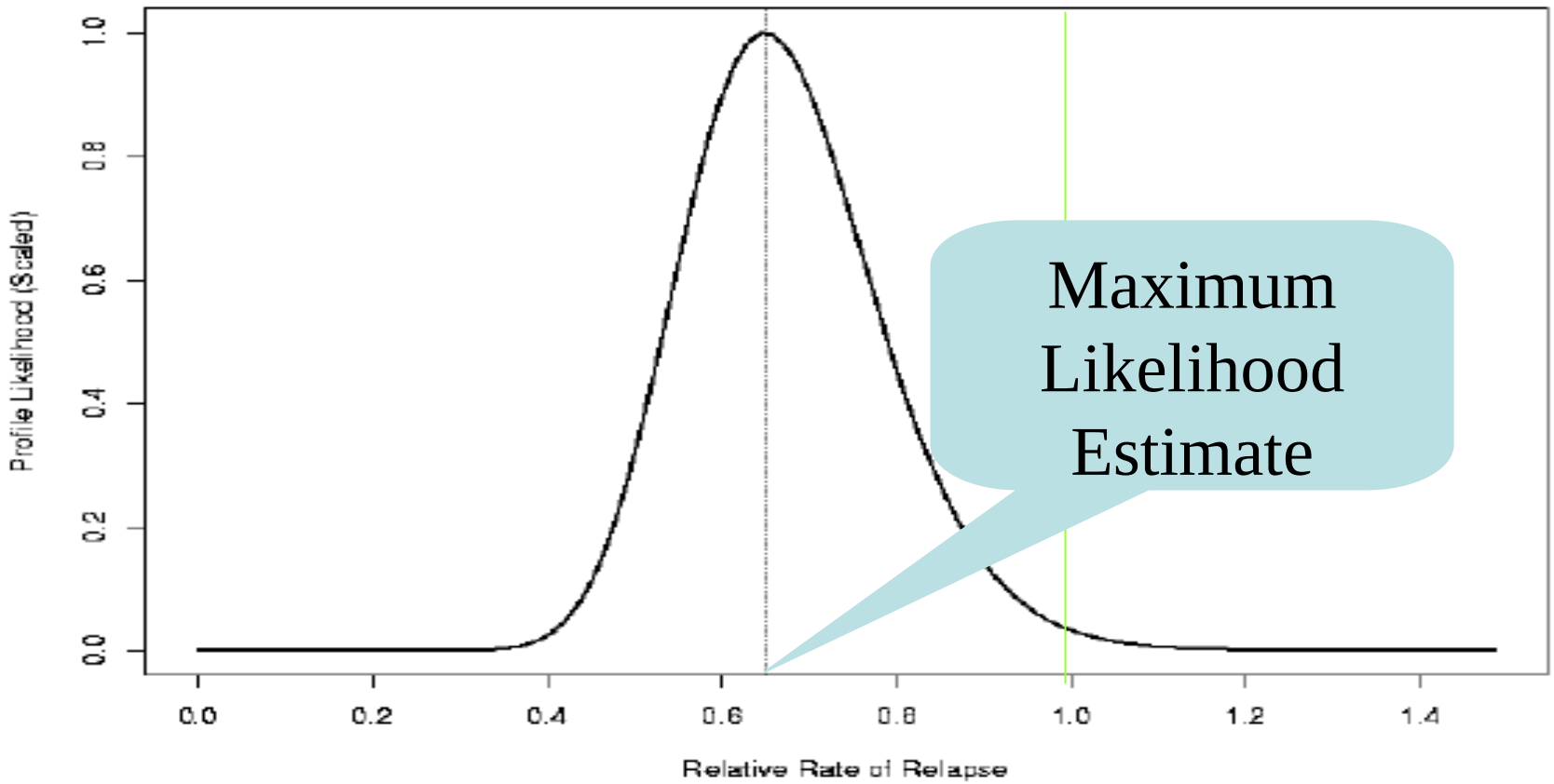
$$(rr \times p)^{41}(1 - rr \times p)^{136}\, p^{67}(1 - p)^{121}$$

likelihood function

Probability of Observed Data



Profile Likelihood Function



Which is the best (true, right) model?

Which is the best (true, right) bottle of beer? (only 20 minutes, Karl)

Which is the best (true) screwdriver?

Who is the best model for this hat?

There is NO Uniformly BEST model

Each has unique merits

Choice depends on your specific objective

How to Choose the Best Model

You can not choose the best model because there isn’t one

You can choose a useful model (think screwdriver)

You can explore and report how your scientific findings vary over a set of other useful models

You can average your results across useful models

How to Make a Useful Model

Approximate mechanisms to an adequate degree

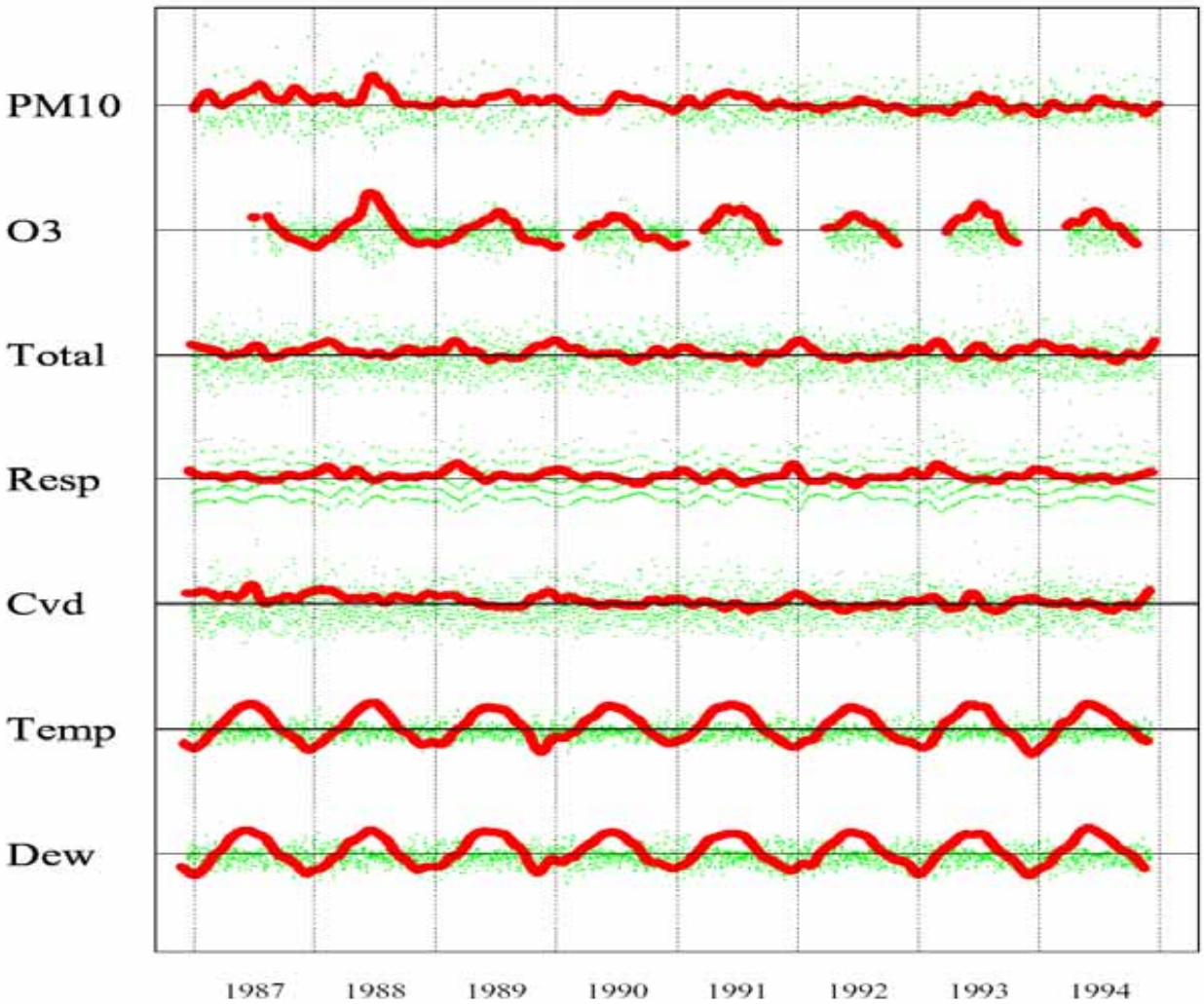
Represent scientific question with one or a few unknown parameters

Be as flexible as possible with respect to the rest of the probability model

Define your model as a member of a broader class so you can use many models, not just one

Does Air Pollution Cause Mortality Cities?

Data for Baltimore, Maryland



Does Air Pollution Cause Mortality?

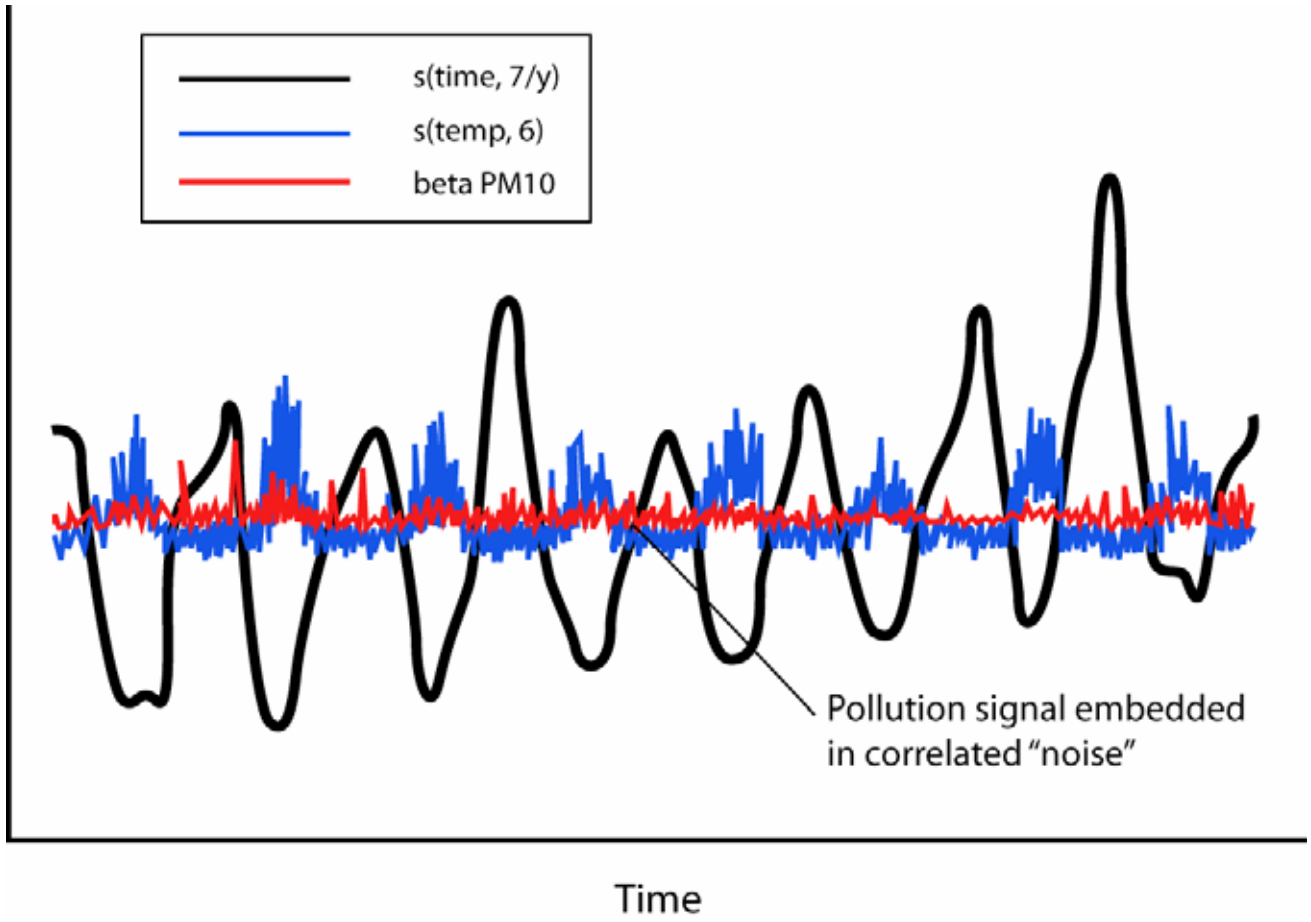
- Y_t^c is the mortality count on day t at location c
- PM_{10t}^c is the level of particulate matter on day t at location c
- β^c is the relative rate of mortality, e.g. the percentage increase in mortality associated with 10 $\mu g/m^3$ increase in PM_{10}
- $s(\text{temp}, df)$ are smooth functions (smoothing splines or regression splines)

$$\begin{aligned}\log E[Y_t^c] &= \text{age-specific intercept} + \beta^c PM_{10t}^c + s(\text{time}, 7/\text{year}) + \\ &+ s(\text{temp}, 6) + s(\text{dewpoint}, 3) + \text{age} \times s(\text{time}, 8) + ...\end{aligned}$$

- estimate $\hat{\beta}^c$ and its statistical variance v^c within each location
- Generalized additive models (GAM) with smoothing splines or Generalized Linear Models (GLM) with natural cubic splines are the methods of choice

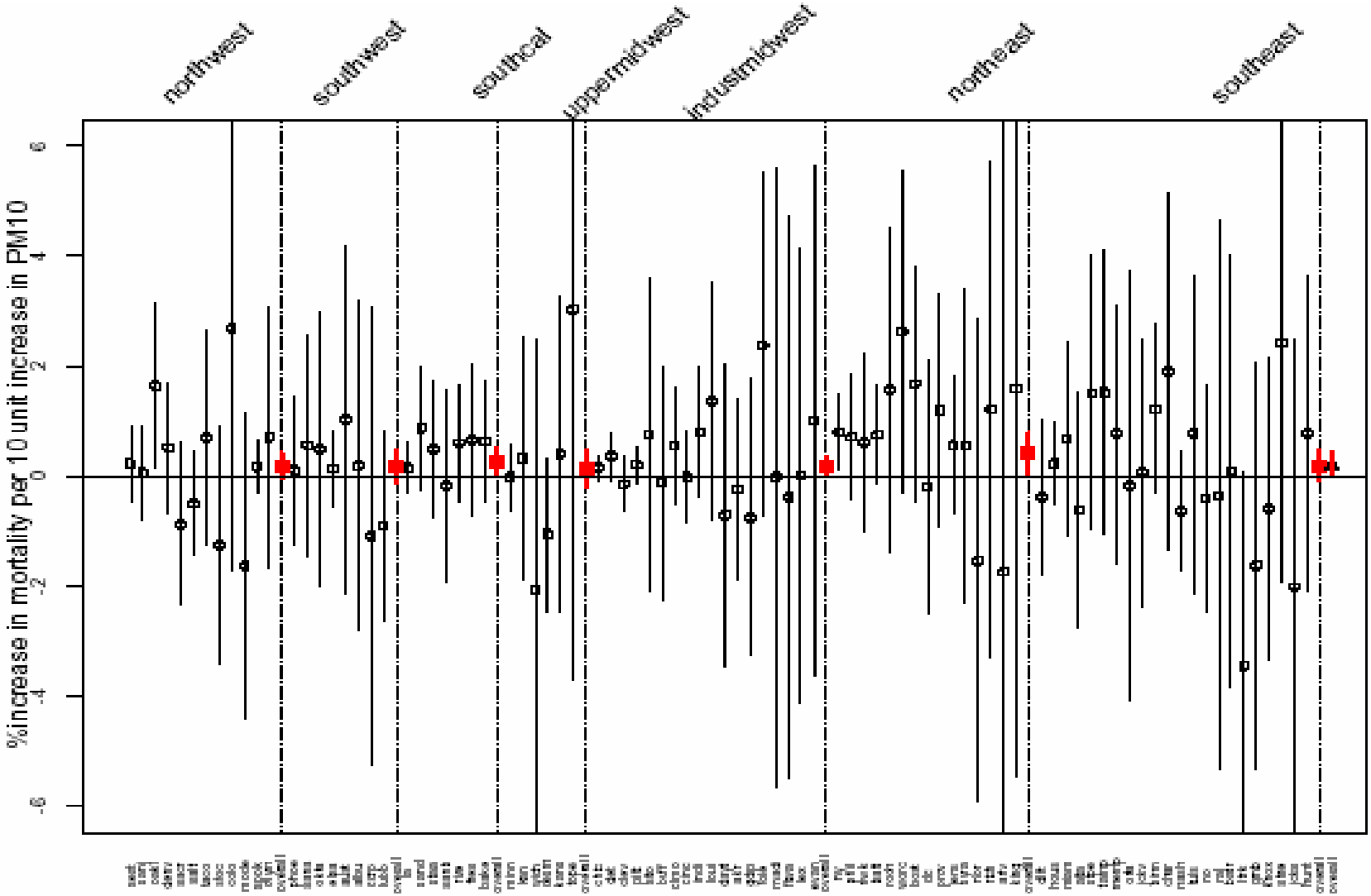
Source: Scott Zeger

Hard (interesting) Statistical Problem



Source: Scott Zeger and CTLT

City-specific Estimates



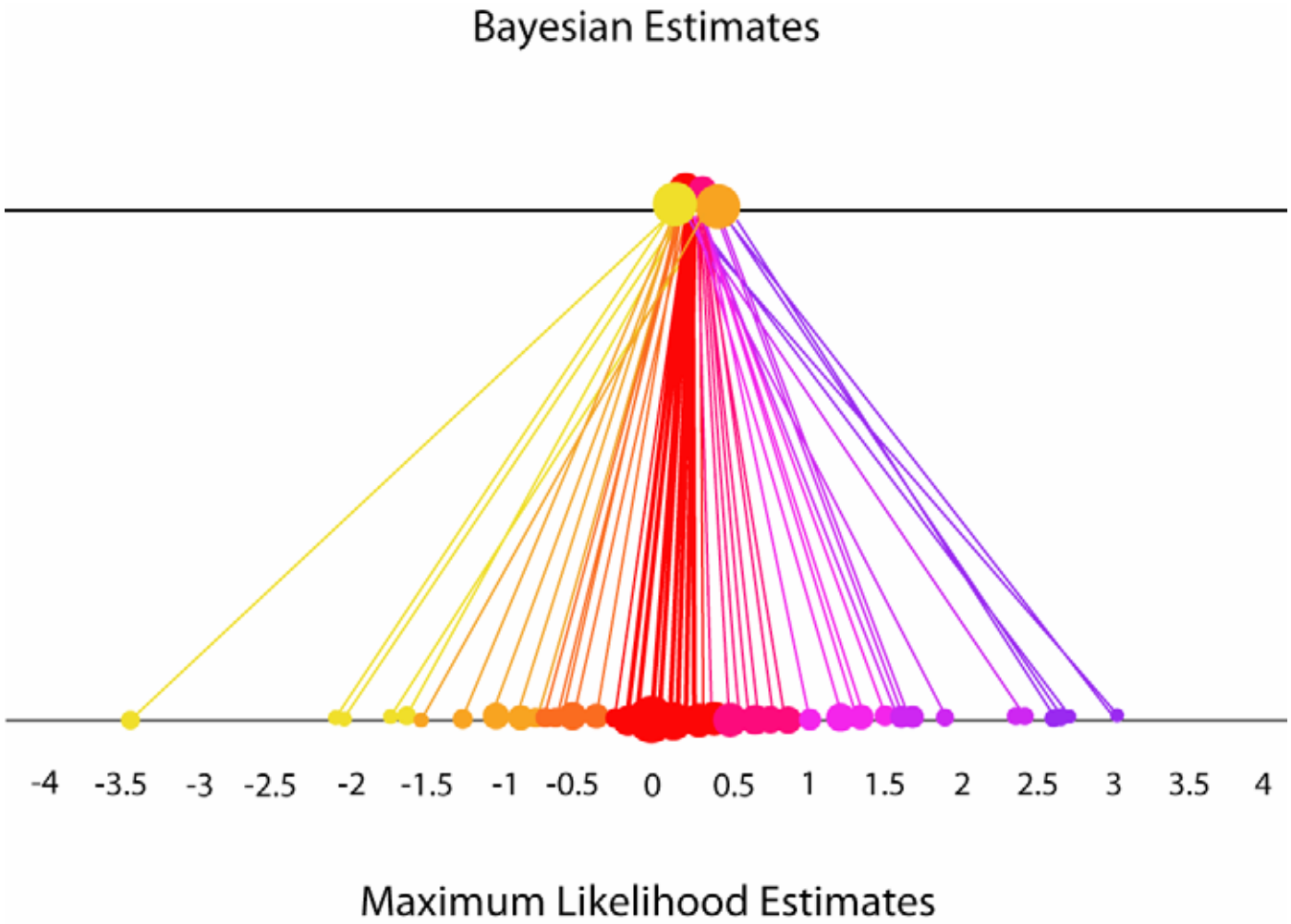
Source: Scott Zeger

Spatial Model for Relative Rates

- β_r^c = log relative rate of mortality associated with PM_{10} in city c in region r
- α_r = average log relative rate in region r
- α^* = national average log relative rate

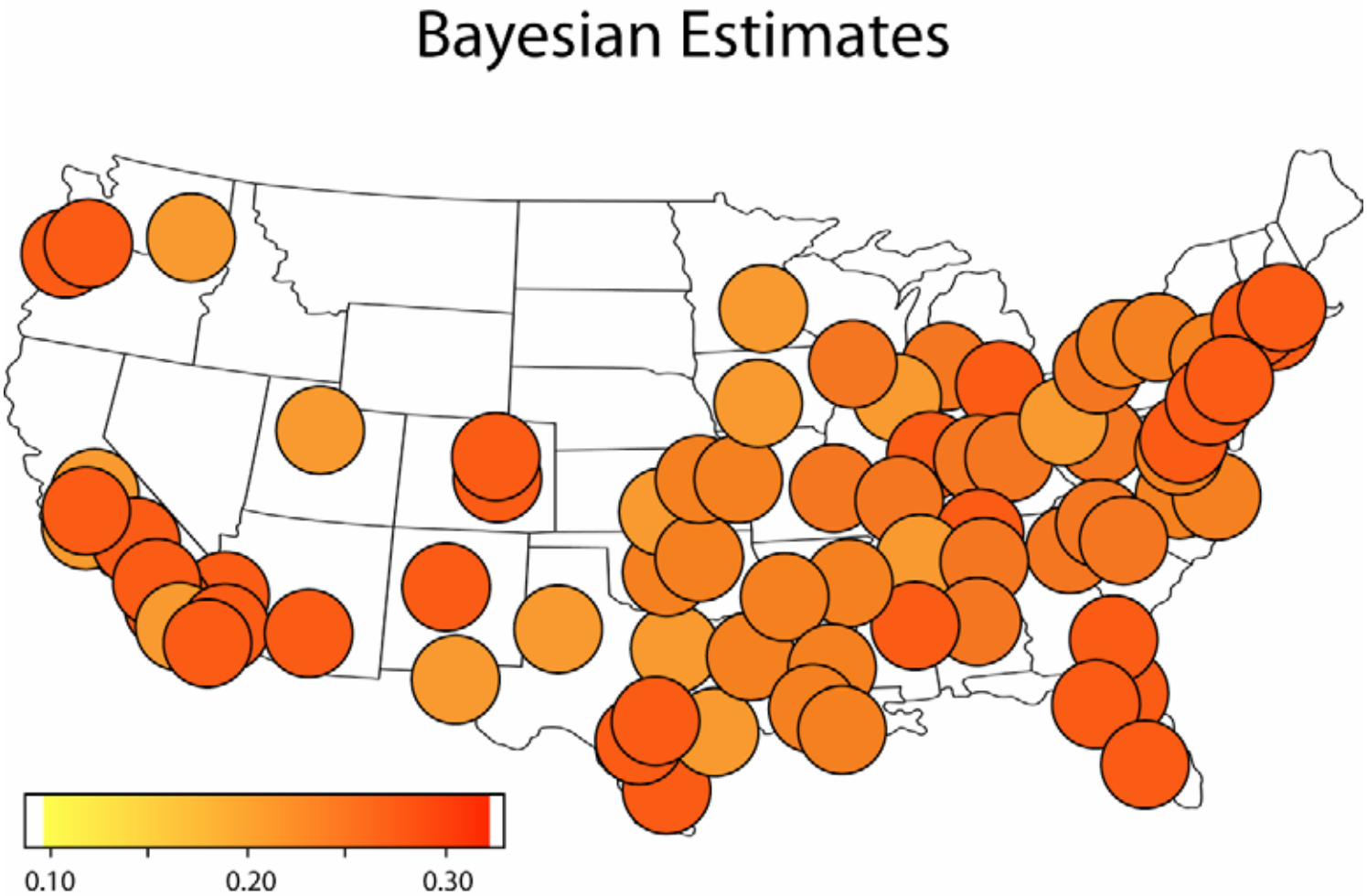
$$\hat{\beta}_R^c = \alpha^* + \underbrace{(\hat{\beta}_r^c - \beta_r^c)}_{\text{statistical error}} + \underbrace{(\beta_r^c - \alpha_r)}_{\text{heterogeneity within region}} + \underbrace{(\alpha_r - \alpha^*)}_{\text{heterogeneity across regions}}$$

Shrinkage



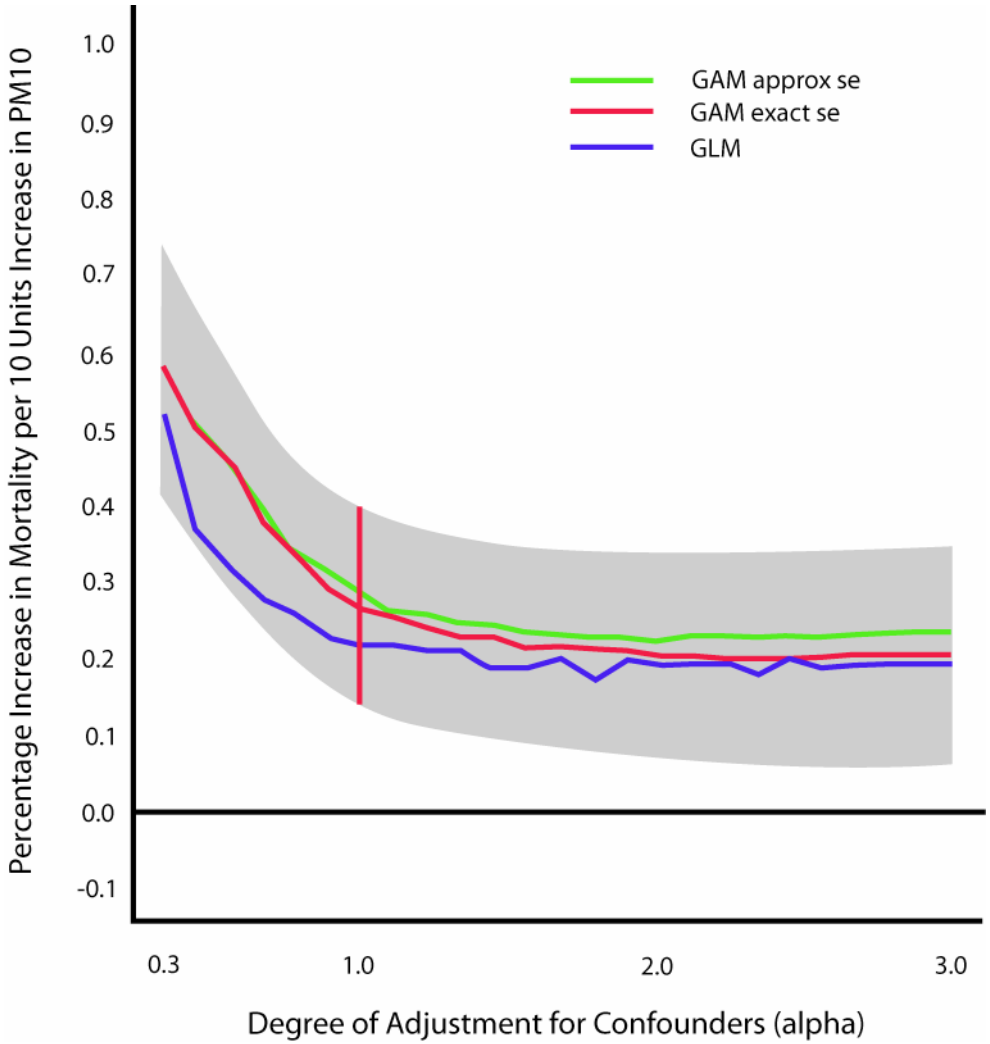
Adapted by CTLT from: Zeger, S.L., Dominici, F., McDermott A., Samet, J.M. "Modeling the Health Effects of Environmental Exposures: Particulate Air Pollution and Mortality in Urban American," In: *Monitoring the Health of Populations: Statistical Methods for Public Health Surveillance*, edited by Brookmeyer R and Stroup D, Oxford University Press, 2004.

Bayes Posterior Estimates



Adapted by CTLT from: Dominici, F., McDermott, A., Zeger, S.L., Samet, M.M., "National Maps of the Effects of Particulate Matter on Mortality: Exploring Geographical Variation," *Environmental Health Perspectives*, Vol. 111(1):39-43, 2003.

Many Models are Better Than the Best



Adapted by CTLT from: Dominici F. McDermott A. Hastie T., “Improved Semi-parametric Time Series Models of Air Pollution and Mortality, *Journal of American Statistical Association*, 468, 938-948, 2004.

Top Ten Ways to Choose the Wrong Model

10. Fail to take home the main message of this talk: there is not one best model; models are useful, not true (you can't get to Chicago in a model airplane)

9. Turn your scientific problem over to a computer that, knowing nothing about your science or your question, is very good at optimizing AIC or BIC or Cross-validated MIC(Key, Mouse)

8. Turn your scientific problem over to your neighborhood statistician, who, not trained at the Bloomberg School and knowing nothing about your science or your question, is very good at optimizing AIC or BIC or Cross-validated MIC(Key, Mouse)

Top Ten Ways to Choose the Wrong Model

7. Choose a model that presumes the answer to your scientific question and ignores what the data say (this is not a useful model in science; it can be quite useful in policy analysis)

6. Choose a model that no one can understand – not the persons who did the study, not their readers (not even yourself)

5. Choose a model that is unique in the world – “it is the only one to use because there is none other like it”

Top Ten Ways to Choose the Wrong Model

4. Use the Cox proportional hazards “model” because it has been used in 14,154 articles cited in PubMed – how could so many people be wrong

3. Since there is no best model of beer, go next door and buy Karl Broman an Amstel Light, which Zeger says must have its own virtues. See what he says about “best”

2. Take a wrong turn on the way to the Paris Versace spring show and end up viewing the Paris Hilton models.