Bayesian and Frequentist Issues in Modern Inference

Bradley Efron

Stanford University

Small Data

- Classical statistics Direct tests and estimates of individual parameters within well-defined models (MLE, Neyman–Pearson)
- **Not much:**
  - data-based model selection
  - Bayesian combination of related problems
- *Today* Methodology (not Philosophy)

Bayesian Inference

- Parameter: $\mu \in \Omega$
- Observed data: x
- Prior: $g(\mu)$
- Probability distributions: $\{f_\mu(x), \mu \in \Omega\}$
- Parameter of interest: $\theta = t(\mu)$

$$E\{\theta|x\} = \int_\Omega t(\mu) f_\mu(x) g(\mu)\, d\mu \Big/ \int_\Omega f_\mu(x) g(\mu)\, d\mu$$

Jeffreysonian Bayes Inference

"Uninformative Priors"

- What if we don't know prior g?
- Jeffreys: $g(\mu) = \left|\mathcal{I}(\mu)\right|^{1/2}$ where $\mathcal{I}(\mu) = \text{cov}\left\{\nabla_\mu \log f_\mu(x)\right\}$

  (the Fisher information matrix)
- Can still use Bayes theorem but how accurate are the estimates?
- Frequentist variability of $E\{t(\mu)|x\}$

General Accuracy Formula

- μ and $x \in \mathcal{R}^p$
- $V_\mu = \mathrm{cov}_\mu(x)$
- $\alpha_x(\mu) = \nabla_x \log f_\mu(x) = \left(\ldots, \frac{\partial \log f_\mu(x)}{\partial x_i}, \ldots\right)^T$

Lemma

$E = E\{t(\mu)|x\}$ has gradient $\nabla_x E = \mathrm{cov}\{t(\mu), \alpha_x(\mu)|x\}$.

Theorem

The delta-method standard deviation of E is

$$\mathrm{sd}(E) = \left[\mathrm{cov}\{t(\mu), \alpha_x(\mu)|x\}^T \, V_x \, \mathrm{cov}\{t(\mu), \alpha_x(\mu)|x\}\right]^{1/2}.$$

Implementation

- Posterior sample from $\mu|x$

  $\{\mu_1, \mu_2, \ldots, \mu_B\}$ (MCMC?)

- Each μ_i gives $t_i = t(\mu_i)$ and

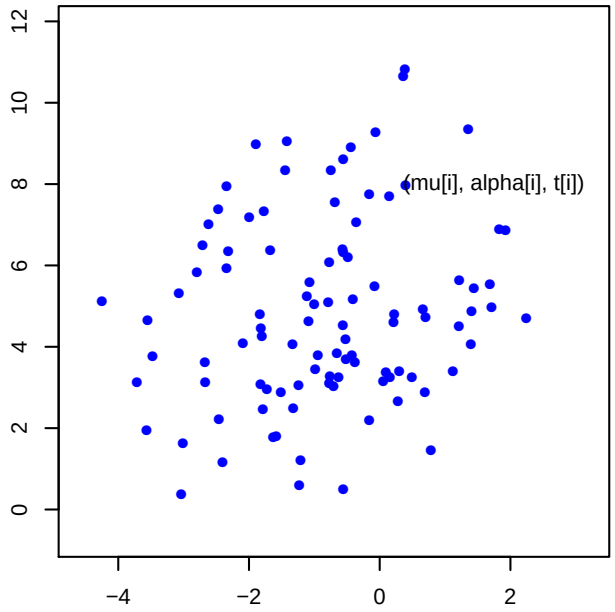
  $\alpha_i = \alpha_x(\mu_i)$

- $\hat{E} = \sum t_i/B \doteq E\{t(\mu)|x\}$

- $\widehat{\text{cov}} = \sum_{i=1}^{B} (\alpha_i - \hat{\alpha})\left(t_i - \bar{t}\right)\Big/B$

- $\boxed{\widehat{\text{sd}} = \left[\widehat{\text{cov}}^T V_x \widehat{\text{cov}}\right]^{1/2}}$

- No additional sampling for $\widehat{\text{cov}}$



Diabetes Data

Efron et al. (2004), "LARS"

- $n = 442$ subjects
- $p = 10$ predictors: age, sex, bmi, glu,...
- Response: $y =$ disease progression at one year
- Model: $\underset{n\times 1}{\boldsymbol{y}} = \underset{n\times p}{\boldsymbol{X}}\ \underset{p\times 1}{\beta} + \underset{n\times 1}{\boldsymbol{e}} \qquad [\boldsymbol{e} \sim \mathcal{N}_n(\boldsymbol{0}, I)]$

Bayesian Lasso

Park and Casella (2008)

- Model: $\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}\beta, I)$
- Prior: $g(\beta) = e^{-\gamma L_1(\beta)}$ $\qquad [\gamma = 0.37]$
- Then posterior mode at Lasso $\hat{\beta}_\gamma$
- Subject 125: $\theta_{125} = \boldsymbol{x}_{125}^T \beta$
- How accurate are Bayes posterior inferences for θ_{125}?

Bayesian Analysis

- MCMC: posterior sample $\{\beta_i \text{ for } i = 1, 2, \ldots, 10,000\}$
- Gives $\left\{\theta_{125,i} = \boldsymbol{x}_{125}^T \beta_i,\ i = 1, 2, \ldots, 10,000\right\}$

$$\boxed{\theta_{125,i} \doteq 0.248 \pm 0.072}$$

- General accuracy formula frequentist sd 0.071 for $E = 0.248$

$$\left[\alpha_x(\mu) = X^T X \beta\right]$$

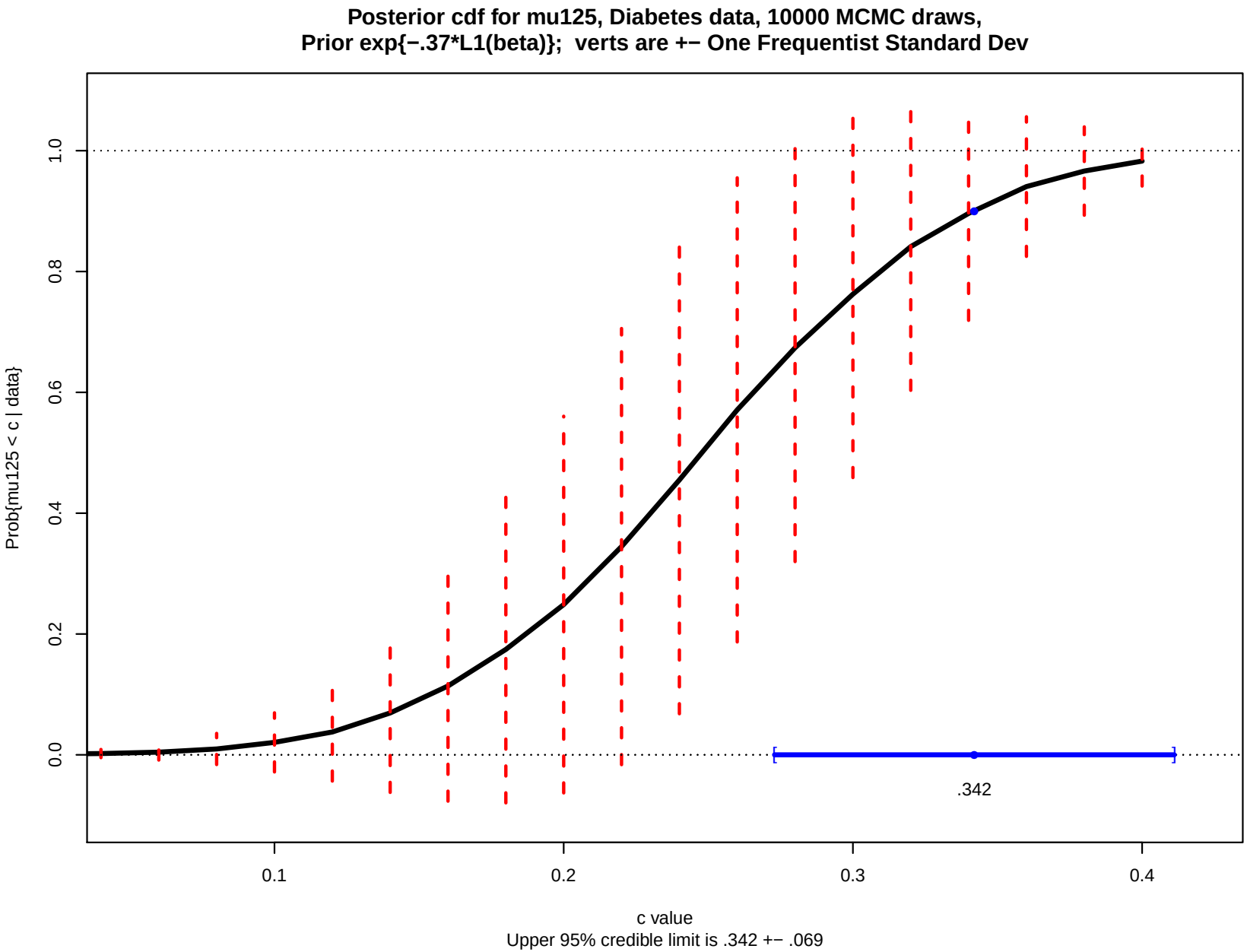
Posterior CDF for Subject 125

- $\text{cdf}_{\boldsymbol{y}}(c) = \Pr\{\theta_{125} \le c|\boldsymbol{y}\}$
- $s_i = \begin{cases} 1 & \text{as } t_i \le c \\ 0 & \text{as } t_i > c \end{cases}$
- $\widehat{\text{cdf}}_{\boldsymbol{y}}(c) = \sum_1^B s_i/B$
- For $c = 0.3$:

$$\widehat{\text{cdf}}_{\boldsymbol{y}}(c) = 0.762 \pm 0.304$$

↗ Bayes estimate

↖ frequentist sd


Posterior cdf for mu125, Diabetes data, 10000 MCMC draws,
Prior exp{−.37*L1(beta)}; verts are +− One Frequentist Standard Dev
Prob{mu125 < c | data}
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.2
0.3
0.4
.342
c value
Upper 95% credible limit is .342 +− .069

Exponential Families

- $f_\alpha\left(\hat{\beta}\right) = e^{\alpha^T\hat{\beta}-\psi(\alpha)} f_0\left(\hat{\beta}\right)$, with $\alpha,\ \hat{\beta}$ in $\mathcal{R}^p$
- Natural parameter α, sufficient statistic $\hat{\beta}$, expectation $\beta = E_\alpha\left\{\hat{\beta}\right\}$

  $\left[\text{Poisson}: \quad f_\mu(x) = e^{-\mu}\mu^x/x!:\ x = \hat{\beta},\ \mu = \beta,\ \alpha = \log(\mu)\right]$
- General accuracy formula For $E = E\left\{t(\beta)|\hat{\beta}\right\}$,

$$\boxed{\widehat{\text{sd}}(E) = \left\{\text{cov}\left(t, \alpha\middle|\hat{\beta}\right)^T V_{\hat{\alpha}}\, \text{cov}\left(t, \alpha\middle|\hat{\beta}\right)\right\}^{1/2}}$$

with $V_{\hat{\alpha}} = \text{cov}_{\alpha=\hat{\alpha}}\left(\hat{\beta}\right)$.

Better Frequentist Inferences

- For $E = E\left\{t(\beta)\middle|\hat{\beta}\right\}$ in exfam $f_\alpha\left(\hat{\beta}\right) = e^{\alpha^T\hat{\beta}-\psi(\alpha)}f_0\left(\hat{\beta}\right)$
- Parametric bootstrap $f_{\hat{\alpha}}(\cdot) \to \left[\hat{\beta}^*_1, \hat{\beta}^*_2, \ldots, \hat{\beta}^*_j, \ldots, \hat{\beta}^*_J\right]$
  $\longrightarrow \left[\cdots E^*_j = E\left\{t(\beta)\middle|\hat{\beta}^*_j\right\}\cdots\right] \longrightarrow$ bootstrap conf int for E
- **Trouble** Need new MCMC sample for each $\hat{\beta}^*_j$
- Shortcut Reweight original MCMC sample (importance sampling)

Digression: *Posterior Exponential Family*

- $f_\alpha\left(\hat{\beta}\right) = e^{\alpha^T\hat{\beta}-\psi(\alpha)} f_0\left(\hat{\beta}\right)$: natural param α, suff stat $\hat{\beta}$, "carrier" $f_0\left(\hat{\beta}\right)$
- Posterior exponential family

$$g(\alpha|\text{suff stat } b) = e^{\left(b-\hat{\beta}\right)^T\alpha-\phi(b)} g\left(\alpha\middle|\hat{\beta}\right)$$

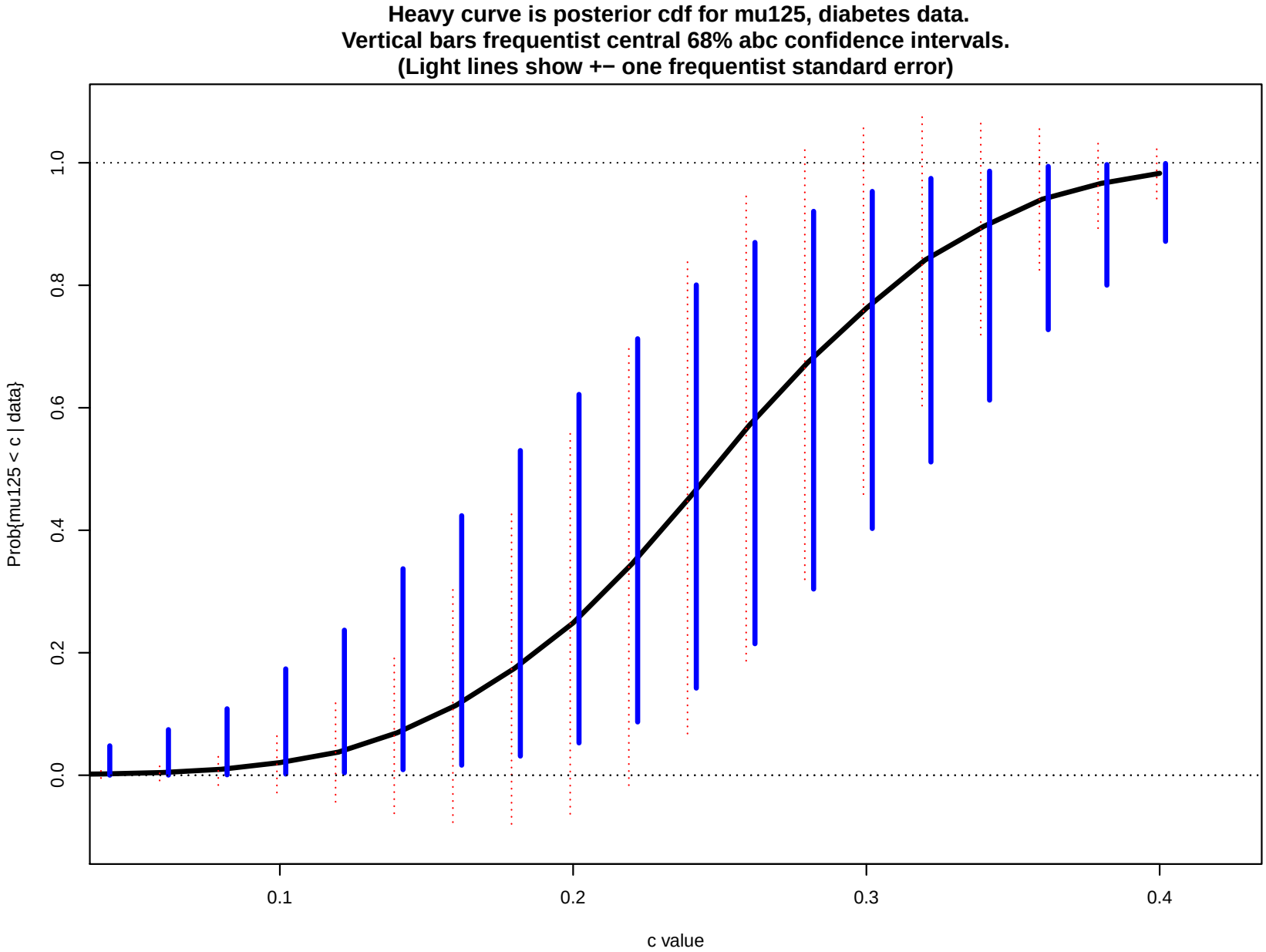
- natural param b, suff stat α, carrier $g\left(\alpha\middle|\hat{\beta}\right)$
- Importance sampling Reweight the original MCMC realizations:

$$\hat{E}\{t_i|b\} = \sum_{i=1}^{B} t_i W_i(b) \Big/ \sum_{i=1}^{B} W_i(b) \qquad \left[W_i(b) = e^{\left(b-\hat{\beta}\right)^T\alpha_i}\right]$$

Bootstrap Intervals Without Bootstrapping

DiCiccio and Efron (1992)

- **"abc"** Investigate $\hat{E}\{t|b\}$ for b *near* $\hat{\beta}$
- Requires $p + 2$ numerical 2nd derivatives of $\hat{E}$ function
- *Next:* Applied to posterior cdf for mu125


Heavy curve is posterior cdf for mu125, diabetes data.
Vertical bars frequentist central 68% abc confidence intervals.
(Light lines show +− one frequentist standard error)
Prob{mu125 < c | data}
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.2
0.3
0.4
c value

Estimation After Model Selection

- **Usually:**

  (a) look at data

  (b) choose model (linear, quad, cubic . . . ?)

  (c) fit estimates using chosen model

  (d) analyze as if pre-chosen

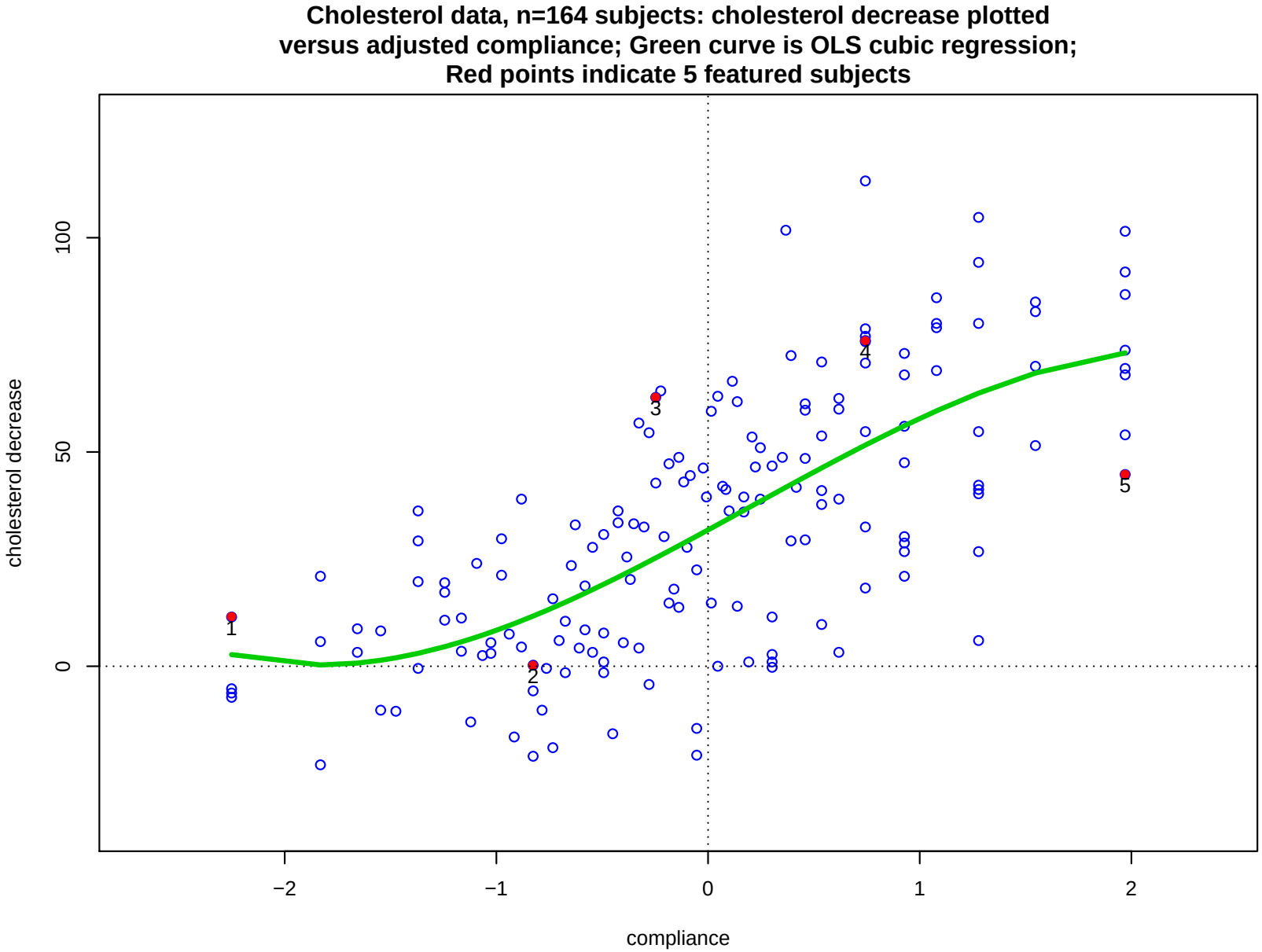
- *Today* Include model selection process in the analysis
- Question Effects on standard errors, confidence intervals, etc.?

Cholesterol Data

- $n = 164$ men took Cholestyramine for ~ 7 years
- $x =$ *compliance measure* (adjusted: $x \sim \mathcal{N}(0,1)$)
- $y =$ *cholesterol decrease*
- Wish to estimate regression values

$$\mu_j = E\{y|x = x_j\} \qquad \text{for } j = 1,2,\ldots,164$$

- $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_{164})^T$


Cholesterol data, n=164 subjects: cholesterol decrease plotted
versus adjusted compliance; Green curve is OLS cubic regression;
Red points indicate 5 featured subjects
cholesterol decrease
100
50
0
−2
−1
0
1
2
compliance
1
2
3
4
5

C_p Selection Criterion

- Regression model $\underset{n\times 1}{\boldsymbol{y}} = \underset{n\times m}{X}\ \underset{m\times 1}{\beta} + \underset{n\times 1}{\boldsymbol{e}} \quad \left[e_i \sim (0, \sigma^2)\right]$
- C_p criterion $\boxed{\left\|\boldsymbol{y} - X\hat{\beta}\right\|^2 + 2m\sigma^2}$

  $\hat{\beta}$ = OLS estimate, m = "degrees of freedom"
- Model selection From possible models $X_1, X_2, X_3, \ldots$

  choose the one minimizing C_p.
- Then use OLS estimate from chosen model.

C_p *for Cholesterol Data*

Model	df	C_p – 80000	(Boot %)
$\mathcal{M}_1$ (linear)	2	1132	(19%)
$\mathcal{M}_2$ (quad)	3	1412	(12%)
$\mathcal{M}_3$ (cubic)	4	667	(34%)
$\mathcal{M}_4$ (quartic)	5	1591	(8%)
$\mathcal{M}_5$ (quintic)	6	1811	(21%)
$\mathcal{M}_6$ (sextic)	7	2758	(6%)

($\sigma = 22$ from "full model" $\mathcal{M}_6$)

Nonparametric Bootstrap Analysis

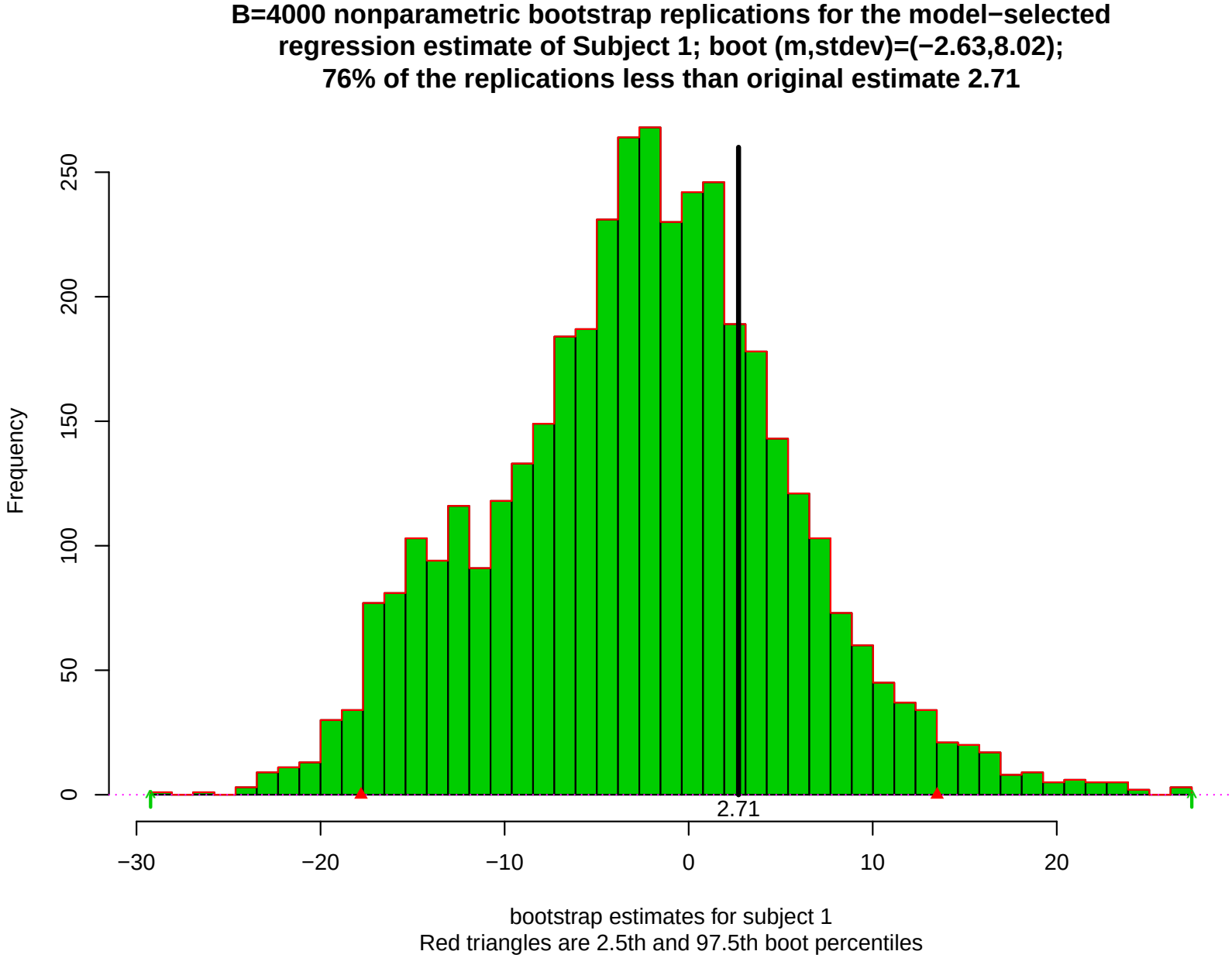
- data $= \{(x_i, y_i),\ i = 1, 2, \ldots, n = 164\}$ gave original estimate

$$\hat{\mu} = X_3 \hat{\beta}_3$$

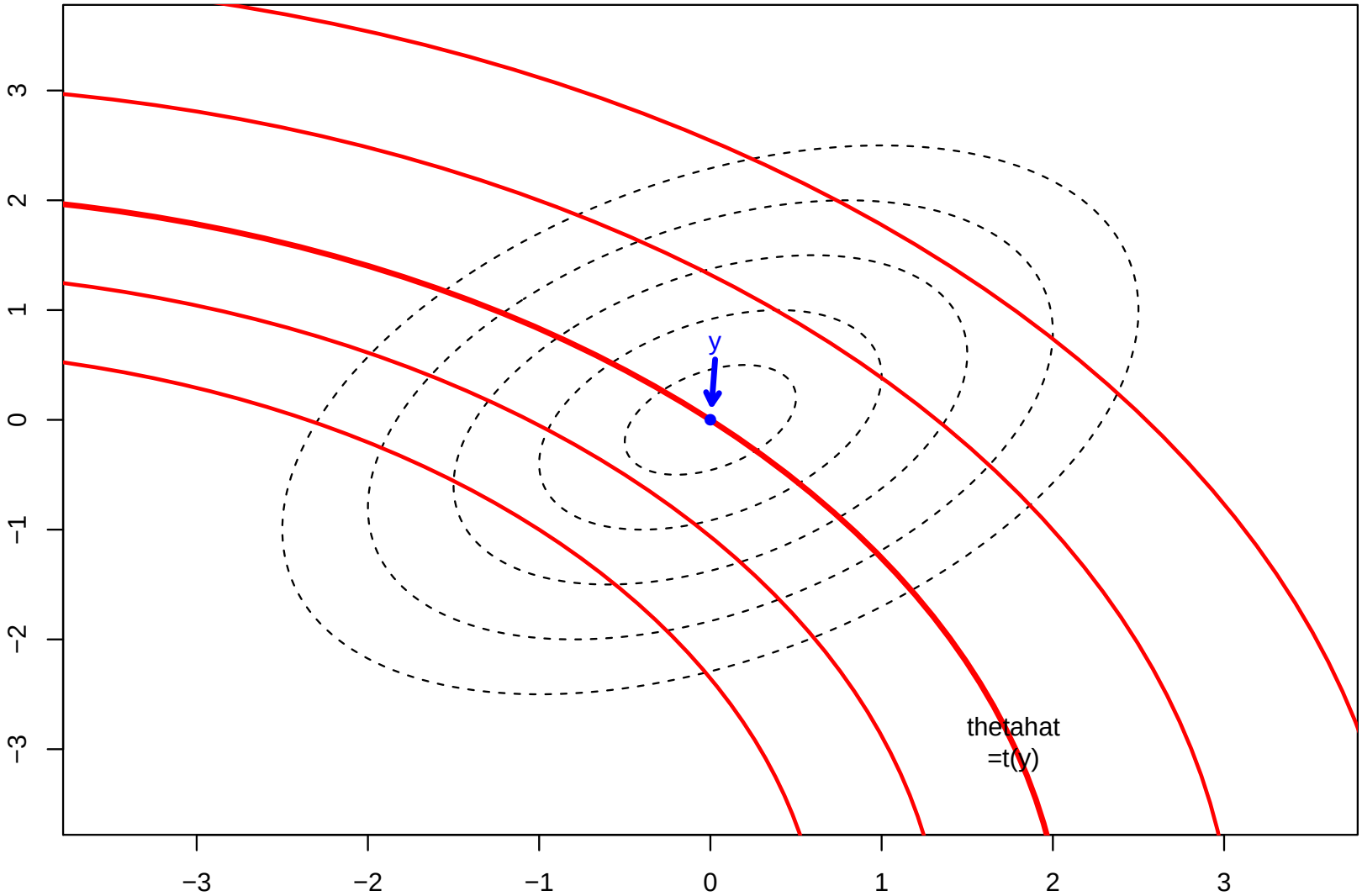
- Bootstrap data set: $\text{data}^* = \left\{(x_j, y_j)^*,\ j = 1, 2, \ldots, n\right\}$ where $(x_j, y_j)^*$ drawn randomly and with replacement from data:

$$\text{data}^* \underset{C_p}{\longrightarrow} m^* \underset{\text{OLS}}{\longrightarrow} \hat{\beta}^*_{m^*} \longrightarrow \hat{\mu}^* = X_{m^*} \hat{\beta}^*_{m^*}$$

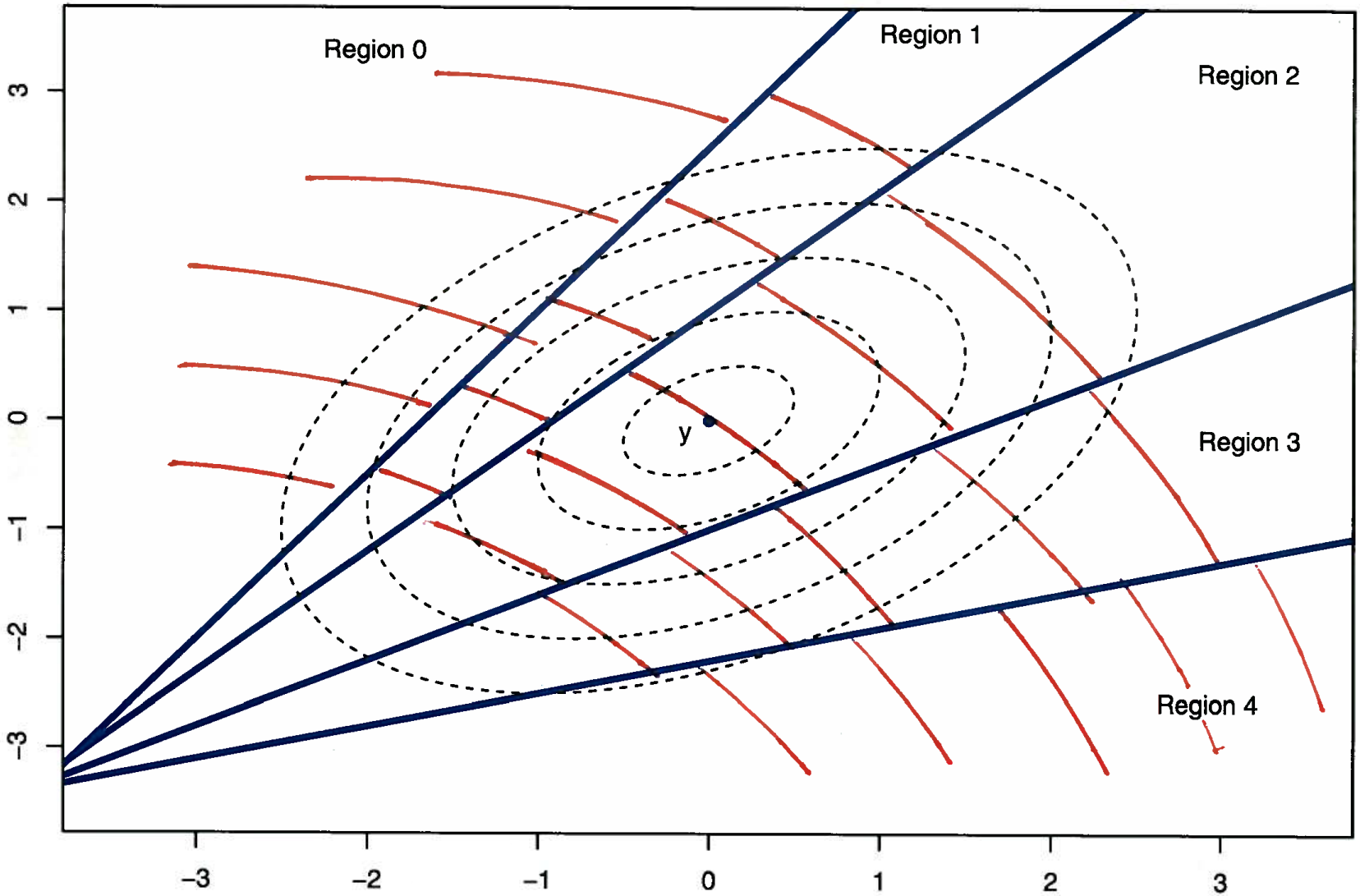
- I did this all $B = 4000$ times.

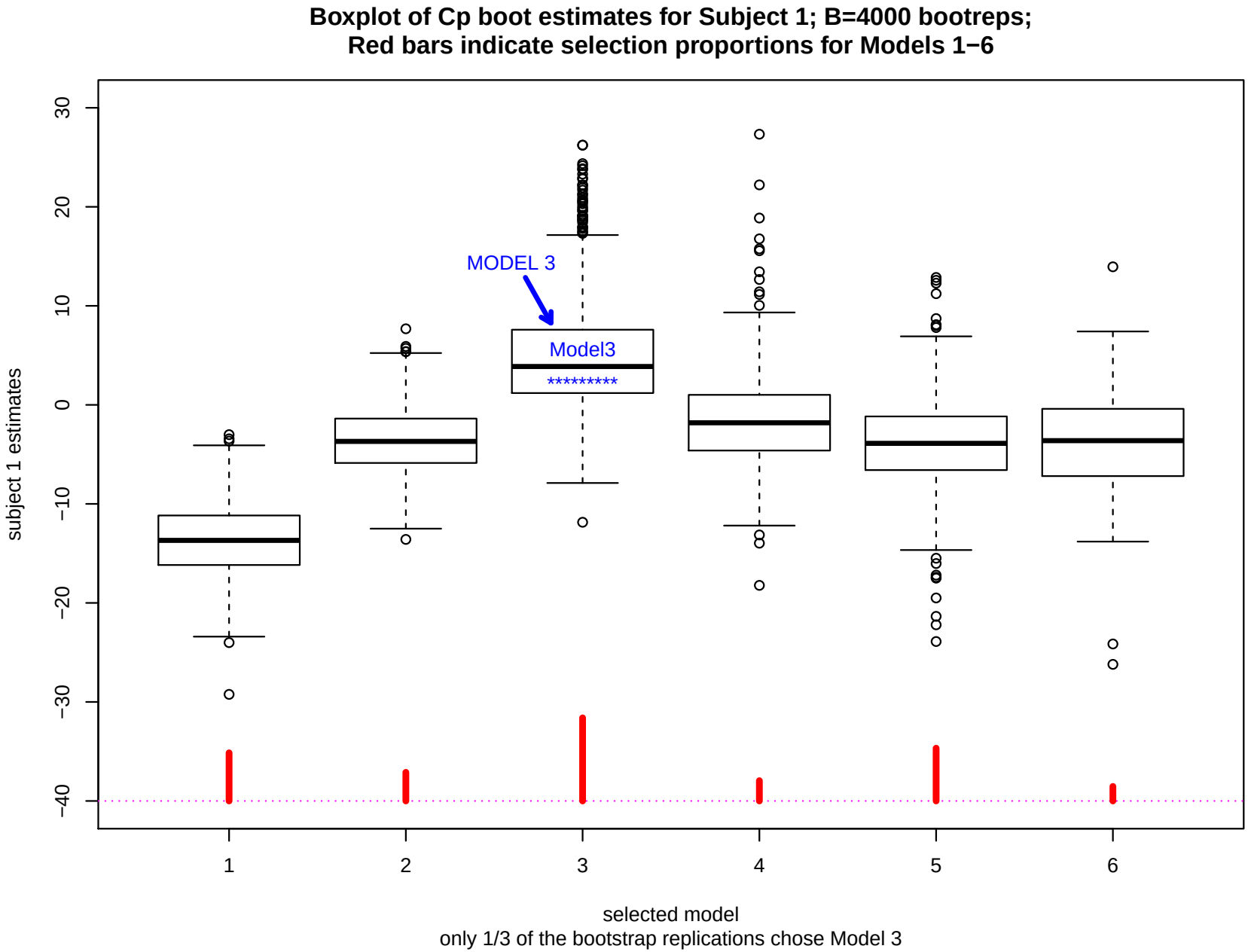

B=4000 nonparametric bootstrap replications for the model−selected
regression estimate of Subject 1; boot (m,stdev)=(−2.63,8.02);
76% of the replications less than original estimate 2.71
Frequency
0
50
100
150
200
250
2.71
−30
−20
−10
0
10
20
bootstrap estimates for subject 1
Red triangles are 2.5th and 97.5th boot percentiles

Smooth Estimation Model: ' y ' is observed data;
Ellipses indicate bootstrap distribution for ' y* ';
Red curves level surfaces of equal estimation for thetahat=t(y)



Estimation with model selection: now the curves of equal estimation are discontinuous across the Model region boundaries




Boxplot of Cp boot estimates for Subject 1; B=4000 bootreps;
Red bars indicate selection proportions for Models 1−6
subject 1 estimates
30
20
10
0
−10
−20
−30
−40
MODEL 3
Model3
****
1
2
3
4
5
6
selected model
only 1/3 of the bootstrap replications chose Model 3

Bootstrap Smoothing

- *Idea* Replace original estimator $t(\boldsymbol{y})$ with bootstrap average

$$\boxed{s(\boldsymbol{y}) = \sum_{i=1}^{B} t\left(\boldsymbol{y}_i^*\right)/B}$$

- Model averaging
- Same as *bagging* ("bootstrap aggregation," Breiman)
- Removes discontinuities, reduces variance
- Approximate confidence interval: $s(\boldsymbol{y}) \pm 1.96 \cdot \widetilde{\text{sd}}$

Accuracy Theorem

- Notation $s_0 = s(\boldsymbol{y}), \quad t_i^* = t(\boldsymbol{y}_i^*),\ i = 1, 2, \ldots B$
- $Y_{ij}^* = \#$ of times jth data point appears in ith boot sample
- $\mathrm{cov}_j = \sum_{i=1}^{B} Y_{ij}^* \cdot \left(t_i^* - s_0\right) \big/ B \quad \left[\text{covariance } Y_{ij}^* \text{ with } t_i^*\right]$

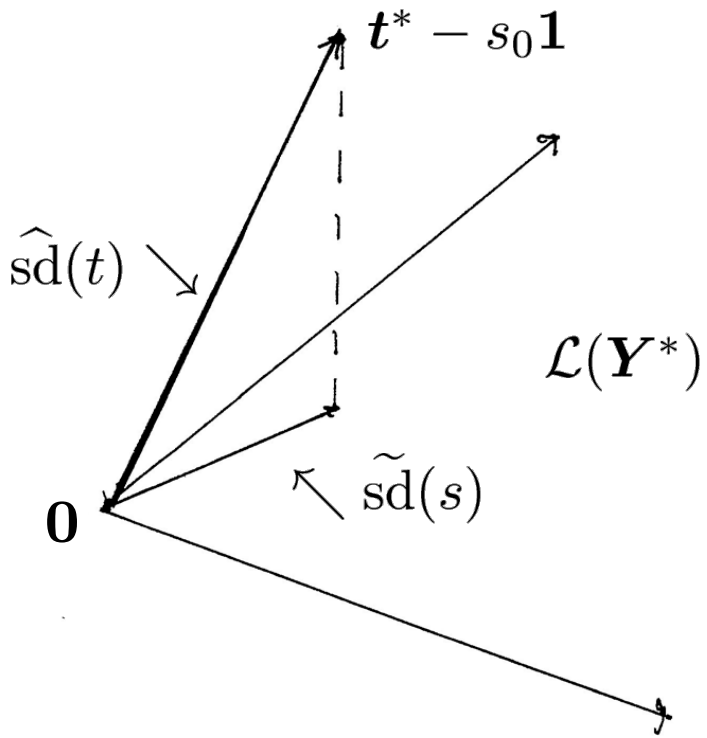
Theorem

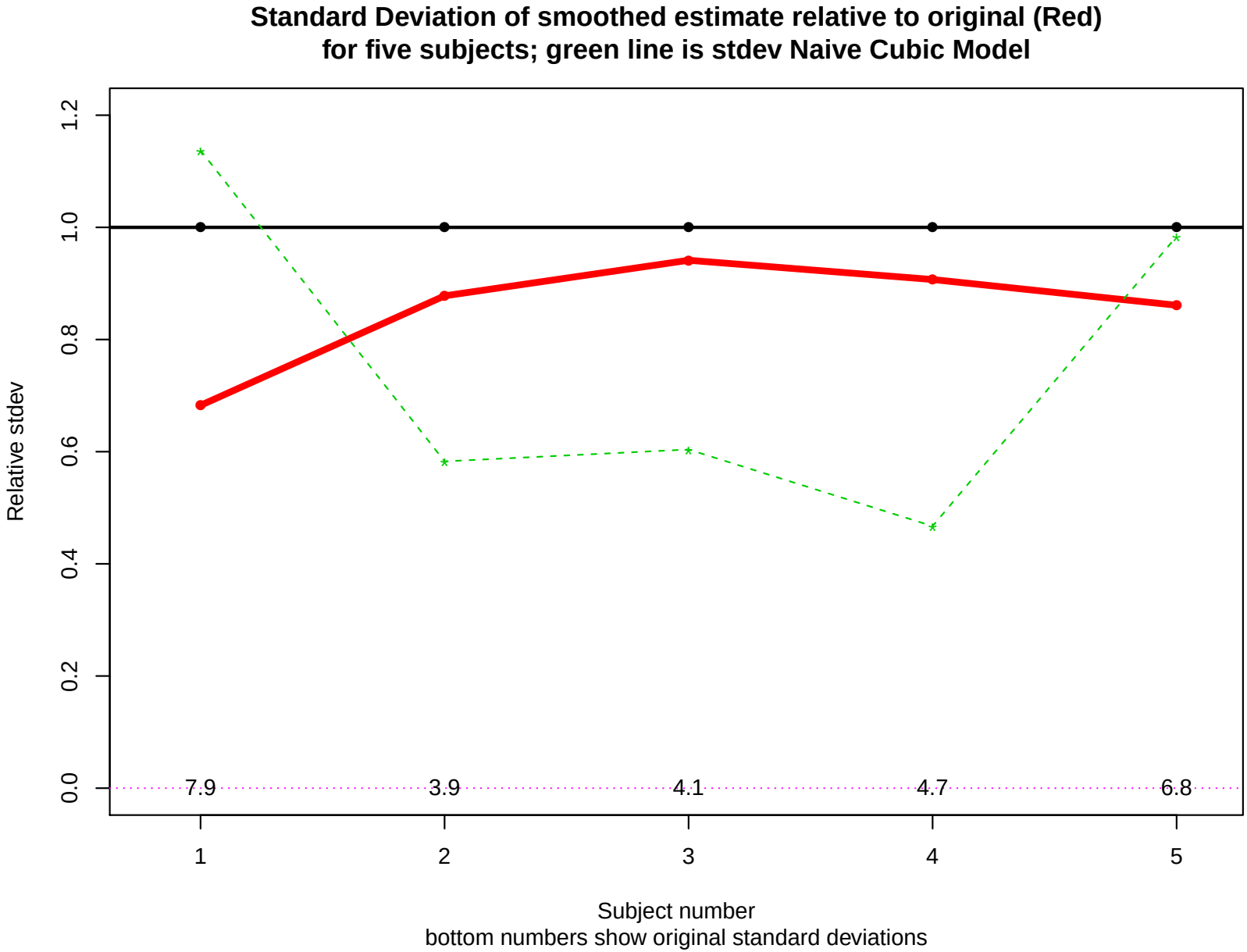
The delta method standard deviation estimate for s_0 is

$$\widetilde{\mathrm{sd}} = \left[\sum_{j=1}^{n} \mathrm{cov}_j^2\right]^{1/2},$$

always $\leq \left[\sum_{i=1}^{B} \left(t_i^* - s_0\right)^2 \big/ B\right]^{1/2}$, *the boot stdev for* $t(\boldsymbol{y})$.

Projection Interpretation




Standard Deviation of smoothed estimate relative to original (Red)
for five subjects; green line is stdev Naive Cubic Model
Relative stdev
0.0
0.2
0.4
0.6
0.8
1.0
1.2
7.9
3.9
4.1
4.7
6.8
1
2
3
4
5
Subject number
bottom numbers show original standard deviations

Model Probability Estimates

- 34% of the 4000 bootreps chose the cubic model
- Poor man's Bayes posterior prob for "cubic"
- How accurate is that 34%?
- Apply accuracy theorem to indicator function for choosing "cubic"

Model	Boot %	± Standard Error
$\mathcal{M}_1$ (linear)	19%	±24
$\mathcal{M}_2$ (quad)	12%	±18
$\mathcal{M}_3$ (cubic)	34%	±24
$\mathcal{M}_4$ (quartic)	8%	±14
$\mathcal{M}_5$ (quintic)	21%	±27
$\mathcal{M}_6$ (sextic)	6%	±6