Three 2nd Thoughts on Empirical Bayes Inference

— Conversations with Carl Morris —

Bradley Efron

Stanford University

Empirical Bayes (Robbins, 1950s)

- Unknown prior density $g(\theta)$ gives $\theta_1, \theta_2, \ldots, \theta_N$ (unobserved)
- Observations $x_1, x_2, \ldots, x_N$ with $x_i \overset{\text{ind}}{\sim} p(x_i|\theta_i)$ (e.g., $x_i \sim \text{Poi}(\theta_i)$)
- The Goal — To estimate the parameters θ_i
- Amazing Fact: For large N we can nearly achieve Bayes risk
- "Large parallel studies contain their own Bayesian information."

"Normal Normal Case":

$\theta_i \sim \mathcal{N}(M, A)$ and $x_i \sim \mathcal{N}(\theta_i, 1)$

- Bayes Rule $\hat{\theta}_i^{\text{Bayes}} = E\{\theta_i | x_i\} = M + (1 - B)(x_i - M)$
- $B = 1/(A + 1)$ [shrinkage factor $A/(A + 1)$]
- Bayes Risk $\boxed{R^{\text{Bayes}} = E\left\{\sum_1^n \left(\theta_i - \hat{\theta}_i^{\text{Bayes}}\right)^2\right\} = N(1 - B)}$
- MLE $\hat{\theta}_i^{\text{MLE}} = x_i : R^{\text{MLE}} = N$
- $R^{\text{Bayes}}/R^{\text{MLE}} = 1 - B$
- Bayes saves proportion B of the risk

The James–Stein Estimator (1961)

- James–Stein $\hat{\theta}_i^{JS} = \hat{M} + \left(1 - \hat{B}\right)\left(x_i - \hat{M}\right)$

  where $\hat{M} = \bar{x}$ and $\hat{B} = (N-3)\Big/\sum_1^N \left(x_i - \hat{M}\right)^2$ (unbiased ests)

- Risk: $R^{JS}\Big/R^{Bayes} = 1 + 3\Big/(N \cdot A)$

- For $N = 18$, $A = 1$, JS loses 1/6 of Bayes savings

- "Shrinkage estimation"

Theorem

$$E_{\boldsymbol{\theta}}\left\{\sum\left(\theta_i - \hat{\theta}_i^{JS}\right)^2\right\} < E_{\boldsymbol{\theta}}\left\{\sum\left(\theta_i - \hat{\theta}_i^{MLE}\right)^2\right\} \textit{ always!}$$

The 18 baseball players

	MLE	Truth	JS
Clemente	.400	.346	.294
F. Robinson	.378	.298	.289
F. Howard	.356	.276	.285
Johnstone	.333	.222	.280
⋮	⋮	⋮	⋮
E. Rodriguez	.222	.226	.256
Campaneris	.200	.285	.252
Munson	.178	.316	.247
Alvis	.156	.200	.242
Squared error:	**.075**		**.021**

Learning From the Experience of Others

- Why does Clemente's good performance increase Munson's estimate?
- Parallel data sets let you "learn from the experience of others"
- Which others?

Relevance (Efron–Morris 1972)

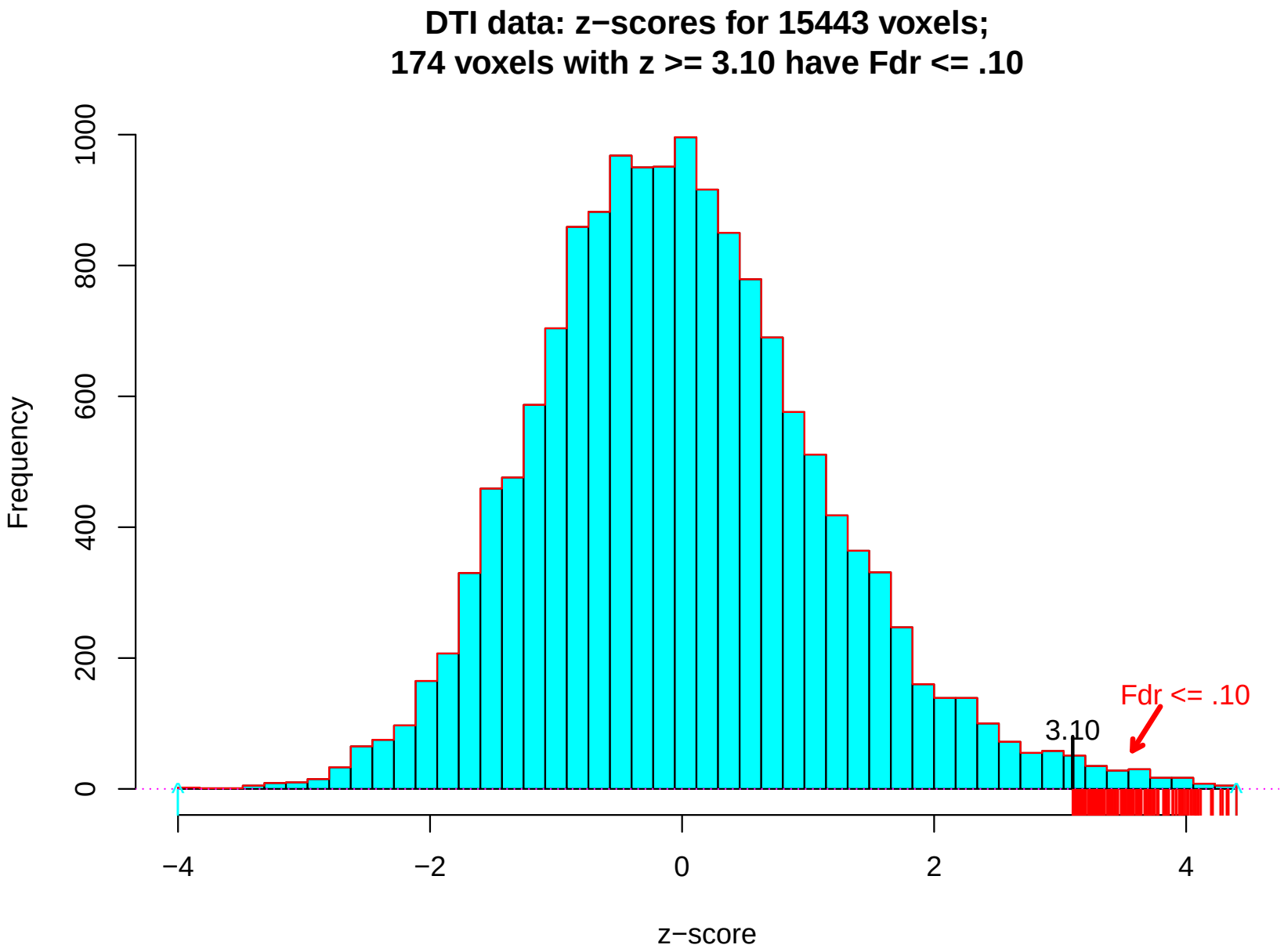
- $\hat{\theta}_i^{\text{rel}} = \hat{M} + \left[1 - \rho(x_i)\hat{B}\right]\left(x_i - \hat{M}\right)$
- Relevance function $\rho(\cdot)$ decreases with $\left|x_i - \hat{M}\right|$
- Limited Translation "Never shrink more than one unit away from MLE" $\longrightarrow$ if $x_i \sim \mathcal{N}(\theta_i, \sigma^2)$ then "unit" $= \sigma$
- Clemente: $\hat{\theta}_i^{\text{rel}} = 0.334$ (cf. $\hat{\theta}_i^{\text{JS}} = 0.294$)
- Loses about 10% of $\hat{\theta}^{\text{JS}}$ savings

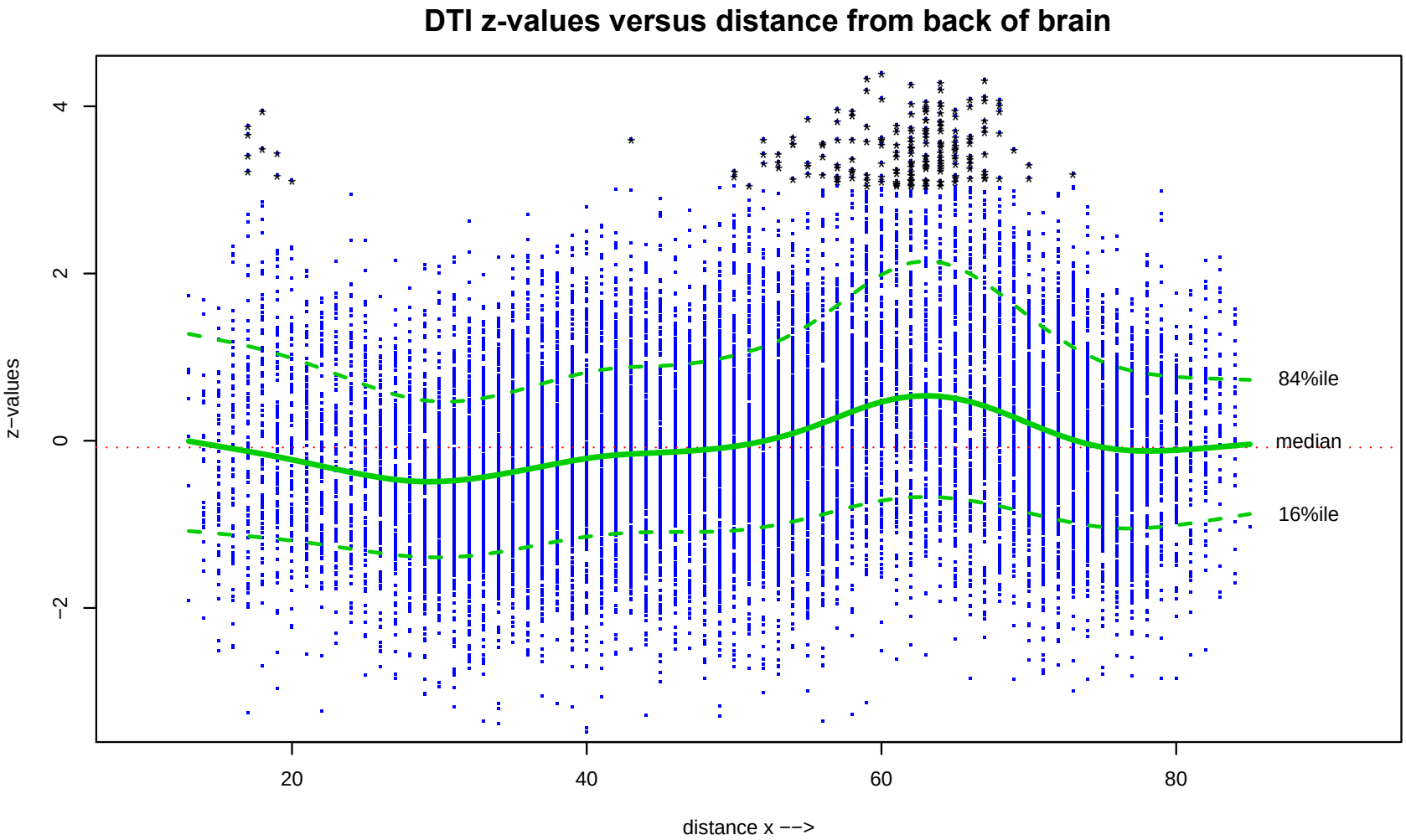
False Discovery Rates (Benjamini–Hochberg 1995)

- EB Testing Which of N cases are "significant"?
- DTI Study: 6 dyslexics vs 6 controls, $N = 15,443$ voxels

$$\boxed{z_i \overset{\text{null}}{\sim} \mathcal{N}(0,1)} \qquad \text{for } i\text{th voxel}$$

- Fdr All z's determine each z_i's significance
- "locfdr": 174 voxels with $z_i \geq 3.10$ have Fdr ≤ 0.10


DTI data: z−scores for 15443 voxels;
174 voxels with z >= 3.10 have Fdr <= .10
Frequency
0
200
400
600
800
1000
−4
−2
0
2
4
z−score
3.10
Fdr <= .10


DTI z-values versus distance from back of brain
z−values
4
2
0
−2
20
40
60
80
distance x −−>
84%ile
median
16%ile

Relevance (Efron 2010, §10.3)

- x_i = distance of voxel i from back of brain
- Target voxel i_0
- Voxel$_i$ counts amount $\rho(x_i - x_{i_0})$ toward Fdr$_i$
- Example $\rho = 1$ if $|x_i - x_{i_0}| \le 10$, $\rho = 0$ otherwise

  or $\rho = \exp(|x_i - x_{i_0}|/10)$
- Kicking the can down the street ...

Second 2nd Thought: Ridge Regression (Hoerl–Kinnard 1970)

- OLS $y = \underset{n\times p}{X}\,\beta + \epsilon$ with $\epsilon_i \overset{\text{ind}}{\sim} (0,1)$
- $\hat{\beta} = S^{-1}X'y$ $[S = (X'X)]$
- Ridge Estimate

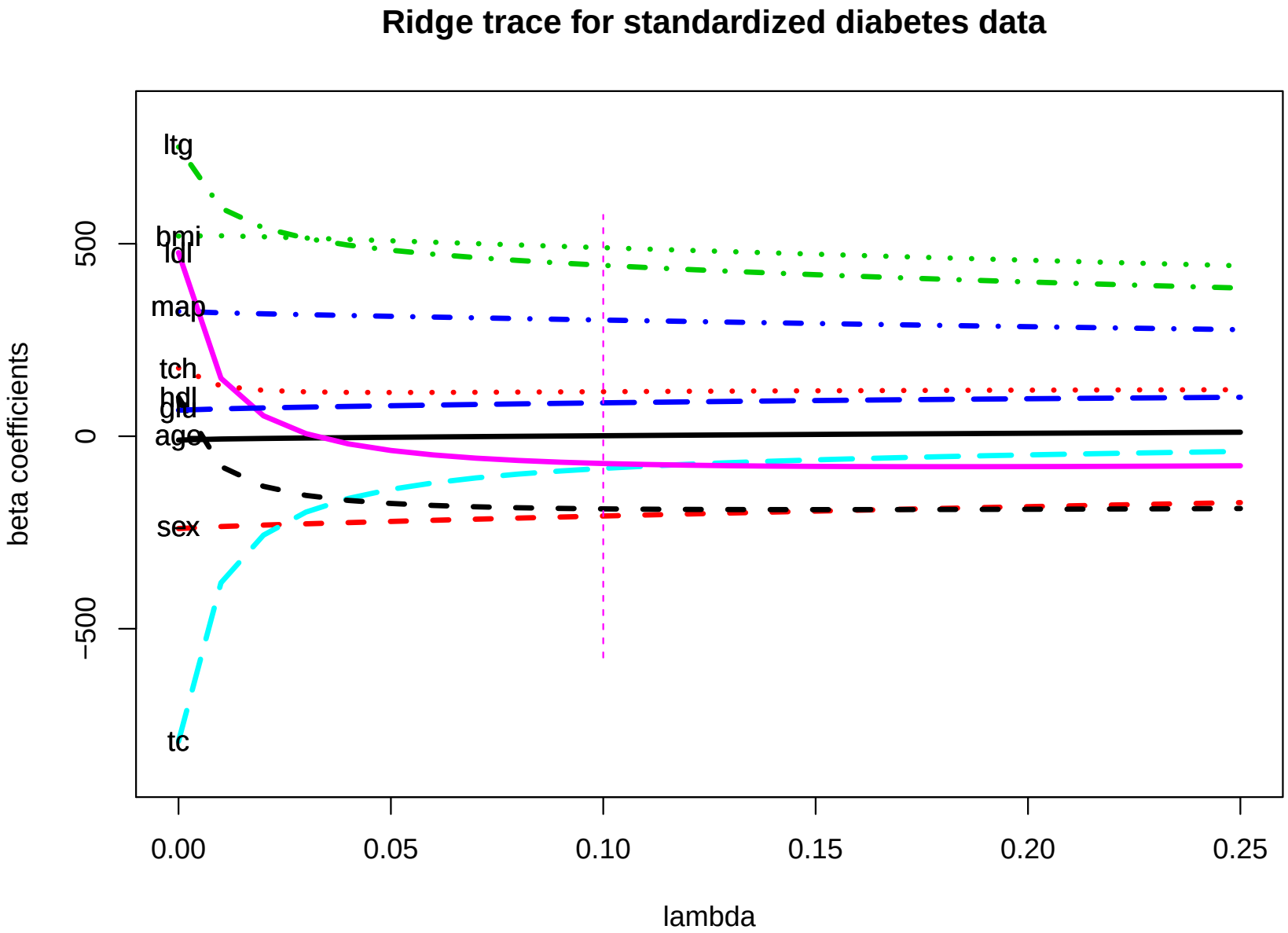
$$\boxed{\hat{\beta}^{(\lambda)} = (S + \lambda I)^{-1} S\,\hat{\beta}}$$

- Shrinks estimate $\hat{\beta}$ toward zero
- Empirical Bayes data-based choice of λ
- Example Diabetes Study: $n = 442$, $p = 10$ predictors

Diabetes study: First 7 of $n = 442$ patients

(predict prog from the 10 covariates)

age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu	**prog**
59	1	32.1	101	157	93.2	38	4	2.11	87	**151**
48	0	21.6	87	183	103.2	70	3	1.69	69	**75**
72	1	30.5	93	156	93.6	41	4	2.03	85	**141**
24	0	25.3	84	198	131.4	40	5	2.12	89	**206**
50	0	23.0	101	192	125.4	52	4	1.86	80	**135**
23	0	22.6	89	139	64.8	61	2	1.82	68	**97**
36	1	22.0	90	160	99.6	50	3	1.72	82	**138**


Ridge trace for standardized diabetes data
beta coefficients
500
0
−500
ltg
bmi
ldl
map
tch
hdl
glu
age
sex
tc
0.00
0.05
0.10
0.15
0.20
0.25
lambda

Comparison with James–Stein

- $\hat{\beta}^{JS} = \left(1 - \dfrac{p-2}{\hat{\beta}' S \hat{\beta}}\right)\hat{\beta}$ $\qquad = 0.97\hat{\beta}$ for diabetes
- $\hat{\mu}^{JS} = X\hat{\beta}^{JS}$ guaranteed to beat $\hat{\mu} = X\hat{\beta}$
- No such result for ridge regression and no automatic choice of λ
- But... there's more than one way to shrink a cat

Estimate and stdevs for components of β

	betamle	betaridge	sdmle	sdridge
age	−10.0	1.3	59.7	52.7
sex	−239.8	−207.2	61.2	53.2
bmi	519.8	489.7	66.5	56.3
map	324.4	301.8	65.3	55.7
tc	−792.2	−83.5	416.2	43.6
ldl	476.7	−70.8	338.6	52.4
hdl	101.0	−188.7	212.3	58.4
tch	177.1	115.7	161.3	70.8
ltg	751.3	443.8	171.7	58.4
glu	67.6	86.7	65.9	56.6

Three Uses of Regression Rules

- Given any vector $\boldsymbol{v}$ of covariates, $\hat{y} = r(\boldsymbol{v})$

  e.g., $\hat{y} = \boldsymbol{v}'\hat{\beta}^{\mathrm{JS}}$ or $\boldsymbol{v}'\hat{\beta}^{(\lambda)}$

  1. Prediction [JS]

     Want small prediction error $E(y - \hat{y})^2$ (y the response at $\boldsymbol{v}$)

  2. Response Surface Estimation

     Want $\hat{y}$ accurate for $E\{y|\boldsymbol{v}\}$

  3. Explanation [Ridge]

     Form of $r(\boldsymbol{v})$ shows importance of the individual covariates

Ridge Regression and Regularization

- Alternate Form

$$\hat{\beta}^{(\lambda)} = \arg\min_{\beta} \left\{ \|y - X\beta\|^2 + \lambda \|\beta\|^2 \right\}$$

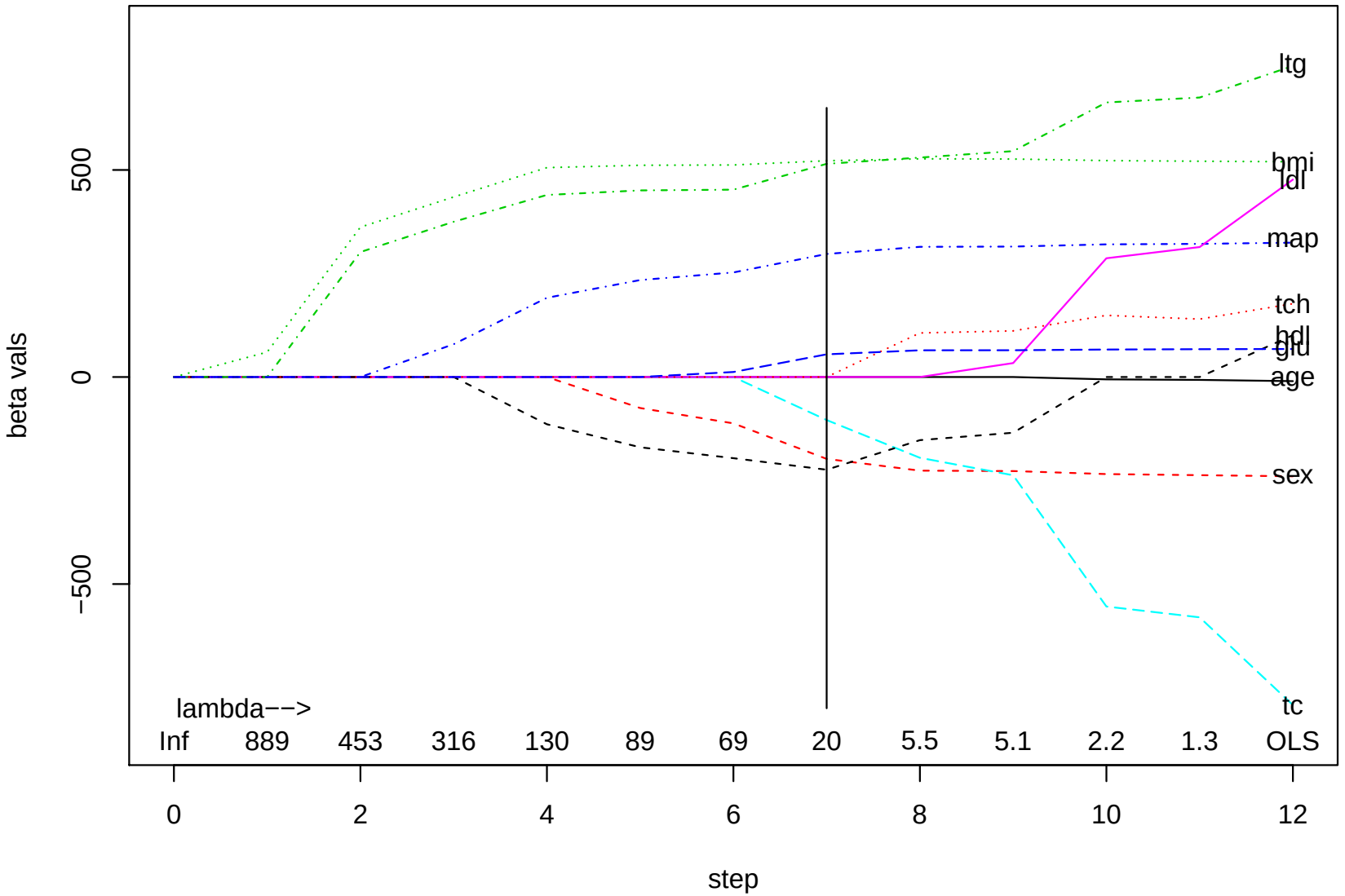
OLS penalty

- "Regularizes" OLS by penalizing large values of $\|\beta\|$
- Equivalently: Bayes vs prior $\mathcal{N}_p(0, I/\lambda)$
- Lasso

$$\tilde{\beta}^{(\lambda)} = \arg\min_{\beta} \left\{ \|y - X\beta\|^2 + \lambda \sum_{1}^{p} |\beta_j| \right\}$$

- Now sometimes shrinkage *all the way to zero* ("Sparsity")

Lasso coefficients as function of regulizer lambda;
Cp calculations say lambda=20 best



Comparison of β estimates

	mle	ridge	lasso	JS
age	−10.0	1.3	0.0	−9.7
sex	−239.8	−207.2	−197.8	−232.6
bmi	519.8	489.7	522.3	504.2
map	324.4	301.8	297.2	314.7
tc	−792.2	−83.5	−103.9	−768.4
ldl	476.7	−70.8	0.0	462.4
hdl	101.0	−188.7	−223.9	98.0
tch	177.1	115.7	0.0	171.8
ltg	751.3	443.8	514.7	728.7
glu	67.6	86.7	54.8	65.6

Robbins versus Stein

- 1956 $\theta_i \sim g(\theta)$ and $x_i \sim \text{Poi}(\theta_i)$ ind for $i = 1, \ldots, N$
- Marginal Density $f(x) = \int_0^\infty g(\theta) e^{-\theta} \theta^x / x!\, d\theta$

  for $x = 0, 1, 2, \ldots$
- Robbins' Formula

$$\boxed{E\{\theta_i | x_i = x\} = (x+1) f(x+1) \Big/ f(x)}$$

- Emp Bayes Estimate $\hat{E}\{\theta_i | x_i = x\} = (x+1) \hat{f}(x+1) \Big/ \hat{f}(x)$

$$\hat{f}(x) = \#\{x_i = x\}/N$$

Comparison (biased)

- James–Stein
  - Elegant airtight theorem
  - Works with or without prior $g(\theta)$
  - Small N
- Robbins-type Empirical Bayes
  - Asymptotic justification
  - Nonparametric (slow)
  - "Need N in the thousands!"

Auto insurance example: $N = 9461$

Claims x	0	1	2	3	4	5	6	7
$\#\{x_i < x\}$	7840	1317	239	42	14	4	4	1
$\hat{E}\{\theta_i \mid x_i = x\}$	**.168**	.363	.527	1.33	1.43	6.00	1.25	
Gamma*	.164	.398	.633	.87	1.10	1.34	1.57	

* Assumes $g(\theta)$ a Gamma density ("parametric EB")

Tweedie's Formula (1956?)

- $\theta_i \sim g(\theta)$ and $x_i \sim \mathcal{N}(\theta_i, 1)$ independent $i = 1, 2, \ldots, N$
- Marginal Density $f(x) = \int_{-\infty}^{\infty} g(\theta) \frac{e^{-(x-\theta)^2/2}}{\sqrt{2\pi}}\, d\theta$

$$\boxed{E\{\theta_i | x_i = x\} = x + \frac{d}{dx} \log f(x)}$$

- Empirical Bayes: fit smooth $\hat{f}(x)$ to histogram of $(x_1, x_2, \ldots, x_N)$;

$$\hat{E}\{\theta_i | x_i = x\} = x + \frac{d}{dx} \log \hat{f}(x)$$

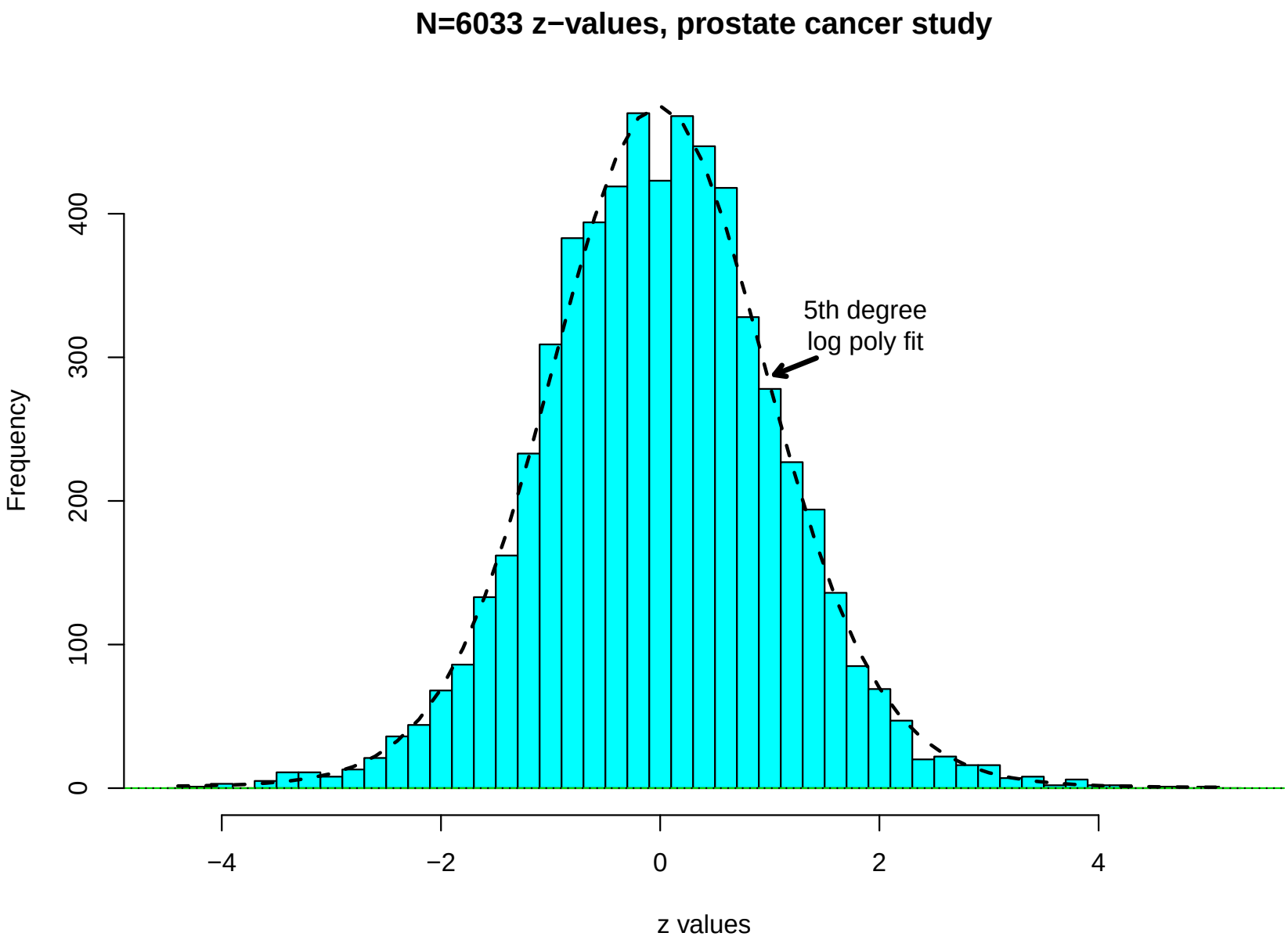
(gives James–Stein if assume $\log g(x)$ quadratic, e.g., normal)

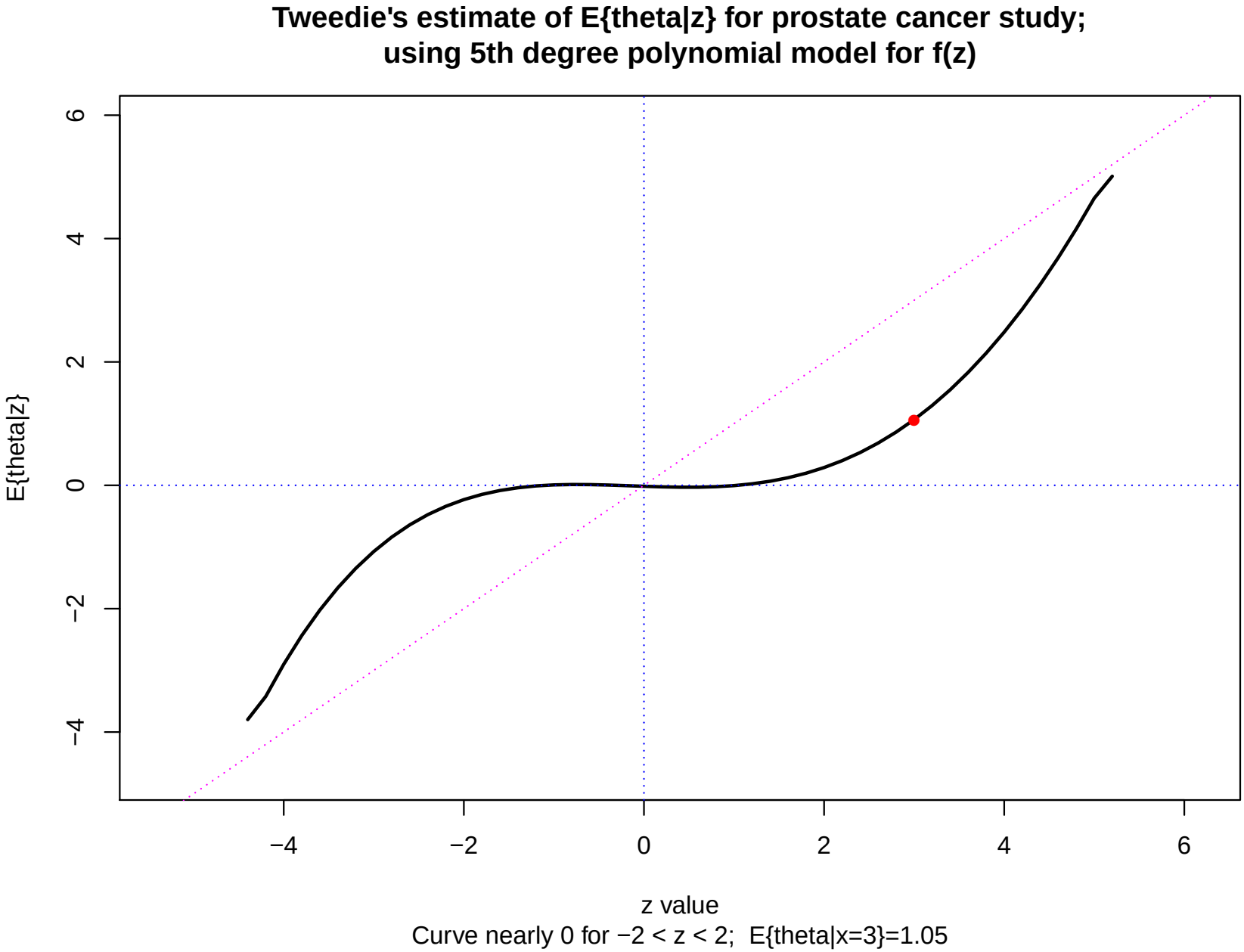
Prostate Cancer Study

- 102 men, 52 prostate cancer and 50 healthy controls
- Each assessed on $N = 6033$ genes
- z_i = two-sample test statistic, patients vs controls (z = "x")

$$\boxed{z_i \sim \mathcal{N}(\theta_i, 1)} \qquad i = 1, 2, \ldots, N$$

- θ_i is "effect size" for ith gene
- Null genes: $\theta_i = 0$


N=6033 z−values, prostate cancer study
Frequency
0
100
200
300
400
5th degree
log poly fit
−4
−2
0
2
4
z values


Tweedie's estimate of E{theta|z} for prostate cancer study;
using 5th degree polynomial model for f(z)
E{theta|z}
6
4
2
0
−2
−4
−4
−2
0
2
4
6
z value
Curve nearly 0 for −2 < z < 2; E{theta|x=3}=1.05

Empirical Bayes Hypothesis Testing (Fdr)

- Prior $g(\theta) \to \theta_i \to z_i \sim \mathcal{N}(\theta_i, 1), \qquad i = 1, 2, \ldots, N$
- $g(\theta)$ has atom π_0 at $\theta = 0$ (the null cases)
- False Discovery Rates

$$\boxed{\text{Fdr}(z_0) = \Pr\{\theta_i = 0 | z_i \geq z_0\} = \pi_0 \bar{\Phi}(z_0)/F(z_0)}$$

$$\bar{\Phi}(z_0) = \Pr\{\mathcal{N}(0,1) \geq z_0\} \quad \text{and} \quad F(z_0) = \Pr_g\{z_i \geq z_0\}$$

Fdr Control Algorithm (1995)

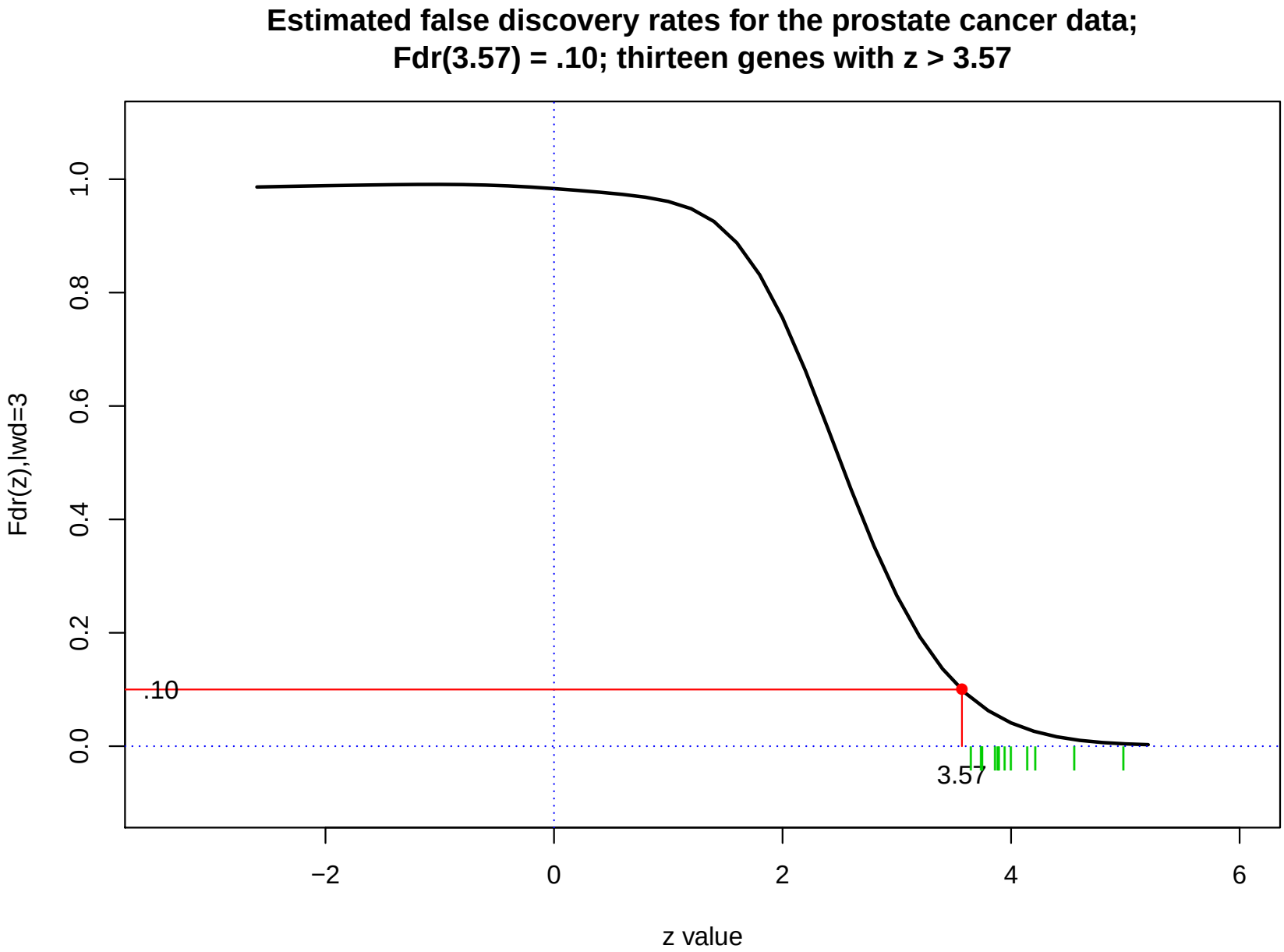
- Estimate Fdr(z) with

$$\widehat{\text{Fdr}}(z) = \bar{\Phi}(z)\Big/\hat{F}(z), \qquad \hat{F}(z) : \#\{z_i \geq z\}/N$$

- Reject $\theta_i = 0$ if $\widehat{\text{Fdr}}(z_i) \leq q$
- Frequentist Theorem

  Expected proportion false discoveries $\leq q$

- Empirical Bayes: $\widehat{\text{Fdr}}(z_i)$ decreases if *more* of the other z_j's exceed z_i


Estimated false discovery rates for the prostate cancer data;
Fdr(3.57) = .10; thirteen genes with z > 3.57
Fdr(z),lwd=3
0.0
0.2
0.4
0.6
0.8
1.0
.10
3.57
−2
0
2
4
6
z value

"There's nobody less Bayesian than an Empirical Bayesian" (Dennis Lindley, 1969)

- Early skepticism of empirical Bayes (relevance)
- Bayes prior experience influences current inference
- Emp Bayes current experience (of others) influences inference
- Crucial idea is not "estimating prior $g(\theta)$" but applying an estimated prior to individual cases

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate . . . . *JRSS-B* 57: 289–300, the original Fdr paper.

Copas, J. B. (1969). Compound decisions and empirical Bayes (with discussion). *JRSS-B* 31: 397–425, Lindley and other skeptics.

Efron, B. (2010). IMS Monographs *1*. Fdr as empirical Bayes.

Efron, B. (2011). *JASA* 106: 1602–1614, Tweedie's estimate.

Efron, B. and Morris, C. (1972). *JASA* 67: 130–139, limited translation and relevance.

Efron, B. and Morris, C. (1973). *JASA* 68: 117–130, the baseball players and other examples.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression . . . . *Technometrics* 12: 69–82.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the 4th Berkeley Symposium*, 361–379, JS estimates.

Morris, C. N. (1983). Parametric empirical Bayes inference . . . . *JASA* 78: 47–65, parametric EB and EB confidence intervals.

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the 3rd Berkeley Symposium*, 157–163.