EASY-TO-USE PROGRAMS FOR BOOTSTRAP CONFIDENCE INTERVALS

Bradley Efron and Balasubramanian Narasimhan

Stanford University

THE STANDARD INTERVALS

$$\boxed{\hat{\theta} \pm z^{(\alpha)}\hat{\sigma}} \qquad \left(z^{(0.975)} = 1.96\right)$$

- $\hat{\theta}$ a point estimate of θ
- $\hat{\sigma}$ an estimate of its standard error

  (Taylor series, jackknife, bootstrap)
- Automatic!
- Used millions of times every year

BETTER APPROXIMATE CONFIDENCE INTERVALS

- Standard intervals can be quite inaccurate
- *Two systems of better intervals* (~ 1985)
  - (A) "adjusted" likelihoods (Barndorff-Nielsen. . .)
  - (B) BCa bootstrap (← today)
- Order of magnitude improved accuracy:

  coverage $0.95 \pm c/n$ rather than $0.95 \pm c \big/ \sqrt{n}$
- Not automatic enough!
- New package `bcaboot`

CORRECTING THE STANDARD INTERVALS

- Standard intervals take literally

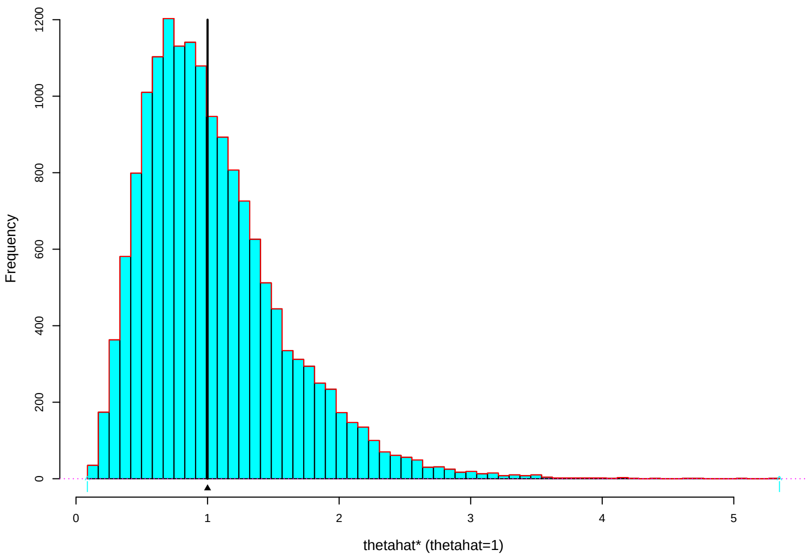
$$\hat{\theta} \sim \underset{\substack{\nearrow \\ \text{normal}}}{\mathcal{N}} \; (\underset{\substack{\uparrow \\ \text{unbiased}}}{\theta}, \quad \underset{\substack{\nwarrow \\ \text{constant} \\ \text{variance}}}{\sigma^2})$$

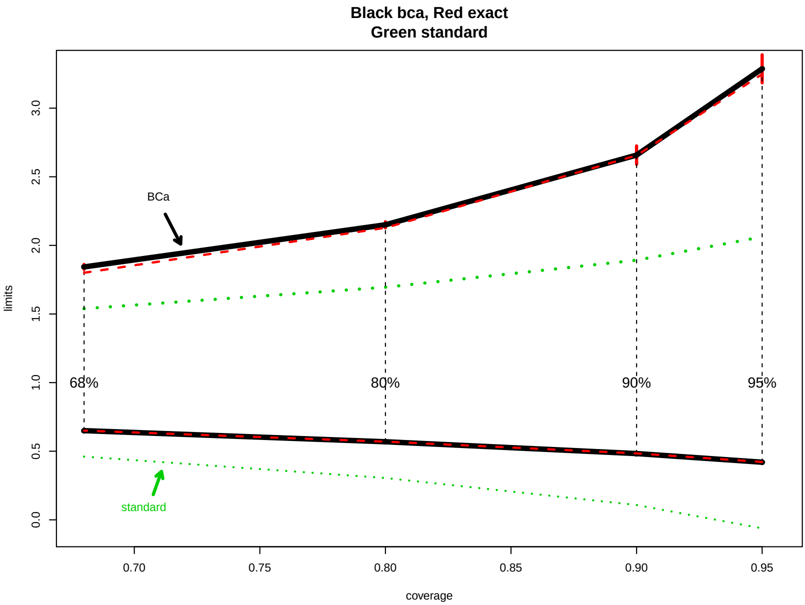
- BCa uses bootstrap resampling to correct all three
- Corrections are of order $c \big/ \sqrt{n}$
- Student-t corrections of order c/n

RATIO OF TWO NORMAL VARIANCE ESTIMATES

- Observe $\hat{\sigma}_1^2 \sim \sigma_1^2 \left(\frac{\chi_{10}^2}{10} \right)$ and $\hat{\sigma}_2^2 = \sigma_2^2 \left(\frac{\chi_{42}^2}{42} \right)$ (independent)
- Parameter of interest: $\theta = \frac{\sigma_1^2}{\sigma_2^2}$ so $\hat{\theta} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$
- Exact confidence limits $\hat{\theta}_{\text{exact}}(\alpha) = \hat{\theta} \Big/ F_{10,42}^{(1-\alpha)}$
- $B = 16,000$ parametric bootstrap gave. . .

16,000 Parametric Bootstrap Replications for the Variance Ratio statistic




Black bca, Red exact
Green standard
limits
coverage
BCa
standard
68%
80%
90%
95%
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.70
0.75
0.80
0.85
0.90
0.95

THE DIABETES DATA

- $n = 442$ patients
- $x_i = (c_i, y_i) \quad \begin{cases} c_i \text{ vector of 10 baseline predictors} \\ y_i \text{ measure of disease progression, 1 year} \end{cases}$
- Data matrix $\quad \boldsymbol{x}\ 442 \times 11$
- $\hat{\theta} =$ adjusted R^2 of y regressed on c

$$\hat{\theta} = \text{lm(y} \sim \text{c)\$ adj.r.squared} = 0.507$$

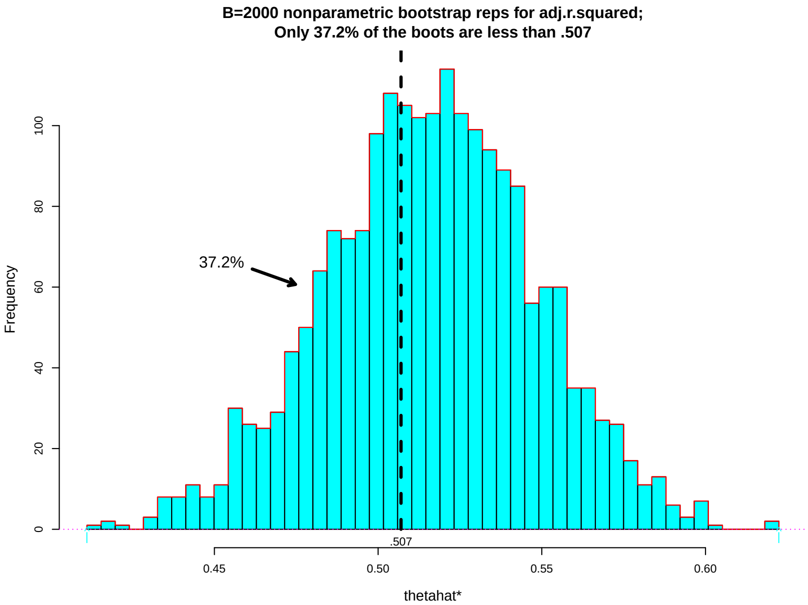
Diabetes Data

$n = 442$ subjects, 10 predictors, response y “progression”.

age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu	spg	y
59	1	32.1	101	157	93.2	38	4	2.11	87	151	−1.13
48	0	21.6	87	183	103.2	70	3	1.69	69	75	−77.13
72	1	30.5	93	156	93.6	41	4	2.03	85	141	−11.13
24	0	25.3	84	198	131.4	40	5	2.12	89	206	53.87
50	0	23.0	101	192	125.4	52	4	1.86	80	135	−17.13
23	0	22.6	89	139	64.8	61	2	1.82	68	97	−55.13
36	1	22.0	90	160	99.6	50	3	1.72	82	138	−14.13
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

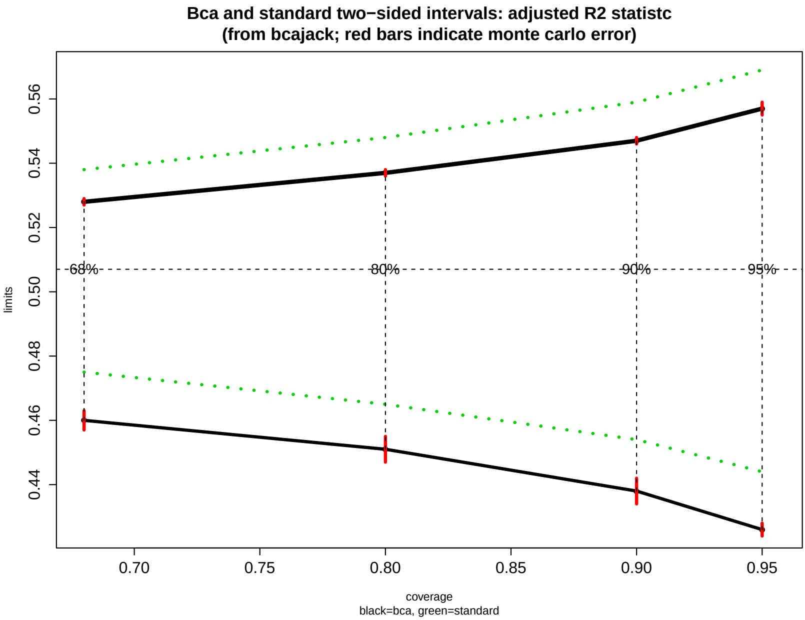
NONPARAMETRIC BOOTSTRAPPING: BCAJACK

- Data matrix $\underset{n\times p}{\boldsymbol{x}}$ e.g., $x_i = (c_i, y_i)$
- $\hat{\theta} = t(\boldsymbol{x})$ estimates parameter of interest θ
- Bootstrap data matrix $\boldsymbol{x}^*$: randomly choose n rows from $\boldsymbol{x}$ *with replacement*
- Gives $\hat{\theta}^* = t(\boldsymbol{x}^*)$
- $\left\{\hat{\theta}^*(1), \hat{\theta}^*(2), \ldots, \hat{\theta}^*(B)\right\}$ provides inferences for θ $\left(\hat{\sigma}_{\text{boot}} = \text{ standard deviation of } \hat{\theta}^*\right)$


B=2000 nonparametric bootstrap reps for adj.r.squared;
Only 37.2% of the boots are less than .507
37.2%
Frequency
0
20
40
60
80
100
.507
0.45
0.50
0.55
0.60
thetahat*

CORRECTING THE STANDARD INTERVALS, CONTINUED

- Standard intervals assume $\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2)$
- BCa makes three corrections:
  1. *For normality* using boot cdf $\widehat{G}$
  2. *For bias* using $\widehat{G}(\hat{\theta})$ (proportion of $\hat{\theta}^*$s less than $\hat{\theta}$)
  3. *For "acceleration"* (i.e., nonconstant σ^2) using jackknife


Bca and standard two−sided intervals: adjusted R2 statistc
(from bcajack; red bars indicate monte carlo error)
limits
0.44
0.46
0.48
0.50
0.52
0.54
0.56
68%
80%
90%
95%
0.70
0.75
0.80
0.85
0.90
0.95
coverage
black=bca, green=standard

THE BCA LEVEL α ENDPOINT

- Let Φ be standard normal cdf and $z^{(\alpha)} = \Phi^{-1}(\alpha)$

$$\hat{\theta}_{\text{bca}}(\alpha) = \widehat{G}^{-1}\Phi\left(\hat{z}_0 + \frac{z^{(\alpha)} + \hat{z}_0}{1 - \hat{a}(z^{(\alpha)} + \hat{z}_0)}\right)$$

- $\widehat{G}$ bootstrap cdf
- $\hat{z}_0 = \Phi^{-1}\widehat{G}\left(\hat{\theta}\right)$ “bias corrector”
- $\hat{a}$ “acceleration” estimate

BCa Properties

- **Transformation invariant**

  if $\lambda = m(\theta)$ then $\hat{\lambda}_{\text{bca}}(\alpha) = m\left(\hat{\theta}_{\text{bca}}(\alpha)\right)$

- *Exact* if $\hat{\lambda} = m\left(\hat{\theta}\right)$ normal translation family

  ($\hat{\lambda} = \frac{1}{2} \log \frac{1+\hat{\theta}}{1-\hat{\theta}}$ for normal correlation coefficients)

- Second-order accurate Claimed coverage accurate to $O(1/n)$

- Reduces to standard intervals if $\widehat{G}$ normal, $\hat{z}_0 = 0$, $\hat{a} = 0$

JACKKNIFE CALCULATION OF ACCELERATION a

- $\boldsymbol{x}_{(i)}$ is $\boldsymbol{x}$ with ith row removed
- $\hat{\theta}_{(i)} = t(\boldsymbol{x}_{(i)})$ and $\hat{\theta}_{(\cdot)} = \sum_1^n \hat{\theta}_{(i)} \Big/ n$
- $D_i = \hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}$
- *Acceleration* $$\hat{a} = \frac{1}{6} \sum_1^n D_i^3 \Bigg/ \left(\sum_1^n D_i^2 \right)^{3/2}$$
- Jackknife standard error $$\left(\frac{n-1}{n} \sum_1^n D_i^2 \right)^{1/2}$$

GROUPED JACKKNIFE

- n can be enormous these days
- partition $\boldsymbol{x}$ into m groups of size $g = n/m$:

$$X_k = \{x_{i_1}, x_{i_2}, \ldots, x_{i_g}\} \qquad (k = 1, 2, \ldots, m)$$

- Now $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$ has just m "points"
- Diabetes $\quad m = 40,\ g = 11 \quad$ (2 left over)

$$\boxed{\texttt{bcajack}(\boldsymbol{x}, 2000, \text{rfun}, m = 40)}$$

PARTIAL OUTPUT OF BCAJACK(X, 2000, RFUN, M=40)

x = DIABETES DATA, RFUN = ADJUSTED R^2

α	bcalims	jacksd	standard
.025	.424	.004	.443
.16	.465	.001	.474
.5	.497	.001	.507
.84	.529	.001	.539
.975	.560	.002	.570

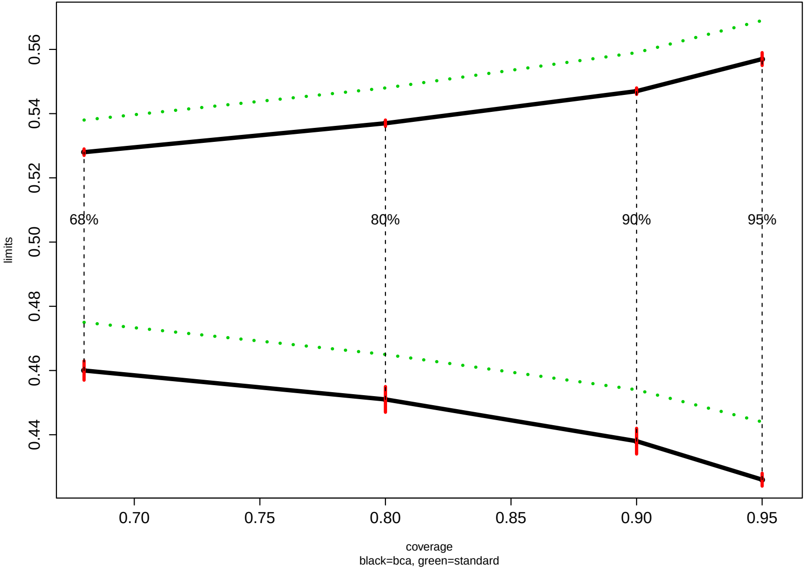
MONTE CARLO ERROR

- *Two types of error* “sampling” and “Monte Carlo”
- $\hat{\sigma}_{\text{boot}}$ concerns sampling error of $\hat{\theta}$
- Was $B = 2000$ enough bootstrap replications?
- Answer Jackknife the bootstrap replications.

JACKKNIFE ESTIMATES OF MONTE CARLO ERROR

- Randomly partition bootstrap replications $\left\{\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, \ldots, \hat{\theta}^*_{(2000)}\right\}$ into 10 groups of 200 each
- Remove one group at a time, and rerun `bcajack`
- Use jackknife formula to get “jacksd”
- No new bootstrapping required

Monte Carlo error is small for upper limits, not so small for lower limits



MORE OUTPUT OF BCAJACK(X, 2000, RFUN, M=40)

	$\hat{\theta}$	sdboot	$\hat{z}_0$	$\hat{a}$	sdjack
estimate	.507	.033	−.327	−.004	.034
jacksd	.000	.001	.027	.000	.000

PARAMETRIC EXAMPLE: THE PEDIATRIC DEATH DATA

- 800 cases African facility for very sick babies
- 600 survived, 200 died
- 11 predictors $\boldsymbol{x}_i = (\text{respiration, heart rate, weight, } \ldots)$
- *Logistic regression model*

π_i death probability $(i = 1, 2, \ldots, 800)$

$$y_i = \begin{cases} 1 & \text{probability } \pi_i \\ 0 & \text{probability } 1 - \pi_i \end{cases} \qquad \text{where } \pi_i = \frac{1}{1 + e^{-x_i'\alpha}}$$

$\theta =$ "resp" coefficient α_1

PARAMETRIC BOOTSTRAPPING

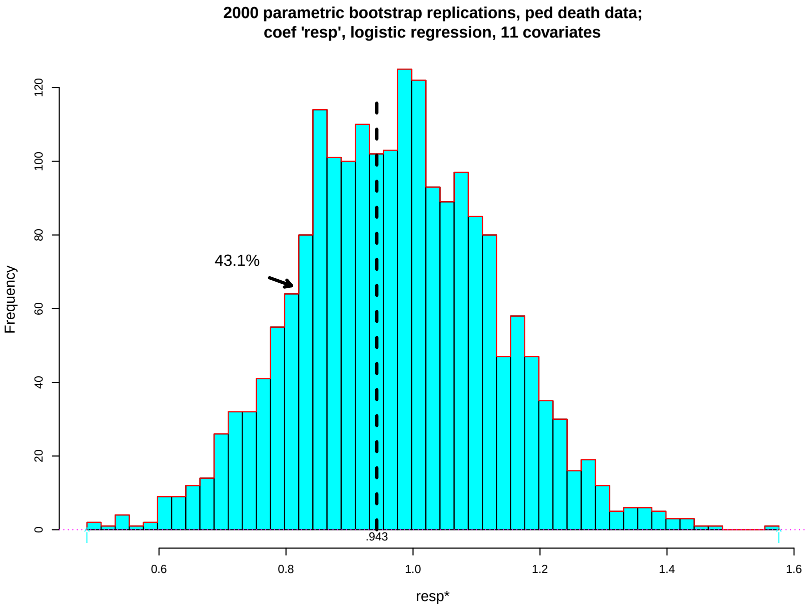
- $\texttt{glm}\left(\boldsymbol{y} \sim \underset{800\times 11}{\boldsymbol{X}}, \text{binomial}\right)$: MLE $\hat{\alpha}$; $\hat{\theta} = \hat{\alpha}_1 = 0.943$
- $\hat{\pi}_i = 1\Big/\left(1 + e^{-x_i'\hat{\alpha}}\right)$
- $y_i^* = 1$ or 0 prob $\hat{\pi}_i$ or $1 - \hat{\pi}_i$
- $\texttt{glm}(\boldsymbol{y}^* \sim \boldsymbol{X}, \text{binomial})$ gives $\hat{\alpha}^*$ and $\hat{\theta}^* = \hat{\alpha}_1^*$
- Also need $b^* = \boldsymbol{X}'\boldsymbol{y}^*$ for $\hat{a}$.

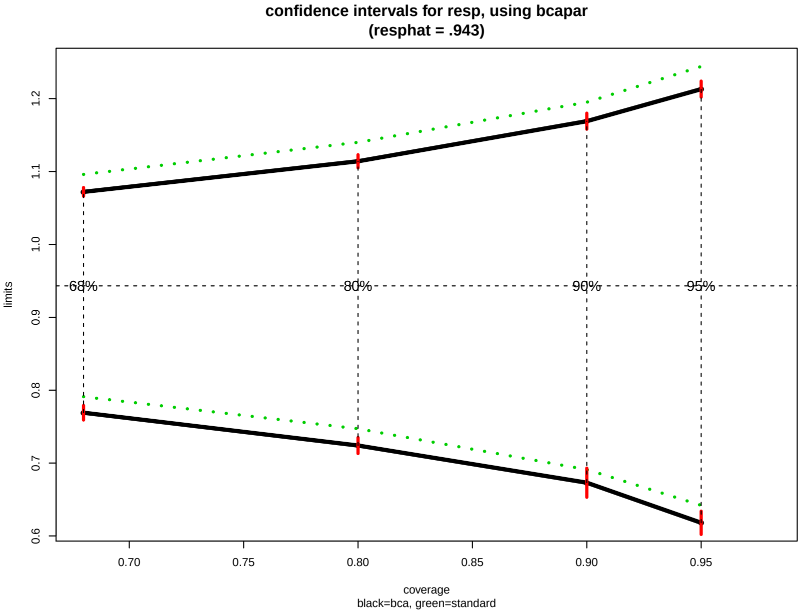
PROGRAM bcapar

- B bootstrap samples give

$$\hat{\boldsymbol{\theta}}^* = \left(\hat{\theta}^*(1), \hat{\theta}^*(2), \dots, \hat{\theta}^*(B)\right) \text{ and } \underset{11\times B}{\boldsymbol{b}^*} = \left(b^*(1), b^*(2), \dots, b^*(B)\right)$$

- `bcapar`$\left(\hat{\theta}, \hat{\boldsymbol{\theta}}^*, \boldsymbol{b}^*\right)$ yields `bca` limits
- $\hat{\theta}$ and $\hat{\theta}^*$ give $\widehat{G}$ and $\hat{z}_0$
- $\boldsymbol{b}^*$ gives acceleration $\hat{a}$

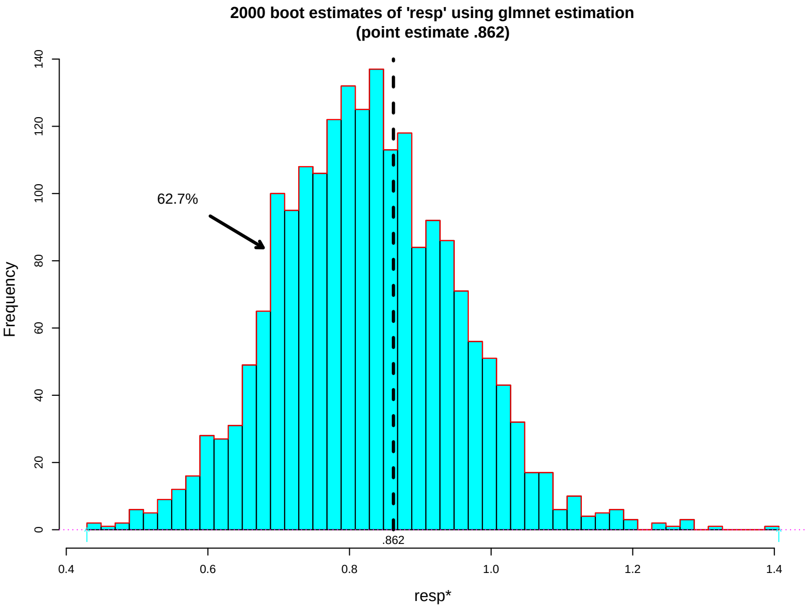

2000 parametric bootstrap replications, ped death data;
coef 'resp', logistic regression, 11 covariates
Frequency
0
20
40
60
80
100
120
43.1%
.943
0.6
0.8
1.0
1.2
1.4
1.6
resp*

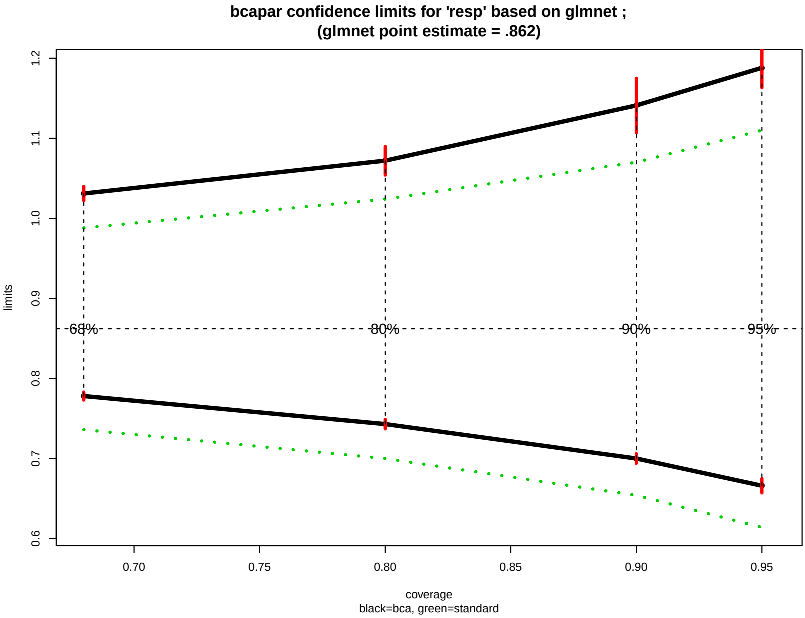

confidence intervals for resp, using bcapar
(resphat = .943)
limits
0.6
0.7
0.8
0.9
1.0
1.1
1.2
68%
80%
90%
95%
0.70
0.75
0.80
0.85
0.90
0.95
coverage
black=bca, green=standard

GLMNET ESTIMATE OF α

(INSTEAD OF MLE)

- `glmnet`$(\boldsymbol{X}, \boldsymbol{y})$ fits increasing succession of logistic regressions; cross-validation used to choose “best” one (*shrinkage*)
- `glmnet` gave $\hat{\alpha}$ and $\hat{\theta} = \hat{\alpha}_1 = 0.862$ (cf 0.943 for MLE)
- Bootstrapping $\hat{\alpha} \longrightarrow \hat{\boldsymbol{\pi}} \longrightarrow \boldsymbol{y}^*$ and then `glmnet`$(\boldsymbol{X}, \boldsymbol{y}^*)$ gives $\hat{\theta}^*$ and $b^* = \boldsymbol{X}'\boldsymbol{y}^*$
- Confidence limits from `bcapar`$\left(\hat{\theta}, \hat{\boldsymbol{\theta}}^*, \boldsymbol{b}^*\right)$


2000 boot estimates of 'resp' using glmnet estimation
(point estimate .862)
Frequency
0
20
40
60
80
100
120
140
62.7%
.862
0.4
0.6
0.8
1.0
1.2
1.4
resp*

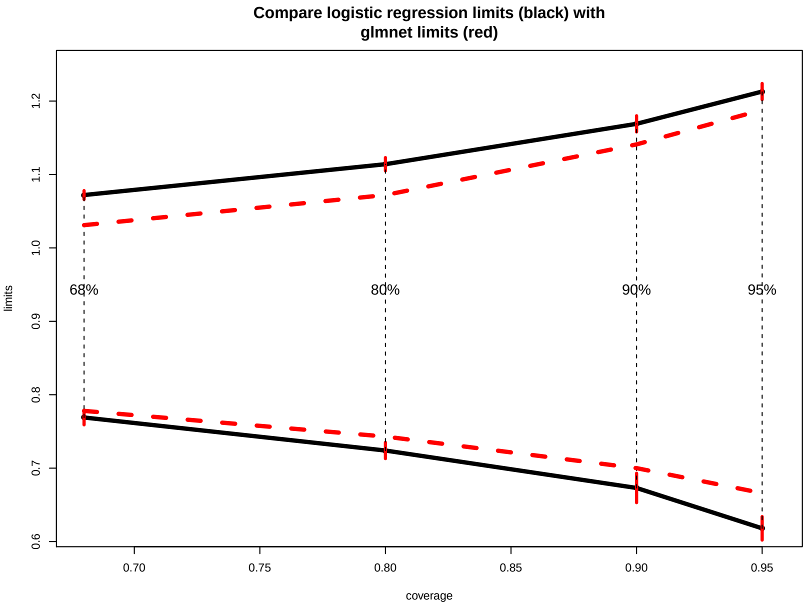

bcapar confidence limits for 'resp' based on glmnet ;
(glmnet point estimate = .862)
limits
0.6
0.7
0.8
0.9
1.0
1.1
1.2
68%
80%
90%
95%
0.70
0.75
0.80
0.85
0.90
0.95
coverage
black=bca, green=standard

COMPARING MLE AND `GLMNET` ESTIMATES OF "RESP"

- `glmnet` estimate $\hat{\theta}$ is smaller than MLE:

`glmnet`	$.86 \pm .13$
MLE	$.94 \pm .15$

- However `bcapar` confidence limits have almost the same centering (but `glmnet` intervals shorter)


Compare logistic regression limits (black) with
glmnet limits (red)
limits
coverage
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.70
0.75
0.80
0.85
0.90
0.95
68%
80%
90%
95%

REFERENCES

DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals. *Statist. Sci.* 11: 189–228.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge: Cambridge University Press, Institute of Mathematical Statistics Monographs (Book 5).

Efron, B. and Narasimhan, B. (2018a). Automatic construction of bootstrap confidence intervals, submitted; available from the author's web site `efron.web.stanford.edu`.

Efron, B. and Narasimhan, B. (2018b). R package `bcaboot`. Available from CRAN.