

PS 787 Assignment 5 (Dec 1, 2007, due Dec 12)

The data in `sub8804.csv` are from the American National Election Study (ANES) surveys of 1988, 1992, 1996, 2000 and 2004. The data in `sub8804a.csv`, `sub8804b.csv` and `sub8804c.csv` are simple random samples (without replacement) of, respectively, 180, 90 and 20 observations taken from `sub8804.csv`. To be precise, `sub8804a.csv` is a sample drawn from `sub8804.csv`, `sub8804b.csv` is a sample drawn from `sub8804a.csv` and `sub8804c.csv` is a sample drawn from `sub8804b.csv`.

The data come from the ANES Cumulative Data File. For more information about the ANES and the surveys, see the sources cited in Assignment 3. The `.csv` files contain four variables:

1. **age**: VCF0101 (“Respondent Age”)
2. **residence**: VCF9002 (“R Length of Residence in Home”); codes: 1 = 4 years or less, 2 = 5-9 years, 3 = 10-19 years, 4 = 20-29 years, 5 = 30 or more years (2004: includes 76 years or more), 6 = ‘All of life’ (regardless of number of years)
3. **turnout**: recoded version of VCF0704 (“Party of R Vote: President- Major Candidates”); codes: TRUE = voted for Democrat, Republican, or Major third party candidate (Perot 1992,1996), FALSE = Did not vote
4. **partisan**: recoded version of VCF0301 (“7-pt Scale Party Identification”); codes: 1 = Strong Democrat, Weak Democrat, Weak Republican, Strong Republican, 0 = Independent-Democrat, Independent-Independent, Independent-Republican,

Observations with code = 0 on variable VCF0702 (“Did R Vote in the November Elections”) are omitted. Also omitted are observations with missing data for **age**, **residence** or **partisan**.

This exercise is about nonparametric bootstrap, highlighting the differences among different kinds of bootstrap confidence intervals.

The task is to model individuals’ choices whether to vote (i.e., $\text{turnout}_i = 1$). The model of interest uses a logit regression model to estimate the following choice probabilities:

$$\text{Prob}(\text{turnout}_i = 1) = b_1 + b_2 \text{partisan}_i$$

To obtain bootstrap interval estimates, we will use a nonparametric approach based on resampling (with replacement) whole observations. That is, the bootstraps are based on resamples of pairs of values (turnout_i , partisan_i). For the samples with 180 and 90 observations we use $R = 1000$ resamples. In the sample with 20 observations we start with $R = 2000$ resamples because, as noted below, a high proportion of the resamples do not produce meaningful parameter estimates.

The **R** program `boot8804.R` estimates the model in all four datasets and computes several kinds of bootstrap confidence intervals for the estimates using each of the smaller datasets. Results from running that program via the command line (in Linux)

```
R CMD BATCH --no-save boot8804.R
```

are in `boot8804.Rout`. The `boot.ci` function is used to compute five kinds of confidence intervals. I use the following notation to define the intervals.

t , parameter estimate in the original (small) sample

v , variance of the parameter estimate in the original sample based on the observed information

$z_\alpha = \Phi^{-1}(\alpha)$, the normal ordinate for $0 < \alpha < 1$ (Φ is the normal cumulative distribution function)

v_R , the variance of the parameter estimate across resamples

b_R , the resampling bias estimate

t^* , parameter estimate in a resample

v^* , variance of the parameter estimate in a resample based on the observed information

$z^* = (t^* - t)/v^{*1/2}$, z -score in a resample

$t_{((R+1)\alpha)}^*$, $(R+1)\alpha$ ordered value of the resample parameter estimates

$z_{(R+1)\alpha}^*$, $(R+1)\alpha$ ordered value of the resample z -scores

The five bootstrap intervals (for $\alpha < .5$) are as follows ($\hat{\theta}_\alpha$ denotes an estimated parameter confidence limit).

“Normal”: $\hat{\theta}_\alpha, \hat{\theta}_{1-\alpha} = t - b_R \pm v^{1/2} z_{1-\alpha}$

“Basic”: $\hat{\theta}_\alpha = 2t - t_{((R+1)(1-\alpha))}^*$, $\hat{\theta}_{1-\alpha} = 2t - t_{((R+1)\alpha)}^*$

“Studentized”: $\hat{\theta}_\alpha = t - v_R^{1/2} z_{((R+1)(1-\alpha))}^*$, $\hat{\theta}_{1-\alpha} = t - v_R^{1/2} z_{((R+1)\alpha)}^*$

“Percentile”: $\hat{\theta}_\alpha, \hat{\theta}_{1-\alpha} = t_{((R+1)\alpha)}^*, t_{((R+1)(1-\alpha))}^*$

“BCa”: BC_a interval

For the smallest dataset (`sub8804c.csv`), a high proportion of the estimation attempts fail across the bootstrap resamples due to perfect or quasiperfect separation. Code in `boot8804.R` excludes those problematic resamples.

Assignment: Explain the differences among the various confidence interval estimates for the b_2 coefficient parameter. The normal theory interval, $\hat{b}_2 \pm v^{1/2} z_{1-\alpha}$, is symmetric around the estimate \hat{b}_2 in each respective sample. Why are the bootstrap intervals not similarly symmetric? Why do the different intervals vary more greatly from one another as the sample size shrinks from 180 down to 20? At the smallest sample size, which intervals are producing the most credible results? Why do you say so?

Extra credit: Using the original dataset `sub8804.csv` as the population (more precisely, fully conditioning on that sample), carry out a Monte Carlo sampling experiment to assess

the performance of the various bootstrap confidence intervals in samples of size 20 drawn from `sub8804.csv`.

For all these questions, present a detailed response supported by the data in `nes8804e.csv` (you need not use **R** to do your analysis).

Background: File `mksubdat8804.R` contains the **R** program used to draw the initial set of random samples from `sub8804.csv`.