PS 787 Assignment 4 (Nov 26, 2008)

This describes the second part of an assignment that explores different ways of making inferences about parameters in the data and GEV models that were the focus of Assignment 2. This part concerns boostrap. A preliminary due date is December 8.

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

The assignment is to compare different bootstrap CIs for some of the parameters in one of the GEV specifications. In particular, to facilitate comparisons with the profile likelihood method highlighted in the first part of this assignment (numbered Assignment 3), consider the same GEV model used there, the one based on the function

$$G_3 = \left(v_T^{1/(1-\sigma)} + v_O^{1/(1-\sigma)}\right)^{1-\sigma} + \left(v_R^{1/(1-\tau)} + v_D^{1/(1-\tau)}\right)^{1-\tau}$$

where as before we define

$$\begin{aligned} Z_T &= X_1 b_1 && \text{(don't vote)} \\ Z_O &= X_2 b_2 && \text{(rolloff)} \\ Z_R &= X_3 b_3 && \text{(Republican vote)} \\ Z_D &= -X_3 b_3 && \text{(Democratic vote)} \end{aligned}$$

with $v_T = \exp(Z_T)$, $v_O = \exp(Z_O)$, $v_R = \exp(Z_R)$ and $v_D = \exp(Z_D)$.

Five kinds of bootstrap intervals are of interest in this assignment, I use the following notation to define the intervals.

t, parameter estimate in the original 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. These definitions give only the form for two-sided intervals for a scalar parameter. $\hat{\theta}_\alpha$ 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": $\hat{\theta}_\alpha, \hat{\theta}_{1-\alpha} = t^*_{((R+1)(\alpha))}$, $t^*_{((R+1)(1-\alpha))}$

Compare parametric and nonparametric forms of these intervals for the same five sets of parameters in focus for the first part of this assignment:

1. a lower-tailed interval for σ (pertinent for testing whether $\sigma = 0$)
2. a two-sided interval for one of the coefficients in the vector b_3
3. a two-sided interval for the difference between the coefficients for the same variable in b_1 and b_2
4. a two-sided interval for the difference between the coefficients for the same variable in b_1 and b_3
5. (extra credit) a two-dimensional 95% confidence region for the joint distribution of $\hat{\sigma}$ and $\hat{\theta}$

Packages and functions to help running various kinds of bootstrap are not hard to find. The `boot` library may be useful for working in **R**.

You may find the amount of computing needed to estimate the GEV model in all the boostrap resamples is excessive. In that case it is acceptable to retreat to a simple bivariate logit model for the choice between the vote-for-Republican and vote-for-Democrat alternatives, choosing single parameters, differences between parameters and maybe the joint confidence region for a pair of parameters as you see fit.

In a straightforward parametric model with nothing special going on, the main benefit of the boostrap can be seen when the different kinds of intervals are compared in small samples. I hesitate to suggest trying this with the GEV models, but with the bivariate logit model there is no good reason not to play with it. So if this avenue interests you, you should try estimating the bivariate logit model and then the different bootstraps with a very small random sample of the observations in the original data set. I suggest trying three smaller sample sizes; random samples of size 120, of size 60 and of size 30. To guarantee the model is estimable in the smallest sample, you may need to stratify the random sample of the ANES observations you select.