Notes on bootstrap Fall 2009

How to generate data and set up confidence intervals.

Bootstrap can be used to construct CI in nonlinear regression models, or when the error distribution is non-normal. Here's a few notes on bootstrap.

1 Generating data

In class we talked about *parametric* and *non-parametric* procedures.

1.1 Parametric bootstrap

Let's say you have a data set consisting of n data objects $(x_i, y_i)_{i=1}^n$ where x_i may be a vector of multiple explanatory variable observations.

- Fit your model to the data ($y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij}$). Record the coefficient estimates $\hat{\beta}_j$, $j = 0, \cdots, p-1$
- Generate a new set of residuals $(e_i^b)_{i=1}^n$ from distribution F_e.
  F_e is for example $N(0, \hat{\sigma}^2)$ (with $\hat{\sigma}^2$ estimated from the model fit). In the lab you can also simulate errors from t_{df} where you choose the degrees of freedom: set e_i to $\hat{\sigma} * t_i$ where $t_i \sim t_{df}$.
- Create the bootstrap data set $(x_i, y_i^b)_{i=1}^n$ where $y_i^b = \hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j x_{ij} + e_i^b$

1.2 Non-parametric bootstrap

Let's say you have a data set consisting of n data objects $(x_i, y_i)_{i=1}^n$ where x_i may be a vector of multiple explanatory variable observations.

- Fit your model to the data ($y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij}$). Record the coefficient estimates $\hat{\beta}_j$, $j = 0, \cdots, p-1$
- Standardize the residuals $(\tilde{e}_i)_{i=1}^n = (e_i/\sqrt{1-h_{ii}})_{i=1}^n$.
- Draw new residuals e_i^b by resampling from $\tilde{e}_i$ (draw n residuals e_i^b from $(\tilde{e}_i)_{i=1}^n$ with replacement).
- Create the bootstrap data set $(x_i, y_i^b)_{i=1}^n$ where $y_i^b = \hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j x_{ij} + e_i^b$

Note, a second variant of non-parametric bootstrap is to resample the data pairs (x_i, y_i) directly, without specifying a model.

1.3 The principle

The idea behind bootstrap is to create a data set for which we know the true model, and can thus investigate (or estimate) the sampling distribution of estimators of interest directly. In the above examples, for bootstrap data y^b the true model is $\hat{\beta}_j$, $j = 0, \cdots, p-1$. The principle underlying bootstrap is thus

$$\frac{\hat{\beta}_j^b - \hat{\beta}_j}{SE(\hat{\beta}_j^b)} =^d \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)}.$$

By estimating, or observing, the sampling distribution of $\hat{\beta}_j^b$ (or the pivotal element $\frac{\hat{\beta}_j^b - \hat{\beta}_j}{SE(\hat{\beta}_j^b)}$) across multiple bootstrap data sets, we get an idea about the sampling distribution of $\hat{\beta}_j$ from the real data.

2 Bootstrap confidence intervals

There are several approaches to constructing bootstrap confidence intervals.

The normal-theory interval assumes that the statistic T (e.g. a regression coefficient estimate) is normally distributed, and uses the bootstrap estimate of sampling variance, and perhaps of bias, to construct a 100(1 - α)-percent confidence interval of the form

$$(T - B^*) \pm z_{1-\alpha/2} \widehat{SE(T)},$$

where z refers to the normal distribution quantiles (e.g., 1.96 for a 95-percent confidence interval, where $\alpha = .05$).
Here, B^* is the bias estimate obtained from the bootstrap: $\sum_{b=1}^{B}(T - T^b)/B$, and $\widehat{SE(T)} = \sqrt{\sum_{b=1}^{B}(T^b - T^*)/(B-1)}$ is the bootstrap estimated standard error of statistics T (where $T^* = \sum_{b=1}^{B} T^b/B$).
An alternative approach, called the bootstrap percentile interval, is to use the empirical quantiles of T^b to form a confidence interval: $[T^b(.025), T^b(.0975)]$, where the end-points of the interval correspond to the 2.5% and 97.5% percentiles of the bootstrap statistic values.

2.1 Pivotal method

The confidence intervals based on

$$(T - B^*) \pm z_{1-\alpha/2} \widehat{SE(T)}$$

assume that the bootstrap T^b have the same sampling distribution as T. However, this is not true if we use the non-pivotal elements. We should instead use the t-statistic $\theta^b = (T^b - T)/SE(T^b)$. We reformulate the confidence intervals as

$$(T - B^*) \pm q_{1-\alpha/2} \widehat{SE(T)},$$

where $q_{1-\alpha/2}$ percentile of θ^b.
If the SE of the estimates T is unknown (like in non-linear regression) we need a second level of

bootstrap to estimate $SE(T^b)$. For each bootstrap sample b, we generate a second set of bootstrap samples $u, u = 1, \cdots, U$. Each bootstrap data u provides an estimate $T^{u(b)}$. We compute $\widehat{SE(T^b)} = \sqrt{\sum_{u=1}^{U}(T^{u(b)} - T^b)/(U-1)}$. Usually we can get away with a smaller value for U than B, e.g. $U = 25$ and $B = 1000$.

In class, we used the most simple pivotal method where we assume $SE(T)$ is known. In linear regression models $SE(\hat{\beta}_j)$ is $\hat{\sigma} * \sqrt{(X'X)_{jj}}$. Each pivotal element is thus $\frac{\hat{\beta}_j^b - \hat{\beta}_j}{SE(\hat{\beta}_j)}$. For lab 3b, you can use this method since you are only working with linear regression models.

2.2 If you're curious: More on the Percentile method

The percentile method is very easy to use. We simple generate bootstrap data and compute the corresponding estimates T^b. The confidence interval is obtained from the lower and upper percentiles of the bootstrap estimates.

However, when the distribution of T is skewed we can get poor coverage of the confidence intervals generated with the percentile methods. There's been work to alleviate this problem using a *bias-corrected and accelerated* bootstrap method proposed by Brad Efron.

- Start by computing the proportion of T^b that is below the original T: P.
- Compute the corresponding normal quantile $\phi^{-1}(P)$. For example, if 50% of the $T^b < T$, $\phi^{-1}(.5) = 0$. (In R, simple take P and use the function `qnorm(P)`). The value $z = \phi^{-1}(P)$ is called the *correction factor.*
- Compute $A = \frac{\sum_{i=1}^{n}(T_{-i} - \bar{T})^3}{6[\sum_{i=1}^{n}(T_{-i} - \bar{T})^2]^{3/2}}$
- Using A we compute

$$a_1 = \phi(z + \frac{z - z_{1-\alpha/2}}{1 - a(z - z_{1-\alpha/2})})$$

and

$$a_2 = \phi(z + \frac{z + z_{1-\alpha/2}}{1 - a(z + z_{1-\alpha/2})})$$

- Pick q_{lower} as the largest integer $< B * a_1$, and q_{upper} and the smallest integer $> B * a_2$. Note that if the correction factors $z = 0$ and $a = 0$, $q_{lower} = \alpha/2$ and $q_{upper} = 1 - \alpha/2$. That is, if the corrections are 0 we will use the regular percentiles of T^b to set up our confidence interval.
- Finally, the confidence interval is obtained as

$$[T^b(q_{lower}), T^b(q_{upper})]$$

For lab 3b, you can use the standard percentile method. That is, use the percentiles of the bootstrap estimates to form the confidence intervals. Be aware however that the coverage of these intervals may not be accurate, and feel free to explore the bias-corrected accelerated methods in your project.

References

Davison AC, Hinkley DV. Bootstrap Methods and their Application. Cambridge University Press: Cambridge, 1997.
Efron, B. 1979. Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics 7:126.
Efron, B. & R. J. Tibshirani. 1993. An Introduction to the Bootstrap. New York: Chapman and Hall.
DiCiccio TJ, Efron B. Bootstrap confidence intervals. Statistical Science 1996; 11:189-212.
Efron B. Better bootstrap confidence intervals. Journal of the American Statistical Association 1987; 82:171-200.