Regression and Simulation Methods

Week 9: Bootstrap Methods

Clara Panchaud, 8th December 2023

What has been covered over the last week?

- Non-parametric bootstrap methods,
- Parametric bootstrapping methods,
- PIT resampling,
- Bootstrapping GLMs.

Talking through today…

> Standard bootstrap

> PIT bootstrap.

Leaving for now…

> Parametric bootstrapping (see Modern Regression and Bayesian Methods SMSTC course)

> Bootstrapping for GLMs (we'll get to in the R Script)

The procedure for the nonparametric bootstrap is as follows:

1. *Resample*. Create B bootstrap samples by sampling with replacement from the original data $\{r_1, \ldots, r_T\}$.[43] Each bootstrap sample has T observations (same as the original sample)

$$\begin{aligned} \left\{r_{11}^*, r_{12}^*, \ldots, r_{1T}^*\right\} &= 1\text{st bootstrap sample} \\ &\vdots \\ \left\{r_{B1}^*, r_{B2}^*, \ldots, r_{BT}^*\right\} &= \text{Bth boostrap sample} \end{aligned}$$

2. *Estimate* θ. From each bootstrap sample estimate θ and denote the resulting estimate $\hat{\theta}^*$. There will be B values of $\hat{\theta}^*$: $\left\{\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*\right\}$.
3. *Compute statistics*. Using $\left\{\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*\right\}$ compute estimate of bias, standard error, and approximate 95% confidence interval.

9.5.1 Probability Integral Transform Resampling

Probability Integral Transform Resampling (PITR) is a method for bootstrapping observations that are independently but not necessarily identically distributed.

Consider observation y_i with cdf $F_i(y_i)$.

1. Transform y_i to $u_i = \hat{F}_i(y_i)$.
2. Resample from the u's with replacement (or randomly permute the u's) — justified because $F_i(y_i) \sim U(0,1)$ for all i. *i.e.* if F_i were known, the u's would be i.i.d.
3. Reassign the u's to observations by assigning the i^{th} u in the resampled (or permuted) list to the i^{th} observation y_i in the original list. If u_j is assigned to observation y_i, bootstrap observation i is calculated as $\hat{F}_i^{-1}(u_j)$.

Running Commentary

Doing a non-parametric bootstrap for whatever you are trying to calculate is the most conservative way to get your answer and the statistics surrounding it.

For example:
1. You have a sample of 1000 patients who may or may not roughly represent the true population.

2. You run some sort of GLM to get the association between the patients and an outcome (for example death).

3. Doing this once with all data will achieve a good estimate but the error (CI) and p-value associated with it could be blind to small changes in the population. That is to say if we miss a few people off the surrounding statistics change as suddenly our data varies a lot and/or the model fit becomes better/worse.

4. Doing a non-parametric bootstrap will allow for different selections of this population which may or may not actually be representative to vary and provide stable estimates on the CI and p-value.

Next Week…

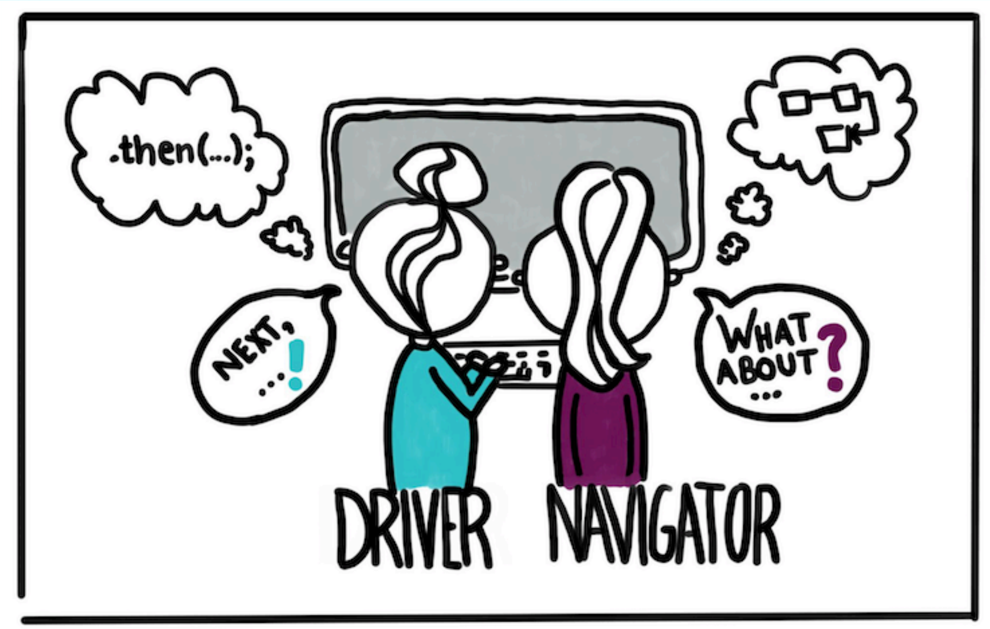
> Now covered all the course content.
> Chance to ask anything else either in or related to the course.
> Will try our best to answer where possible.
> Bring along any questions from throughout the course.
> If we run out of questions, we'll go through the second question on last years assignment which is different but equally useful to look at.

A big note to make is that the SMSTC courses are for your learning. They shouldn't be seen as typical undergrad/masters courses and therefore its down to you what you take away.

With that said try to think more broadly about what you've learnt and not fixate on the assessment too much (it is useful but not the be all and end all).

Rest of the tutorial...

- In pairs work on the ninth notebook. Paired programming will continue!



Rest of the week...

Wait for the assignment and recap everything you have covered in this course!