Lecture 4, Part II: Bootstrapping Regression Models

Text references: Chapter 6.4 in Shalizi

Review: Recall the bootstrap principle for assessing the uncertainty of a statistic, such as, an estimator $\hat{\theta} = g(X_1, ..., X_n)$ of a parameter θ:

$$\begin{array}{rcccccl} \text{Real world:} & F & \Rightarrow & X_1, ..., X_n & \Rightarrow & \hat{\theta} = g(X_1, ..., X_n) \\ \text{Bootstrap world:} & \widehat{F} & \Rightarrow & X_1^*, ..., X_n^* & \Rightarrow & \hat{\theta}^* = g(X_1^*, ..., X_n^*) \end{array}$$

By simulating B bootstrap samples (each of size n), we compute B bootstrap replications $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$. The sample variance v_{boot} of $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$ is an estimator for $\mathbb{V}(\hat{\theta})$.

We can also use the B bootstrap samples to construct a $1-\alpha$ *bootstrap pivotal confidence interval* for θ:

$$C = (2\hat{\theta} - q^*_{1-\alpha/2}, 2\hat{\theta} - q^*_{\alpha/2})$$

Bootstrapping Regression Models

Next we will use the bootstrap approach to assess the variability of the estimates and predictions from a statistical learning method. We recall the basic regression set-up.

Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$ we have two goals:

estimation: Find an estimate $\hat{r}(x)$ of the regression function $r(x) = \mathbb{E}(Y|X = x)$.

prediction: Given a new X, predict Y; we use $\hat{Y} = \hat{r}(X)$ as the prediction.

Questions:

- How do we choose the best model $\hat{r}$ including tuning parameters? (hint: cross-validation)
- How do we assess the variability of coefficient estimates in regression models; such as, the estimated intercept $\hat{\beta}_0$ and estimated slope $\hat{\beta}_1$ in linear regression?
- How do we construct approximate confidence intervals for the regression curve?
- Is the difference in MSE between two regression fits $\hat{r}_1$ and $\hat{r}_2$ significant?

To bootstrap regression models, we need to simulate new data or pairs of observations

$$(X_1^*, Y_1^*) ..., (X_n^*, Y_n^*),$$

refit the regression curve to the bootstrap sample, and finally repeat this whole procedure B times for B bootstrap samples.

There are many options for how to simulate bootstrap samples and how much trust to put in each model (see the next page, Shalizi Sec 6.4, and lecture slides for an example of how to bootstrap a linear regression model by resampling residuals and by resampling cases).

Things to think about for linear regression:

- What are the usual standard linear regression assumptions built into R's `lm` function? How do we check if these assumptions are reasonable?
- What is the difference between the different bootstrap options? How do we check if the model assumptions are reasonable?

Bootstrapping Regressions (cont'd)

A regression is a model for Y conditional on X:

How do we simulate new data $(X_1^*, Y_1^*) ..., (X_n^*, Y_n^*)$? Here are some approaches, in decreasing order of reliance on the model:

- *Fully parametric or model-based bootstrapping.* When does it make sense?

- *Bootstrapping by resampling residuals.* When does it make sense?

- *Nonparametric bootstrapping by resampling cases.* When does it make sense?

So at the end of the day, what are the considerations in choosing parametric versus nonparametric bootstraps? Well, this is just another bias-variance tradeoff, like those we have seen in regression. To cite Shalizi Sec 6.3.1:

"When we have a properly specified model, simulating from the model gives more accurate results (at the same n) than does re-sampling the empirical distribution — parametric estimates of the distribution converge faster than the empirical distribution does. If on the other hand the model is mis-specified, then it is rapidly converging to the wrong distribution."