Stat 200: Introduction to Statistical Inference **Autumn 2018/19**

Lecture 17: Bootstrap II, permutations, hold-outs

Lecturer: Art B. Owen *November 29*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They are meant as a memory aid for students who took stat 200 at Stanford University. They may be distributed outside this class only with the permission of the instructor. Also, Stanford University holds the copyright.*

Abstract

These notes are mnemonics about what was covered in class. They don't replace being present or reading the book. Reading ahead in the book is very effective.

17.1 Bootstrap confidence intervals

The main use of the bootstrap is to get a confidence interval. We can start with the plug-in principle and suppose that

$$\mathrm{Distn}(\hat\theta-\theta; X_i\sim F)\approx\mathrm{Distn}(\hat\theta^*-\hat\theta; X_i^*\sim\hat F). \tag{17.1}$$

Here Distn is a function that takes the distribution F of the data and returns the distribution of our error $\hat\theta-\theta$. Distribution in and distribution out. If you prefer you can work with a function that takes the distribution of F and returns a quantile, such as the 97.5'th percentile of $\hat\theta-\theta$. Then we are back to distribution in and real number out.

As before we generate a large number B of bootstrap samples and we get $\hat\theta^{*b}$ for $b=1,\dots,B$. This time we sort them getting

$$\hat\theta^{*(1)}\leqslant\hat\theta^{*(2)}\leqslant\hat\theta^{*(3)}\leqslant\cdots\leqslant\hat\theta^{*(B)}.$$

We will work with 95% confidence because it is familiar. We get the 2.5% and 97.5% quantiles of the bootstrap sample, calling them $\hat\theta^{*(0.025B)}$ and $\hat\theta^{*(0.975B)}$ respectively. We get to choose B so we can make sure that $.025B$ and 0.975 are integers, but even if they were not, we could define our quantiles by interpolation. We will use $Q^{.975}(\hat\theta;F)$ to represent the given quantile of $\hat\theta$ when $X_1,\dots,X_n\overset{\mathrm{iid}}{\sim}F$.

Now, working in slow motion (slower than we did in class) we develop a confidence interval.

$$\begin{aligned}0.95&=\Pr(Q^{.025}(\hat\theta;F)-\theta\leqslant\hat\theta-\theta\leqslant Q^{.975}(\hat\theta;F)-\theta) &&\text{(defn of quantiles)}\\&\approx\Pr(Q^{.025}(\hat\theta^*;F^*)-\hat\theta\leqslant\hat\theta^*-\hat\theta\leqslant Q^{.975}(\hat\theta^*;F^*)-\hat\theta) &&\text{(plug in)}\\&\doteq\Pr(\hat\theta^{*(.025B)}-\hat\theta\leqslant\hat\theta^*-\hat\theta\leqslant\hat\theta^{*(.97B)}-\hat\theta;\ X_i^*\sim\hat F) &&\text{(bootstrap sampling).}\end{aligned}$$

We then are left with an approximate 95% confidence interval for $\hat\theta-\theta$ of the form

$$0.95\approx\Pr(\hat\theta^{*(.025B)}-\hat\theta\leqslant\hat\theta-\theta\leqslant\hat\theta^{*(.975B)}-\hat\theta).$$

Next we unravel it to get θ by itself in the middle. The result is

$$0.95\approx\Pr(2\hat\theta-\hat\theta^{*(.975B)}\leqslant\theta\leqslant2\hat\theta-\hat\theta^{*(.025B)}).$$

Notice that the upper limit of the confidence interval depends on the lower quantile of the bootstrap sampled values and vice versa.

17.2 Bootstrap bias estimate

Some of our statistics are biased. The bootstrap leads to a bias adjustment. The plug-in idea is

$$\mathbb{E}(\hat\theta - \theta; X_i \sim F) \approx \mathbb{E}(\hat\theta^* - \hat\theta; X_i^* \sim F^*).$$

We can estimate the bias by

$$\widehat{\text{bias}}(\hat\theta) = \frac{1}{n}\sum_{b=1}^{B} \hat\theta^{*b} - \hat\theta \equiv \bar{\hat\theta}^* - \hat\theta$$

and if we subtract the estimated bias from $\hat\theta$ we get

$$\hat\theta - (\bar{\hat\theta}^* - \hat\theta) = 2\hat\theta - \bar{\hat\theta}^*.$$

17.3 Percentile method

Perhaps the most common bootstrap confidence interval is the percentile method which simply takes

$$0.95 \approx \Pr(\hat\theta^{*(.025B)} \leqslant \theta \leqslant \hat\theta^{*(.975B)}).$$

That is, we simply sort the resampled values and use the interval spanned by the central 95% of them.

This does not really have a good plug-in story. Rice remarks about this too. It matches the interval given above if

$$\hat\theta^{*(.025B)} \approx 2\hat\theta - \hat\theta^{*(.975B)},$$

or equivalently

$$\hat\theta \approx \frac{\hat\theta^{*(.025B)} + \hat\theta^{*(.975B)}}{2}.$$

The above approximation is often quite good. For large n we might have a central limit theorem making $\hat\theta^*$ approximately normal and approximately symmetric around $\hat\theta$.

So the percentile method has three components to its justification: plug-in, random sampling, and a symmetry argument.

17.4 Parametric bootstrap

Instead of resampling from the empirical CDF $\hat F$ we can run our Monte Carlo simulations by sampling from $f(x; \hat\theta)$ instead. Of course this requires that we have an available method to sample from our parametric family. That is very often the case. With the parametric bootstrap we stand to gain some precision if our parametric model is good, while losing some if the parametric model is bad.

17.5 Permutations

Suppose we have data $X_1, \dots, X_n \overset{\text{iid}}{\sim} F_x$ independent of $Y_1, \dots, Y_m \overset{\text{iid}}{\sim} F_y$. We might want to test the null hypothesis

$$H_0 : \mathbb{E}(X) = \mathbb{E}(Y).$$

The natural way to do this is to compute $\bar{Y} - \bar{X}$ and reject H_0 if $|\bar{Y} - \bar{X}|$ is unusually large. That leads to the question of defining how large is unusually large.

We can make progress under a somewhat stronger assumption

$$\mathcal{H}_0 : F_x = F_y,$$

which implies H_0 (when the expectations exist). This stronger assumption does not force F_x or F_y to have a parametric form. It just requires them to be equal.

Under $\mathcal{H}_0$ our $n+m$ observations are IID from the common distribution $F = F_x = F_y$. We could write them in order as $(X_1, X_2, \dots, X_n, Y_1, Y_2, \dots, Y_m)$. There are $(n+m)!$ ways to permute the order of these values. For instance one such permutation is $(X_1, Y_3, X_4, Y_n, \dots, X_2)$. If we decided to make $X_1^*, \dots, X_n^*$ be the first n values after permutation and $Y_1^*, \dots, Y_n^*$ be the last m values the distribution of those $n+m$ values would be exactly the same for any permutation under $\mathcal{H}_0$.

The permutation reference distribution is the distribution of $|\bar{Y}^* - \bar{X}^*|$ under all $(n+m)!$ permutations. We only have to do $N = \binom{n+m}{n}$ permutations because the ordering within the first n (or last m) permuted observations does not change $|\bar{Y}^* - \bar{X}^*|$. The permutation p-value is

$$p = \frac{1}{N}\sum_{\ell=1}^{N} 1\{|\bar{Y}^{*\ell} - \bar{X}^{*\ell}| \geqslant |\bar{Y} - \bar{X}|\},$$

where ℓ represents one of the ways of choosing n of the data points to be X_i^* values. For full details of the permutation test (that you might want to read about ***after*** this course) see “Testing Statistical Hypotheses” (3rd edition) by Lehmann and Romano.

Often N is too large for the full test to be done. Then we sample M permutations at random and report

$$p = \frac{1}{M+1}\left(1 + \sum_{\ell=1}^{N} 1\{|\bar{Y}^{*\ell} - \bar{X}^{*\ell}| \geqslant |\bar{Y} - \bar{X}|\}\right).$$

The plus one in numerator and denominator represent the actually observed data values and with out them you might get $p = 0$ (which is very wrong).

Permutation analysis is often used in A/B trials where the allocation of a subject to treatment A or B really is made at random.

We can use permutations on more complicated statistics than $\bar{Y} - \bar{X}$. For instance we could compare the 10’th percentiles.

17.6 Hold outs

In linear regression we estimate the mean square future errors by

$$\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

The future might not match our model for it. Sometimes we can suppose that holding out say 10% of the present data is a reasonable approximation to the future distribution. In that case we can fit our regression to 90% held in data and average the squared errors on the 10% held out.

We don't have to use squared error either. If the problem gives us a more meaningful error measure, perhaps $|Y - \hat{\beta}_0 - \hat{\beta}_1 x|$, we can average that. One of the main reasons for using squared error was that it make the theory easy. Holdouts let us study quantities where the theory would not be so easy.

In more advanced courses you will see that a holdout set can be used to compare two ways of predicting to see which works best. For instance logistic versus logit models.