Week 5 Notes: The Bootstrap

The contents of this week of notes are based largely on Chapter 8 of *All of Statistics* by Larry Wasserman.

1 Some Motivation

Our Goalposts

Recall that the goal of Frequentist inference is to obtain estimators, intervals, and hypothesis tests that have strong properties with respect to the *sampling* distribution (as opposed to the posterior distribution). Given data $\mathcal{D}$ a Frequentist approach might be to construct an interval estimate for a parameter ψ such that

$$G_\theta\{L(\mathcal{D}) \leq \psi \leq U(\mathcal{D})\} = 1 - \alpha,$$

for a desired *confidence level* $1-\alpha$. Such intervals are often of the form $\widehat{\psi} \pm z_{\alpha/2}\, s_{\widehat{\psi}}$, where where $\widehat{\psi}$ is a point estimate, $s_{\widehat{\psi}}$ is an estimate of the standard deviation of $\widehat{\psi}$, and $z_{\alpha/2}$ corresponds to an appropriate quantile of the standard normal distribution. While rarely possible, we would like coverage to hold exactly and without depending on θ.

Misspecified GLMs

For the past several weeks, we have been learning about how to perform inference using generalized linear models (GLMs). In particular, we have learned about how to construct confidence intervals and perform hypothesis tests using the *asymptotic properties* of the likelihood, which allow us to derive sampling distributions like $\widehat{\beta} \overset{\bullet}{\sim} \text{Normal}\{\beta_0, \mathcal{I}(\beta_0, \phi_0)^{-1}\}$, where (β_0, ϕ_0) are the "true" values of the regression parameters β and dispersion parameter ϕ.

An important justification for using the asymptotic properties of the likelihood is that the model is correctly specified. The model might be misspecified in (at least) three ways:

- Maybe we misspecified the stochastic component of the model, e.g., we assumed Y_i was Poisson but really it was negative binomial.
- Maybe we misspecified the systematic component, e.g., maybe we omitted some covariates or the "linear predictor" is actually a nonlinear function of the covariates.
- Maybe the link function was specified incorrectly, e.g., maybe we should have used a log link rather than the logit link in a binomial regression.

In the first case, where the stochastic component is misspecified, note that even though the model is misspecified it still makes sense to try to estimate β. It turns out, however, that the MLE $\widehat{\beta}$ will still generally be asymptotically normal and centered on β_0, but the asymptotic variance will no longer be given by $\mathcal{I}(\beta_0, \phi_0)^{-1}$. *Question:* **is there a simple procedure that will automatically let us estimate the correct variance of the MLE even under this type of model misspecification?**

Complicated Sampling Distributions

More generally, given a statistic $T = T(\mathcal{D})$, one might want an automatic procedure for obtaining the sampling distribution of T, while making minimal assumptions about the data generating process. For example, consider the iid sampling model $X_1, \ldots, X_N \overset{\text{iid}}{\sim} F$. We might aim to construct a confidence interval for, say, the median of F, or a robust measure of scale like the median absolute deviation $\text{MAD} = \text{median}(|X - \text{median}(X)|)$. For *non-linear functionals* $\psi(F)$ (see below for a definition of a linear functional) of the data generating process F this may not be straight-forward to obtain (e.g., the median, inter-quartile range, MAD, standard deviation).

A general approach for obtaining standard errors and confidence intervals for non-linear functionals is the *functional delta method*, which I am not going to cover; suffice to say, it is a generalization of the delta method that is applicable a bit more generally, but is inconvenient in that we often have to sit down and do some math to get it to work. *Question:* **is there a simple procedure that will give us reasonable inference for arbitrary (or nearly arbitrary) non-linear functionals of our data generating process?**

2 The Bootstrap Principle

It is not always possible, given a sample $\mathcal{D} \sim G$, to determine the sampling distribution of a statistic $T = T(\mathcal{D})$. This is because we do not know G; of course, if we knew G, we would not need to do any inference.

The bootstrap gets around this problem by using the data to estimate G from the data to obtain some $\widehat{G}$. Given $\widehat{G}$, we can compute the sampling distribution of $T^\star = T(\mathcal{D}^\star)$ where $\mathcal{D}^\star \sim \widehat{G}$.

The Bootstrap Principle: Suppose that $\mathcal{D} \sim G$, $\psi = \psi(G)$ is some parameter of the distribution G of interest, and $T = T(\mathcal{D})$ is some statistic aimed at estimating ψ. Then we can evaluate the sampling distribution of $T(\mathcal{D})$ by

1. estimating G with some $\widehat{G}$; and
2. using the sampling distribution of $T^\star = T(\mathcal{D}^\star)$ as an estimate of the sampling distribution of T, where $\mathcal{D}^\star \sim \widehat{G}$.

Implementing the bootstrap principle has two minor complications. First, how do we estimate G? Second, how do we compute the sampling distribution of $T(\mathcal{D}^\star)$?

How we estimate G typically depends on the structure of the problem. Suppose, for example, that $\mathcal{D} = (X_1, \ldots, X_N)$ which are sampled iid from F (so that $G = F^N$). Then a standard choice is to use the empirical distribution function $\widehat{F} = \mathbb{F}_N = N^{-1} \sum_{i=1}^N \delta_{X_i}$ where δ_x is the point mass at x (so that $\widehat{G} = \widehat{F}^N$); this is referred to as the *nonparametric bootstrap* because it does not depend on any parametric assumptions about F. Alternatively, we could estimate F using a *parametric* model: that is, we may set $\widehat{F} = F_{\widehat{\theta}}$ where $\widehat{\theta}$ is the maximum likelihood estimate of θ in some parametric family $\{F_\theta : \theta \in \Theta\}$. This is referred to as the *parametric bootstrap*. Other methods for estimating F lead to different bootstrap procedures that may be more appropriate in any given situation (such as the *residual bootstrap* if you are willing to make some assumptions about the correctness of the regression model, or the *block bootstrap* for time-series data).

In all but the simplest settings, Monte Carlo is used to approximate the sampling distribution of $T^\star$. That is, we sample $\mathcal{D}_1^\star, \ldots, \mathcal{D}_B^\star$ independently from $\widehat{G}$ and take $\frac{1}{B} \sum_{b=1}^B \delta_{T_b^\star}$ as our approximation of the sampling distribution of T, where $T_b^\star = T(\mathcal{D}_b^\star)$.

Exercise 1: Bootstrapping Linear Functionals

Suppose that $X_1, \ldots, X_N \overset{\text{iid}}{\sim} F$ and let $\psi(F)$ denote the population mean of F, i.e., $\psi(F) = \mathbb{E}_F(X_i) = \int x\, F(dx)$. We consider bootstrapping the sample mean $\bar{X}_N = N^{-1} \sum_{i=1}^N X_i$ using the approximation $\widehat{F} = \mathbb{F}_N$. That is, we consider the sampling distribution of $\bar{X}^\star = N^{-1} \sum_{i=1}^N X_i^\star$ where $X_1^\star, \ldots, X_N^\star$ are sampled independently from $\mathbb{F}_N$. **Note: none of the answers to these questions involve considering simulated datasets.**

a. What is $\psi(\mathbb{F}_N)$?

b. The *actual* bias of $\bar{X}_N$ is $\mathbb{E}_F\{\bar{X}_N - \psi(F)\} = 0$. What is the *bootstrap estimate* of the bias $\mathbb{E}_{\mathbb{F}}(\bar{X}_N^\star - \bar{X})$?

c. The variance of $\bar{X}_N$ is σ_F^2/N where σ_F^2 is $\mathrm{Var}_F(X_i)$. What is the *bootstrap estimate* of the variance of $\bar{X}$, $\mathrm{Var}_{\mathbb{F}_N}(\bar{X}^\star)$?

d. A parameter ψ is said to be *linear* if it can be written as $\psi(F) = \int t(x)\, F(dx)$ for some choice of $t(x)$. In this case it is natural to estimate ψ using $\bar{T} = N^{-1} \sum_i t(X_i)$. Write down the bootstrap estimate of the bias and variance of $\bar{T}$ in this setting.

3 Bootstrap Variance Estimation

The bootstrap can be used to approximate the variance (or standard deviation) of T as follows:

1. Draw $\mathcal{D}^\star \sim \widehat{G}$ (for example, if $X_1, \ldots, X_N \overset{\text{iid}}{\sim} F$ then we take $X_1^\star, \ldots, X_N^\star \overset{\text{iid}}{\sim} \mathbb{F}$).
2. Compute $T^\star = T(\mathcal{D}^\star)$.
3. Repeat steps 1 and 2 B times to get $T_1^\star, \ldots, T_B^\star$.
4. Let $v_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\left(T_b^\star - \bar{T}\right)^2$ where $\bar{T} = \frac{1}{B}\sum_b T_b^\star$.

Wasserman gives the following pseudo-code for estimating the variance of the sample median of $X_1, \ldots, X_N$:

```
## Let X be a vector of size N, sampled from F
T <- median(X)
Tboot <- numeric(N)
for(i in 1:B) {
  ## Sample N iid draws from the empirical distribution of X
  Xstar <- sample(x = X, size = N, replace = TRUE)
  ## Save the result
  Tboot[i] <- median(Xstar)
}
v_boot <- var(Tboot)
se_boot <- sqrt(v_boot)
```

In addition to the sample median, Wasserman considers estimating the variance of the *sample skewness* given by

$$\frac{\frac{1}{N}\sum_i (X_i - \bar{X})^3}{s^3}$$

where s denotes the sample standard deviation and $\bar{X}$ the sample mean.

4 Types of Bootstrap Intervals

The Normal Interval

Given the sampling distribution of T, we can do things like construct confidence intervals for ψ. For example, it is often the case that T is asymptotically normal and centered at ψ. We can then use the bootstrap estimate of $\text{Var}(T)$ to make the confidence interval

$$T \pm z_{\alpha/2}\sqrt{v_{\text{boot}}}.$$

This is a pretty commonly used approach, but you might guess it only works well if T is approximately normal.

The Basic Percentile Interval

A commonly-taught bootstrap-based $100(1-\alpha)\%$ interval is to take $(T^\star_{\alpha/2}, T^\star_{1-\alpha/2})$ where $T^\star_\gamma$ denotes the $100\gamma^{\text{th}}$ percentile of $T^\star$ (which is again approximated by Monte Carlo). Pseudo-code for this approach would replace the pseudo-code for the variance estimation with

```
quantile(Tboot, c(0.025, 0.975))
```

A downside of this approach is that, despite appearing simple, it is actually not very easy to justify; it sort of treats the samples of $T^\star$ as though they were *samples of ψ from a posterior distribution*, and given that there are no priors/posteriors in sight this seems dubious.

Pivotal Intervals

The next problem motivates the use of *pivotal intervals*. We recall the *delta method* approach to computing standard errors. Suppose that $\widehat{\mu}$ has mean μ and variance τ^2 and that we want to approximate the mean and variance of $g(\widehat{\mu})$. The delta method states that, if τ is sufficiently small, then $\mathbb{E}\{g(\widehat{\mu})\} \approx g(\mu)$ and $\text{Var}\{g(\widehat{\mu})\} \approx g'(\mu)^2\tau^2$. This is based on the somewhat crude approximation

$$g(\widehat{\mu}) \approx g(\mu) + (\widehat{\mu} - \mu)g'(\mu) + \text{remainder}$$

with the remainder being of order $O(\tau^2)$. The delta method approximation is obtained by ignoring the remainder.

Exercise 2: Bootstrapping a Log-Normal

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Normal}(\mu, 1)$ and let $\psi = e^\mu$ and $T = e^{\bar{X}_n}$ be the MLE of ψ. Create a dataset using $\mu = 5$ consisting of $n = 20$ observations.

(a) Use the delta method to get the standard error and 95% confidence interval for ψ.

(b) Use the nonparametric bootstrap to get a standard error and 95% confidence interval for ψ using the normal interval.

(c) The *parametric bootstrap* makes use of the assumption that F (in this case) is a normal distribution. Specifically, we take $\widehat{F}$ equal to its maximum likelihood estimate, the $\text{Normal}(\bar{X}_n, 1)$ distribution. Using the parametric bootstrap, compute the standard error and a 95% confidence interval for ψ.

(d) Plot a histogram of the bootstrap replications for the parametric and nonparametric bootstraps, along with the approximation of the sampling distribution of T from the

delta method (i.e., Normal$(T, \widehat{s}^2)$). Compare these to the true sampling distribution of T. Which approximation is closest to the true distribution?

(e) Depending on the random data generated for this exercise, you most likely will find that the sampling distribution of T estimated by both the bootstrap and the delta method are not so good; the biggest problem is that the sampling distribution will be location-shifted by $T - \psi$. Repeat part (d), but instead comparing the sampling distribution of $T - \psi$ to the bootstrap estimates obtained by sampling $T^\star - T$.

The lesson of part (e) is that the bootstrap approximation is likely to be best when we apply it to *pivotal quantities*. A quantity $S(T, \psi)$ (which is allowed to depend on ψ) is said to be pivotal if it has a distribution that is independent of ψ. For example, in Exercise 2 the statistic $\sqrt{n}(\bar{X} - \mu)$ is a pivotal quantity, and in general $Z = \frac{\sqrt{n}(\bar{X}-\mu)}{s}$ is asymptotically pivotal (where s is the sample standard deviation).

Exercise 3: A Better Pivot

While we saw an improved approximation for $T - \psi$, argue that this is nevertheless not a pivotal quantity. Propose a pivotal quantity $S(T, \psi)$ which is more suitable for bootstrapping.

Both the normal and percentile intervals exercise rely on asymptotic normality, which we may like to avoid. An alternative approach is to apply the bootstrap to $\zeta = T - \psi(F)$ rather than to T directly, so that $\psi(F) = T - \zeta$. If we knew the $\alpha/2$ and $(1 - \alpha/2)$ quantiles of ζ (say, $\zeta_{\alpha/2}$ and $\zeta_{1-\alpha/2}$), then we could form a confidence interval

$$G_\theta(T - \zeta_{1-\alpha/2} \leq \psi \leq T - \zeta_{\alpha/2}) = 1 - \alpha.$$

The *empirical bootstrap* (or *pivotal interval*) confidence interval estimates these quantiles from the quantiles of $T^\star - \psi(\widehat{F})$, which are computed by simulation. More generally, we could use this approach for any pivotal quantity; for example, since $\xi = T/\psi$ is pivotal in Exercise 2, we could use the interval $(T/\xi_{1-\alpha/2}, T/\xi_{\alpha/2})$ as our interval. Roughly speaking, the closer that $S(T, \psi)$ is to pivotal, the better we expect the interval to perform.

Exercise 4: Better Bootstrap for Lognormal

Use the nonparametric bootstrap to make a 95% confidence interval using the pivotal quantity ξ described above.

5 Exercises

Exercise 5: Wasserman 8.1

Consider the following dataset:

```r
# Create a data frame
df <- data.frame(
  LSAT = c(576, 635, 558, 578, 666, 580, 555, 661, 651, 605, 653, 575, 545,
           572, 594),
  GPA = c(3.39, 3.30, 2.81, 3.03, 3.44, 3.07, 3.00, 3.43, 3.36, 3.13, 3.12,
          2.74, 2.76, 2.88, 3.96)
)

# Print the data frame
print(df)
```

```
   LSAT  GPA
1   576 3.39
2   635 3.30
3   558 2.81
4   578 3.03
5   666 3.44
6   580 3.07
7   555 3.00
8   661 3.43
9   651 3.36
10  605 3.13
11  653 3.12
12  575 2.74
13  545 2.76
14  572 2.88
15  594 3.96
```

which are LSAT scores (for entrance to law school) and GPA. Estimate the standard error of the correlation coefficient ρ using the bootstrap. Find a 95 percent confidence interval using the normal, pivotal, and percentile methods.

Exercise 6: Wasserman 8.2

Conduct a simulation to compare the various bootstrap confidence interval methods. Let $N = 50$ and let $\psi = \frac{1}{\sigma^3}\int (x-\mu)^3 F(dx)$ be the skewness. Draw $Y_1, \ldots, Y_N \sim \text{Normal}(0, 1)$ and set $X_i = e^{Y_i}$, $i = 1, \ldots, N$. Construct the three types of bootstrap 95 percent intervals for ψ from the data $X_1, \ldots, X_N$. Repeat this whole thing many times and estimate the true coverage of the three intervals.