Lecture 3: Bayesian Bootstrap*

Bryan S. Graham, UC - Berkeley & NBER

September 4, 2015

The previous chapters emphasized optimal prediction under complete knowledge of the population data distribution. Specifically, given knowledge of the joint distribution of X and Y how does one predict Y for a random draw with $X = x$? Under specific parameterizations of the *loss* associated with prediction error, the linear, mean and quantile regression functions are all optimal prediction functions. By optimal I mean that they minimize average loss, or *risk*, across many replications of our prediction problem.

Evaluating the risk attached to a specific prediction procedure requires knowledge of the data distribution function. Consequently, when this distribution function is unknown, a risk-minizing procedure is generally unavailable. In this lecture we consider prediction when only a random sample of size N from the target population is available. While our random sample provides information about the underlying population, it does not perfectly reveal its properties. We must proceed under varying degrees of uncertainty about the true population distribution. Our inference problem is created by this lack of knowledge.

One approach to handling this uncertainty would be to calculate the risk associated with a given prediction rule under all logically possible data distributions. We could then choose rules that have low maximal risk or *minimax rules* (e.g., Ferguson, 1967, Chapter 2; Lehmann and Casella, 1998, Chapter 5). Here we will develop a different approach. We begin with the observation that the sample provides information about the relative plausibility of different logically possible data distributions. If the sample includes many draws from a certain region of the sample space, then we might reasonably conclude that the true population data distribution attaches a large amount of probability mass to that region. Likewise we may conclude that data distributions which attach little mass to that region are less plausible candidates for the true distribution.

One way to formalize the above intuition is to conceptualize the data distribution as a

[*] This note is adapted from my teaching materials from Ec 240a at Berkeley and makes references to other materials not included here.

random draw from the set of all logically possible data distributions. The frequency with which specific possible data distributions are drawn is governed by a *prior*. After observing the sample we update our beliefs about from which distribution, among all those logically possible, our sample was, in fact, drawn.

Treating the sampled data distribution as random is not straightfoward. Doing so requires placing a distribution on the space of probability measures (i.e., we need a distribution for distributions). We will use results from Ferguson (1973, 1974) for this purpose. The basic idea, however, dates to the work of Pierre Simon Laplace in the late 18th century (Laplace 1774; Stigler, 1986a, 1986b). Laplace, considered a bag of approximately fair coins. It may be that prior to flipping any specific coin we believe that the true probability of observing heads is in the neighborhood of one half. We formalize this, as Laplace did, by assuming that the distribution of heads probabilities across all coins in the bag is uniform over a narrow region centered on 1/2. Once we draw a *specific* coin, and begin to flip it, our beliefs about *its* heads probability will change. A coin that comes up heads more often then tails is more likely to have a true heads probability is excess of 1/2 than below 1/2. Of course, even after many flips of the coin we remain uncertain about its true heads probability.

Our post-sample beliefs are summarized by a *posterior* distribution, again on the space of probability measures. This distribution is computed using the prior, the likelihood for the data, and Bayes rule. Assume, for example, the ultimate object of interest is a vector of linear predictor coefficients. We may compute a predictive distribution for this coefficient vector as follows. First, we draw a candidate distribution from our posterior distribution (on the space of probability measures). Second, using the drawn distribution we compute the linear predictor coefficients. If we repeat this process many times we will get a distribution of linear predictor coefficients. This distribution captures our uncertainty about the magnitudes of the linear predictor coefficients in the population actually sampled. This predictive distribution will reflect our uncertainty about the underlying population data distribution. We attach greater probability mass to predictions that are optimal (i.e., risk minimizing) under joint distributions that, *after having observed the sample in hand*, we believe are more likely to characterize the true sampled population. We attach less probability mass to predictions that are optimal under distributions that, again after having observed the sample in hand, we believe are less likely. In situations where a *single* prediction is called for we may use the average, median or some other quantile of our predictive distribution. These types of point estimates are called Bayes' predictions or *Bayes' rules* (e.g., Ferguson, 1967, Chapter 2; Lehmann and Casella, 1998, Chapter 4).

1 Discrete distributions

Our development will focus on the case where $Z = (X', Y)'$ is a discrete random variable, possibly vector-valued, with J points of support (i.e, $Z \in \{z_1, \ldots, z_J\}$). Importantly, the number of support points may be extremely large. In many cases our assumption of discreteness will be unrestrictive, since even nominally continuously-valued random variables can only be measured (and stored) with finite precision (we will develop some limitations of this argument below; cf., Efron (1982)).

Available is a random sample of size N from the population of interest, $\{Z_i\}_{i=1}^{N}$. Let $\mathbf{Z} = (Z_1', \ldots, Z_N')'$ be the vector containing all N random draws. The probability that a generic random draw takes on the j^{th} possible value is

$$\Pr\left(Z = z_j \middle| \theta\right) = \theta_j, \ j = 1, \ldots, J,$$

where $\theta = (\theta_1, \ldots, \theta_J)'$ is the collection of probabilities attached to each possible realized value of Z. This vector fully characterizes the sampled population. The set of all logically possible population distributions or the *parameter space* is given by the $(J-1)$ probability (unit) simplex

$$\Theta = \mathbb{S}^{J-1} = \left\{ \theta = (\theta_1, \ldots, \theta_J) \in \mathbb{R}^J \, : \, \theta_j \geq 0, \sum_{j=1}^{J} \theta_j = 1 \right\}.$$

Our parameter space is a subset of a finite dimensional Euclidean space, in this sense our approach is ‘parametric’. On the other hand, conditional on the multinomial assumption, it places no restrictions on the form of the true data distribution. In this sense our setup is ‘nonparametric’. Exploiting the dual parametric/nonparametric interpretation of the multinomial likelihood often leads to interesting insights.

If we knew θ with certainty we could proceed as in the previous lectures. For example the vector of coefficients indexing the best linear predictor of Y given X equals, for $z_j = \left(x_j', y_j\right)'$,

$$\begin{aligned} \beta(\theta) &= \mathbb{E}\left[XX'\right]^{-1} \times \mathbb{E}\left[XY\right] \\ &= \left[\sum_{j=1}^{J} \theta_j x_j x_j'\right]^{-1} \times \left[\sum_{j=1}^{J} \theta_j x_j y_j\right], \end{aligned}$$

which is a function of θ. Unfortunately we do not know which value of θ indexes the sample population, and hence $\beta(\theta)$.

The gamma function

The function

$$\Gamma\left(x\right)=\int_{t=0}^{t=\infty}t^{x-1}\exp\left(-t\right)\mathrm{d}t, \tag{1}$$

which appears in the Dirichlet density function (4), is called the is the *gamma function.* It has several special properties. First, it is finite if $x>0$. Second, integration by parts (with $u\left(t\right)=t^{x}$ and $v\left(t\right)=-\exp\left(-t\right)$) gives

$$\int_{t=a}^{t=b}t^{x}\exp\left(-t\right)\mathrm{d}t=\left[-t^{x}\exp\left(-t\right)\right]_{a}^{b}+x\int t^{x-1}\exp\left(-t\right)\mathrm{d}t.$$

Letting $a\rightarrow0$ and $b\rightarrow\infty$ we get the recursive relationship $\Gamma\left(x+1\right)=x\Gamma\left(x\right)$ for $x>0$. Since $\int_{t=0}^{t=\infty}\exp\left(-t\right)\mathrm{d}t=\left[-\exp\left(-t\right)\right]_{0}^{\infty}=1$ we further have

$$\begin{array}{rcccl}\Gamma\left(1\right) & = & 1 & = & 0!\\ \Gamma\left(2\right) & = & 1\cdot\Gamma\left(1\right) & = & 1!\\ \Gamma\left(3\right) & = & 2\cdot\Gamma\left(2\right) & = & 2!\\ \Gamma\left(4\right) & = & 3\cdot\Gamma\left(3\right) & = & 3!\end{array}$$

So that, third, the gamma function is an extension of the factorial function.

The *gamma distribution,* with location parameter α, and scale parameter β, has density

$$f\left(x;\alpha,\beta\right)=\frac{1}{\beta^{\alpha}}\frac{x^{\alpha-1}\exp\left(-x/\beta\right)}{\Gamma\left(\alpha\right)} \tag{2}$$

for $x>0$ and zero otherwise. A useful feature of gamma random variables is that if

$$X_{j}\sim\mathcal{G}\left(\alpha_{j},\beta\right),$$

then

$$\sum_{j=1}^{J}X_{i}\sim\mathcal{G}\left(\sum_{j=1}^{J}\alpha_{j},\beta\right). \tag{3}$$

Chamberlain (2012) calls this a reproductive stable property. When $\alpha=1$ the gamma distribution coincides with an exponential distribution with scale parameter β.

We will model our uncertainty about the true value of θ, by assuming that it is itself random. Specifically we assume that the θ indexing the sampled population corresponds to a random draw from a collection of possible data distributions. The probability measure we place on the parameter space Θ is called the *prior distribution.* A convenient way to assign probabilities to the $(J-1)$ unit simplex is to use the a *Dirichlet* distribution (cf., Ferguson, 1973, 1974;

Chamberlain and Imbens, 2003; Ng, Tian and Tang, 2011). The Dirichlet distribution is a distribution over $\mathbb{S}^{J-1}$ with density

$$\pi\left(\theta_1,\ldots,\theta_J\right)=\frac{\Gamma\left(\sum_{j=1}^{J}\alpha_j\right)}{\prod_{j=1}^{J}\Gamma\left(\alpha_j\right)}\left[\prod_{j=1}^{J}\theta_j^{\alpha_j-1}\right] \tag{4}$$

if $\theta\in\Theta=\mathbb{S}^{J-1}$ and zero otherwise. The parameter indexing the Dirichlet, $\alpha=(\alpha_1,\ldots,\alpha_J)'$, is a vector of strictly positive real numbers. We will discuss how to choose α in practice below.

2 Bayes rule

Let $N_j=\sum_{i=1}^{N}\mathbf{1}\left(Z_i=z_j\right)$ equal the number of units in our sample taking the j^{th} possible value. Conditional on θ, the *likelihood*, of our data is multinomial with density

$$f\left(\mathbf{z}|\,\theta\right)=\frac{N!}{N_1!\cdots N_J!}\prod_{j=1}^{J}\theta_j^{N_j}.$$

The value of $\mathbf{Z}$ in hand reveals information about likely (population) values of θ. The sample allows us to learn, or update our beliefs, about the true population distribution. For example, if draws with $Z=z_j$ are numerous in our sample, then we might sensibly conclude that θ_j is 'likely' to be 'large'. We update our beliefs using the laws of probability, specifically Bayes' rule. We call the conditional distribution of θ given $\mathbf{Z}=\mathbf{z}$, or the distribution representing our beliefs about the plausibility of different values for θ *after* having observed the sample in hand, the *posterior distribution.*

Using Bayes' rule the posterior density is given by

$$\bar{\pi}\left(\theta|\,\mathbf{z}\right) \;=\; \frac{f\left(\mathbf{z}|\,\theta\right)\pi\left(\theta\right)}{\int f\left(\mathbf{z}|\,\theta\right)\pi\left(\theta\right)d\theta},$$

which, given the multinomial likelihood and Dirichlet prior, takes the specific form

$$\bar{\pi}\left(\theta|\,\mathbf{z}\right)=\frac{\prod_{j=1}^{J}\theta_j^{N_j+\alpha_j-1}}{\int_{\mathbb{S}^{J-1}}\prod_{j=1}^{J}\theta_j^{N_j+\alpha_j-1}\mathrm{d}\theta_1\cdots\mathrm{d}\theta_J}.$$

Following calculations similar to those used to derive the characterization of the beta function given by (5) we can show that

$$\int_{\mathbb{S}^{J-1}} \prod_{j=1}^{J} \theta_j^{N_j+\alpha_j-1} \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_J = \frac{\prod_{j=1}^{J} \Gamma\left(N_j+\alpha_j\right)}{\Gamma\left(\sum_{j=1}^{J} N_j+\alpha_j\right)}$$

and consequently that our posterior is itself a member of the Dirichlet family with density

$$\bar{\pi}\left(\theta \middle| \mathbf{z}\right) = \left[\frac{\Gamma\left(\sum_{j=1}^{J} N_j+\alpha_j\right)}{\prod_{j=1}^{J} \Gamma\left(N_j+\alpha_j\right)}\right] \prod_{j=1}^{J} \theta_j^{N_j+\alpha_j-1}.$$

The Dirichlet is what is known as the conjugate prior of the multinomial distribution: when our prior beliefs about the distribution of probability mass across a finite set of support points takes the Dirichlet form, then our posterior beliefs, after observing a random sample, will take the same form.

Let $\bar{\alpha}_j = \alpha_j + N_j$ for $j = 1, \ldots, J$, $\bar{\alpha}_0 = \sum_{j=1}^{J} \bar{\alpha}_j$ and $\bar{\theta}_j = \frac{\alpha_j+N_j}{\sum_{j=1}^{J} \alpha_j+N_j} = \frac{\bar{\alpha}_j}{\bar{\alpha}_0}$. The posterior mean of θ is given by

$$\begin{aligned} \mathbb{E}\left[\theta \middle| \mathbf{Z}=\mathbf{z};\alpha\right] &= \begin{pmatrix} \bar{\theta}_1 \\ \vdots \\ \bar{\theta}_J \end{pmatrix}, \\ &= \bar{\theta}. \end{aligned}$$

while the posterior covariance is

$$\begin{aligned} \mathbb{V}\left(\theta \middle| \mathbf{Z}=\mathbf{z};\alpha\right) &= \frac{1}{1+\bar{\alpha}_0} \begin{pmatrix} \bar{\theta}_1\left(1-\bar{\theta}_1\right) & \cdots & -\bar{\theta}_1\bar{\theta}_J \\ \vdots & \ddots & \vdots \\ -\bar{\theta}_J\bar{\theta}_1 & \cdots & \bar{\theta}_J\left(1-\bar{\theta}_J\right) \end{pmatrix}. \\ &= \frac{1}{1+\bar{\alpha}_0}\left[\operatorname{diag}\left\{\bar{\theta}\right\} - \bar{\theta}\bar{\theta}'\right]. \end{aligned}$$

The sample allows us to sharpen our beliefs about which distribution functions are more likely to characterize the sampled population, but uncertainty remains; the posterior distribution is non-degenerate.

The beta function

The connection between the gamma and factorial functions suggests we may use the latter to generalize the bionomial coefficient $\left(\begin{array}{c} x+y \\ x \end{array}\right) = \frac{(x+y)!}{x!y!}$. Such a generalization is provided by the reciprocal of the *beta function:*

$$\beta(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

There is a representation of the beta function as a definite integral which will frequently prove useful. Using (1) we have

$$\Gamma(x)\Gamma(y) = \int_{t=0}^{t=\infty}\int_{s=0}^{s=\infty} t^{x-1}\exp(-t)\, s^{y-1}\exp(-s)\,\mathrm{d}s\mathrm{d}t.$$

Applying the change-of-variable $s = tu$ to the second integral yields

$$\begin{aligned}\Gamma(x)\Gamma(y) &= \int_{t=0}^{t=\infty}\int_{u=0}^{u=\infty} t^{x-1}\exp(-t)\, t^{y-1}u^{y-1}\exp(-tu)\cdot t\cdot \mathrm{d}u\mathrm{d}t \\ &= \int_{t=0}^{t=\infty}\int_{u=0}^{u=\infty} t^{x+y-1}u^{y-1}\exp(-t(1+u))\cdot \mathrm{d}u\mathrm{d}t.\end{aligned}$$

Next we let $v = t(1+u)$, yielding

$$\begin{aligned}\Gamma(x)\Gamma(y) &= \int_{v=0}^{v=\infty}\int_{u=0}^{u=\infty} \left(\frac{v}{1+u}\right)^{x+y-1} u^{y-1}\exp(-v)\cdot(1+u)\cdot \mathrm{dudv} \\ &= \int_{v=0}^{v=\infty}\int_{u=0}^{u=\infty} \frac{v^{x+y-1}}{(1+u)^{x+y}}u^{y-1}\exp(-v)\cdot \mathrm{dudv} \\ &= \Gamma(x+y)\int_{u=0}^{u=\infty}\frac{u^{y-1}}{(1+u)^{x+y}}\mathrm{d}u.\end{aligned}$$

Finally we apply the change-of-variables $u = \theta/(1-\theta)$ to get

$$\begin{aligned}\Gamma(x)\Gamma(y) &= \Gamma(x+y)\int_{\theta=0}^{\theta=1}\frac{\left(\frac{\theta}{1-\theta}\right)^{y-1}}{\left(1+\frac{\theta}{1-\theta}\right)^{x+y}}\frac{1}{(1-\theta)^2}\mathrm{d}\theta \\ &= \Gamma(x+y)\int_{\theta=0}^{\theta=1}\theta^{y-1}(1-\theta)^{x-1}\mathrm{d}\theta,\end{aligned}$$

which implies that

$$\beta(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} = \int_{\theta=0}^{\theta=1}\theta^{y-1}(1-\theta)^{x-1}\mathrm{d}\theta. \tag{5}$$

3 Posterior simulation

Our posterior for θ summarizes our beliefs, after observing $\mathbf{Z} = \mathbf{z}$, about the relatively plausibility of all possible joint distributions of X and Y. Consequently we can use this distribution to summarize our beliefs about any functional of the data distribution. For example, the posterior mean of the vector of coefficients indexing the best linear predictor of Y given X is

$$\begin{aligned}\bar{\beta} &= \mathbb{E}\left[\beta(\theta)\middle|\,\mathbf{Z}=\mathbf{z};\alpha\right]\\ &= \int_{\mathbb{S}^{J-1}}\left[\sum_{j=1}^{J}\theta_j x_j x_j'\right]^{-1}\times\left[\sum_{j=1}^{J}\theta_j x_j y_j\right]\left[\frac{\Gamma\left(\sum_{j=1}^{J}N_j+\alpha_j\right)}{\prod_{j=1}^{J}\Gamma\left(N_j+\alpha_j\right)}\right]\prod_{j=1}^{J}\theta_j^{N_j+\alpha_j-1}\mathrm{d}\theta_1\cdots\mathrm{d}\theta_J.\end{aligned}$$

In principle we could compute this integral directly. In practice an easier approach is to use simulation. To develop this simulation approach it is useful to use a connection between gamma random variables and Dirichlet ones. Let $\{W_j\}_{j=1}^{J}$ be J independent random variables with $W_j \sim \mathcal{G}(\alpha_j, 1)$. Define

$$V_j = W_j / \sum_{j=1}^{J} W_j.$$

It turns out that $(V_1, \ldots, V_J)$ coincides with a random draw from a Dirichlet distribution with parameter $\alpha = (\alpha_1, \ldots, \alpha_J)'$ (e.g., Ng, Tian and Tang, 2011, p. 40).

Showing this result, which involves an application of the change of variables formula, is instructive. Let $U = \sum_{j=1}^{J} W_j$. This gives $W_j = UV_j$ for $j = 1, \ldots, J-1$ and $W_J = U\left(1 - \sum_{j=1}^{J-1} V_j\right)$. The Jacobian is

$$\begin{pmatrix} V_1 & U & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ V_{J-1} & 0 & \cdots & U \\ 1-\sum_{j=1}^{J-1}V_j & -U & \cdots & -U \end{pmatrix},$$

with determinant U^{J-1}. The joint density of $\mathbf{W} = (W_1, \ldots, W_J)'$ is

$$f(\mathbf{w};\alpha) = \prod_{j=1}^{J} w_j^{\alpha_j - 1}\frac{\exp(-w_j)}{\Gamma(\alpha_j)},$$

so we have a density for $V_1, \ldots, V_{J-1}, U$ of

$$\begin{aligned} f\left(v_1, \ldots, v_{J-1}, u; \alpha\right) &= \prod_{j=1}^{J-1} \left(uv_j\right)^{\alpha_j - 1} \frac{\exp\left(-uv_j\right)}{\Gamma\left(\alpha_j\right)} \\ &\times \left[\left(u\left(1 - \sum_{j=1}^{J-1} v_j\right)\right)^{\alpha_J - 1} \frac{\exp\left(-u\left(1 - \sum_{j=1}^{J-1} v_j\right)\right)}{\Gamma\left(\alpha_J\right)}\right] u^{J-1} \\ &= \frac{1}{\prod_{j=1}^{J} \Gamma\left(\alpha_j\right)} \left[\prod_{j=1}^{J-1} v_j^{\alpha_j - 1}\right] \left[1 - \sum_{j=1}^{J-1} v_j\right]^{\alpha_J - 1} \\ &\times u^{\left(\sum_{j=1}^{J} \alpha_j - 1\right) + N - 1} \exp\left(-u \sum_{j=1}^{J-1} v_j - u\left(1 - \sum_{j=1}^{J-1} v_j\right)\right) \\ &= \frac{1}{\prod_{j=1}^{J} \Gamma\left(\alpha_j\right)} \left[\prod_{j=1}^{J-1} v_j^{\alpha_j - 1}\right] \left[1 - \sum_{j=1}^{J-1} v_j\right]^{\alpha_J - 1} u^{\left(\sum_{j=1}^{J} \alpha_j\right) - 1} \exp\left(-u\right). \end{aligned}$$

Integrating out u yields a marginal distribution for $V_1, \ldots, V_{J-1}$ of

$$\begin{aligned} f\left(v_1, \ldots, v_{J-1}; \alpha\right) &= \int f\left(v_1, \ldots, v_{J-1}, u; \alpha\right) \mathrm{d}u \\ &= \frac{1}{\prod_{j=1}^{J} \Gamma\left(\alpha_j\right)} \left[\prod_{j=1}^{J-1} v_j^{\alpha_j - 1}\right] \left[1 - \sum_{j=1}^{J-1} v_j\right]^{\alpha_J - 1} \int u^{\left(\sum_{j=1}^{J} \alpha_j\right) - 1} \exp\left(-u\right) \mathrm{d}u \\ &= \frac{\Gamma\left(\sum_{j=1}^{J} \alpha_j\right)}{\prod_{j=1}^{J} \Gamma\left(\alpha_j\right)} \left[\prod_{j=1}^{J-1} v_j^{\alpha_j - 1}\right] \left[1 - \sum_{j=1}^{J-1} v_j\right]^{\alpha_J - 1}, \end{aligned}$$

Since $\Gamma\left(\alpha\right) = \int_0^\infty u^{\alpha - 1} \exp\left(u\right) \mathrm{d}u$. This is the density of a Dirichlet distribution with parameter $\alpha = \left(\alpha_1, \ldots, \alpha_J\right)'$ as claimed.

Recalling that $\bar{\alpha}_j = \alpha_j + N_j$ and letting $V_j = W_j / \sum_{j=1}^{J} W_j$ for $W_j \sim \mathcal{G}\left(\bar{\alpha}_j, 1\right)$, both for $j = 1, \ldots, J$, we may represent our posterior distribution for θ as

$$\theta | \, \mathbf{Z} \sim \left(V_1, \ldots, V_J\right)'.$$

Note that

$$V_j = \frac{W_j}{W_j + \sum_{k \neq j} W_k}$$

and hence that the marginal posterior distribution of θ_j is a member of the beta family: $\mathcal{B}\left(\bar{\alpha}_j, \sum_{k \neq j} \bar{\alpha}_k\right)$.

4 Improper prior

We will continue our discussion under the assumption that the prior takes an improper form with $\alpha_j = 0$ for $j = 1, \ldots, J$. Our prior is improper under this parameterization because it does not integrate to one (and hence is not a valid density function). Nevertheless the posterior distribution is proper and we will base inference on it. See Chamberlain and Imbens (2003) and Chamberlain (2012) for a discussion of this choice of prior. Briefly, its two advantages are that it is 'soft' (i.e., generally dominated by the data) and, relatedly, does not place any probability mass on support points not realized in the sample. This latter property simplifies simulation.

To see this second point observe that a random draw from our posterior distribution of probability measures is, for V_j as defined above,

$$F^{\mathrm{BB}}(z) = \sum_{j=1}^{J} V_j \mathbf{1}(z_j \leq z).$$

But under our improper prior $V_j = W_j / \sum_{j=1}^{J} W_j$ with $W_j \sim \mathcal{G}(N_j, 1)$ and N_j equalling the number of observations taking on the j^{th} possible value of Z. Let $W_i^* \sim \mathcal{G}(1,1)$ and $V_i^* = W_i^* / \sum_{i=1}^{N} W_i^*$, an alternative representation of a posterior draw is

$$\begin{aligned}
F_{\mathrm{N}}^{\mathrm{BB}}(z) &= \sum_{i=1}^{N} V_i^* \mathbf{1}(Z_i \leq z) \\
&= \sum_{i=1}^{N} V_i^* \sum_{j=1}^{J} \mathbf{1}(Z_i = z_j) \mathbf{1}(z_j \leq z) \\
&= \sum_{j=1}^{J} \mathbf{1}(N_j > 0) \left(\sum_{i=1}^{N} \mathbf{1}(Z_i = z_j) V_i^* \right) \mathbf{1}(z_j \leq z) \\
&\sim \sum_{j=1}^{J} \mathbf{1}(N_j > 0) V_j \mathbf{1}(z_j \leq z) \\
&= \sum_{j=1}^{J} V_j \mathbf{1}(z_j \leq z), \\
&= F^{\mathrm{BB}}(z),
\end{aligned}$$

since by the reproductive stable property of the gamma distribute $V_j \sim \sum_{i=1}^{N} \mathbf{1}(Z_i = z_j) V_i^*$ (also note that V_j is degenerate at zero if $N_j = 0$ under our improper prior).

Say we wish to simulate the posterior distribution of the coefficient vector indexing the best linear predictor of Y given X. Using our first representation a random draw from this distribution is given by

$$\beta = \left[\sum_{j=1}^{J} V_j x_j x_j'\right]^{-1} \times \left[\sum_{j=1}^{J} V_j x_j y_j\right].$$

Using the second representation, in contrast, a random draw is given by

$$\beta = \left[\sum_{i=1}^{N} V_i^* X_i X_i'\right]^{-1} \times \left[\sum_{i=1}^{N} V_i^* X_i Y_i\right],$$

which corresponds to the weighted least squares fit of $\mathbf{Y}$ onto $\mathbf{X}$ with the vector $\mathbf{V}^* = (V_1^*, \ldots, V_N^*)'$ containing the weights.

This suggests the following approach to inference.

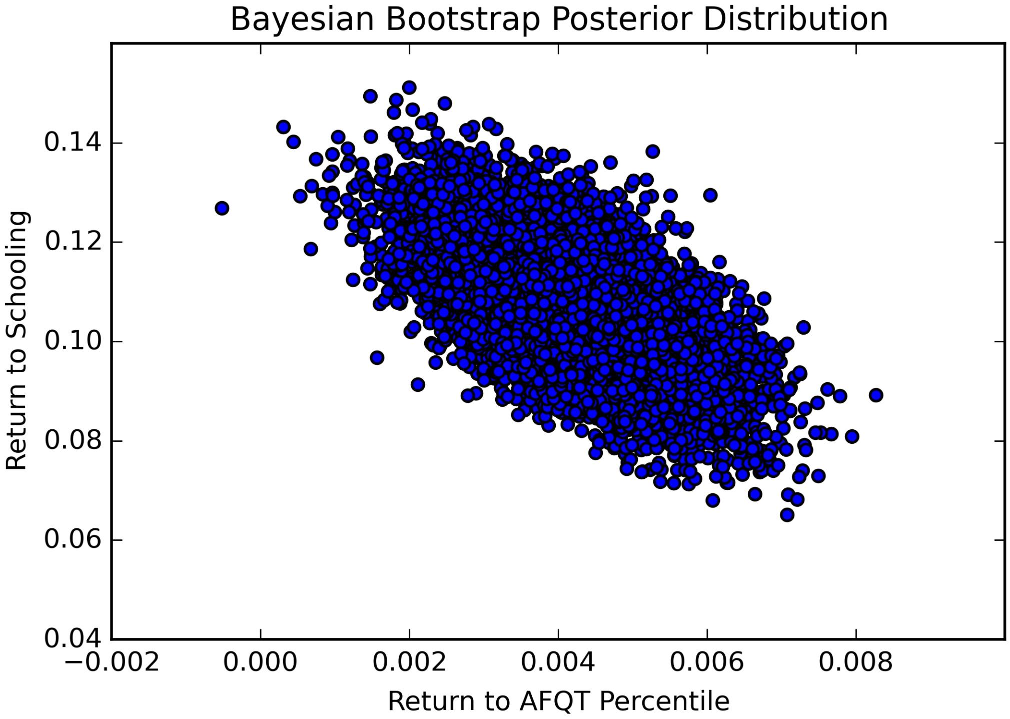
1. Draw an N vector of independent $\mathcal{G}(1,1)$ random variables. Form the sum normalized vector $\mathbf{V}_{(b)}^*$;
2. Compute $\beta_{(b)} = \left[\sum_{i=1}^{N} V_i^* X_i X_i'\right]^{-1} \times \left[\sum_{i=1}^{N} V_i^* X_i Y_i\right]$.

Repeat this process for $b = 1, \ldots, B$ times with B large (>1,000). The posterior probability that an element of β lies in some interval can be approximated by the fraction of its B posterior draws that lie in the same interval.[1] An interval which will contain the true linear predictor coefficient vector with probability 0.95 can be formed by taking the 0.025 and 0.975 sample quantiles of the B posterior draws. This is sometimes called a credible interval in Bayesian statistics.

As an illustration of posterior simulation using the Bayesian Bootstrap consider the coefficients on schooling and AFQT percentile in the linear predictor of log earnings given a constant, schooling and AFQT percentile. The sample is NLSY79 extract of about 2,000 employed white males. The first column reports the OLS estimate of the coefficient on schooling and AFQT (as well as the ratio of these two estimates). Below each coefficient estimate is an heteroscedastic robust standard error estimate and a 95 percent confidence interval. The second column reports posterior means, standard deviations and 95 percent credible intervals for these same three objects. These are approximated using 10,000 "Bayesian Bootstrap" posterior draws. Note that the posterior distribution for the ratio of the schooling-to-AFQT

[1] We will develop methods which formalize the degree to which this approximation becomes more and more accurate as B, the number of posterior draws, increase in later chapters.

Figure 1: Bayes Bootstrap



linear predictor coefficients differs appreciably from its corresponding (estimated) asymptotic sampling distribution.

	OLS	BB
	0.1064	0.1065
YrsSch	(0.012)	(0.012)
	$[0.084, 0.129]$	$[0.084, 0.130]$
	0.0042	0.0042
AQFT	(0.001)	(0.001)
	$[0.002, 0.006]$	$[0.002, 0.006]$
	25.31	27.91
YrsSch/AQFT	(8.36)	(12.9)
	$[8.9, 41.7]$	$[14.2, 56.8]$

Figure 1 plots the 10,000 draws of the two linear predictor coefficients from the posterior distribution.

5 Discussion

A convenient feature of the approach to measuring uncertainty outlined above is that it may be used to make degree of belief statements about complex functionals of the data distribution. The discussion above emphasized linear predictor coefficients, but its application to conditional quantiles and other, more exotic, objects is similarly simple. We will develop this point by means of an example in a later chapter. Rubin (1981) introduced the above simulation method, terming it the 'Bayesian Bootstrap'.

In later chapters we will develop an alternative, large sample, approach to inference. This approach will not require the specification of a prior distribution. It is also conceptually distinct. In many cases the Bayesian Bootstrap and the large sample approach will give similar answers. This has been formalized by demonstrating a large sample equivalence in certain leading cases (for example Efron (1982), Lo (1987) and Hahn (1997)).

The exposition here has made the assumption that Z may take on only a finite number of values. The case of infinite support can also be handled and was first explored by Ferguson (1973, 1974).

References

[1] Chamberlain, Gary. (2012). "Lecture Notes on the Bayesian Bootstrap," *Mimeo.*

[2] Chamberlain, Gary and Guido W. Imbens. (2003). "Nonparametric applications of Bayesian inference," *Journal of Business and Economic Statistics* 21 (1): 12 - 18.

[3] Efron, Bradley. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: Society for Industrial and Applied Mathematics.

[4] Ferguson, Thomas S. (1967). *Mathematical Statistics: A Decision Theoretic Approach.* New York: Academic Press.

[5] Ferguson, Thomas S. (1973). "A Bayesian analysis of some nonparametric problems," *Annals of Statistics* 1 (2): 209 - 230.

[6] Ferguson, Thomas S. (1974). "Prior distributions on spaces of probability measures," *Annals of Statistics* 2 (4): 615 - 629.

[7] Hahn, Jinyong. (1997). "Bayesian bootstrap of the quantile regression estimator: a large sample study," *International Economic Review* 38 (4): 795 - 808.

[8] Johnson, Norman L., Samuel Kotz and N. Balakrishnan. (1995). *Continuous Univariate Distributions 2* New York: Wiley-Interscience.

[9] Lo, Albert Y. (1987). "A large sample study of the Bayesian Bootstrap," *Annals of Statistics* 15 (1): 360 - 375.

[10] Ng, Kai Wang, Guo-Liang Tian, and Man-Lai Tang. (2011). *Dirichlet and Related Distributions: Theory, Methods and Applications.* West Sussex, UK: John Wiley & Sons.

[11] Laplace, Pierre Simon. (1774,1986). "Memoir on the probability of causes of events," *Statistical Science* 1 (3): 364 - 378.

[12] Lehmann, Erich L. and George Casella. (1998). *Theory of Point Estimation.* New York: Springer.

[13] Rubin, Donald B. (1981). "The Bayesian bootstrap," *Annals of Statistics* 9 (1): 130 - 134.

[14] Stigler, Stephen M. (1986a). *The History of Statistics: The Measurement of Uncertainty before 1900.* Cambridge, MA: Harvard University Press.

[15] Stigler, Stephen M. (1986b). "Laplace's 1774 memoir on inverse probability," *Statistical Science* 1 (3): 359 - 363.