7
Repeated-sampling inference

Much of the discussion regarding econometric analysis of endogenous relations centers around identification issues. In this chapter we review the complementary matter of inference. Exchangeability or symmetric dependence and de Finetti's theorem lie at the heart of most (perhaps all) statistical inference. A simple binomial example illustrates. Exchangeability says that a sequence of coin flips has the property

$$\begin{aligned} & Pr\left(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1\right) \\ = \; & Pr\left(X_3 = 1, X_4 = 0, X_2 = 1, X_1 = 1\right) \end{aligned}$$

and so on for all permutations of the random variable index. de Finetti's theorem [1937, reprinted in 1964] provides justification for typical statistical sampling from a population with unknown distribution based on a large number of *iid* draws from the unknown distribution. That is, if ex ante the analyst assesses that samples are exchangeable (and from a large population), then the samples can be viewed as independent and identically distributed from an unknown distribution function. Perhaps it is instructive to consider whether (most) specification issues can be thought of as questions of the validity of some exchangeability conditions. While we ponder this, we review repeated-sampling based inference with particular attention to bootstrapping and Bayesian simulation.[1]

[1] MacKinnon [2002] suggests three fruitful avenues for exploiting abundant computing capacity: (1) structural models at the individual level that frequently draw on simulation, (2) Markov chain Monte Carlo (*McMC*) analysis, and (3) bootstrap inference.

7.1 Monte Carlo simulation

Monte Carlo simulation can be applied when the statistic of interest is pivotal.

Definition 7.1 *A pivotal statistic is one that depends only on the data and no unknown parameters.*

Monte Carlo simulation of pivotal statistics produces exact tests.

Definition 7.2 *Exact tests are tests for which a true null hypothesis is rejected with probability precisely equal to α, the nominal size of the test.*

However, if the test statistic is not pivotal (for instance, the distribution is unknown), a Monte Carlo test doesn't apply.

7.2 Bootstrap

Inference based on bootstrapping is simply an application of the *fundamental theorem of statistics*. That is, when randomly sampled with replacement the empirical distribution function is consistent for the population distribution function (see appendix).

To bootstrap a single parameter such as the correlation between two random variables. say x and y, we simply sample randomly with replacement from the pair (x, y). Then, utilize the empirical distribution of the statistic (say, sample correlation) to draw inferences, for instance, about the mean, etc. (see Efron [1979, 2000]).

7.2.1 Bootstrap regression

For a regression that satisfies standard *OLS* (spherical) conditions, bootstrapping involves first estimating the regression via *OLS* $X_i\hat{\beta}$ and calculating the residuals.[2] The second step involves randomly sampling with replacement a residual for each estimated regression observation $X_i\hat{\beta}$. Pseudo responses $\widehat{Y}$ are constructed by adding the sampled residual to the estimated regression $X_i\hat{\beta}$ for each draw desired (often this is simply n, the original sample size). Next, b_k is estimated via *OLS* regression of $\widehat{Y}$ on the matrix of regressors. Steps two and three are repeated B times to produce an empirical sample of b_k, $k = 1, \ldots .B$. Davidson and MacKinnon [2003] recommend choosing B such that $\alpha (B + 1)$ is an integer where α is the proposed size of the test. Inferences (such as interval estimates) are then based on this empirical sample.

[2] The current and next section draw heavily from Freedman [1981] and Freedman and Peters [1984].

7.2.2 Bootstrap panel data regression

If the errors are heteroskedastic and/or correlated, then the bootstrapping procedure above is modified to accommodate these features. The key is we bootstrap exchangeable partitions of the data. Suppose we have panel data stacked by time series of length T by J cross-sectional individuals in the sample (the sample size is $n = T * J$).

Heteroskedasticity

If we suppose the errors are independent but the variance depends on the cross-sectional unit,

$$\Sigma = \begin{bmatrix} \sigma_1^2 I_T & 0 & \cdots & 0 \\ 0 & \sigma_2^2 I_T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_J^2 I_T \end{bmatrix}$$

then random draws with replacement of the first step residuals (whether estimated by *OLS* or *WLS*, weighted least squares) are taken from the size T sample of residuals for each cross-sectional unit or group of cross-sectional individuals with the same variance. As these partitions are exchangeable, this preserves the differences in variances across cross-sectional units. The remainder of the process remains as described above for bootstrapping regression.

When the nature of the heteroskedasticity is unknown, Freedman [1981] suggests a *paired bootstrap* where $[Y_i, X_i]$ are sampled simultaneously. MacKinnon [2002, p. 629-631] also discusses a *wild bootstrap* to deal with unknown heteroskedasticity.

Correlated errors

If the errors are serially correlated but the variance is constant across cross-sectional units,

$$\Sigma = \begin{bmatrix} V & 0 & \cdots & 0 \\ 0 & V & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V \end{bmatrix}$$

where

$$V = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_T \\ \rho_1 & 1 & \cdots & \rho_{T-1} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_T & \rho_{t-1} & \cdots & 1 \end{bmatrix}$$

then random vector (of length T) draws with replacement of the first step residuals (whether estimated by *OLS* or *GLS*, generalized least squares) are taken from the cross-sectional units.[3] As these partitions are exchangeable, this preserves the serial correlation inherent in the data. The remainder of the process is as described above for bootstrapping regression.[4]

Heteroskedasticity and serial correlation

If the errors are serially correlated and the variance is nonconstant across cross-sectional units,

$$\Sigma = \begin{bmatrix} V_1 & 0 & \cdots & 0 \\ 0 & V_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V_J \end{bmatrix}$$

where

$$V_j = \sigma_j^2 \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_T \\ \rho_1 & 1 & \cdots & \rho_{T-1} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_T & \rho_{t-1} & \cdots & 1 \end{bmatrix}$$

then a combination of the above two sampling procedures is employed.[5] That is, groups of cross-section units with the same variance-covariance structure are identified and random vector (of length T) draws with replacement of the first step residuals (whether estimated by *OLS* or *GLS*) are taken from the groups of cross-sectional units. As these partitions are exchangeable, this preserves the heteroskedasticity and serial correlation inherent in the data. The remainder of the process is as described above for bootstrapping regression.

[3] For cross-sectional correlation (but independent errors through time)

$$\Sigma = \sigma^2 \begin{bmatrix} I_T & \rho_{12} I_T & \cdots & \rho_{1J} I_T \\ \rho_{12} I_T & I_T & \cdots & \rho_{2J} I_T \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1J} I_T & \rho_{2J} I_T & \cdots & I_T \end{bmatrix}$$

simply apply the same ideas to the length J vector of residuals over cross-sectional units in place of the length T vector of residuals through time.

[4] When the nature of the serial correlation is unknown, as expected the challenge is greater. MacKinnon [2002] discusses two approaches: *sieve bootstrap* and *block bootstrap*. Not surprisingly, when the nature of the correlation or heteroskedasticity is unknown the bootstrap performs more poorly than otherwise.

[5] Cross-sectional correlation and heteroskedasticity

$$\Sigma = \begin{bmatrix} \sigma_1^2 I_T & \rho_{12}\sigma_1\sigma_2 I_T & \cdots & \rho_{1J}\sigma_1\sigma_J I_T \\ \rho_{12}\sigma_1\sigma_2 I_T & \sigma_2^2 I_T & \cdots & \rho_{2J}\sigma_2\sigma_J I_T \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1J}\sigma_1\sigma_J I_T & \rho_{2J}\sigma_2\sigma_J I_T & \cdots & \sigma_J^2 I_T \end{bmatrix}$$

again calls for sampling from like variance-covariance groups.

7.2.3 Bootstrap summary

Horowitz [2001] relates the bootstrap to asymptotically pivotal statistics in discussing effective usage of the bootstrap.

Definition 7.3 *An asymptotically pivotal statistic is a statistic whose asymptotic distribution does not depend on unknown population parameters.*

Horowitz concludes

- If an asymptotically pivotal statistic is available, use the bootstrap to estimate the probability distribution of the asymptotically pivotal statistic or a critical test value based on the asymptotically pivotal statistic.
- Use an asymptotically pivotal statistic if available rather than bootstrapping a non-asymptotically pivotal statistic such as a regression slope coefficient or standard error to estimate the probability distribution of the statistic.
- Recenter the residuals of an overidentified model before applying the bootstrap.
- Extra care is called for when bootstrapping models for dependent data, semi- or non-parametric estimators, or non-smooth estimators.

7.3 Bayesian simulation

Like bootstrapping, Bayesian simulation employs repeated sampling with replacement to draw inferences. Bayesian sampling in its simplest form utilizes Bayes' theorem to identify the posterior distribution of interest $p(\theta \mid Y)$ from the likelihood function $p(Y \mid \theta)$ and prior distribution for the parameters of interest $p(\theta)$.

$$p(\theta \mid Y) = \frac{p(Y \mid \theta)\, p(\theta)}{p(Y)}$$

The marginal distribution of the data $p(Y)$ is a normalizing adjustment. Since it does not affect the kernel of the distribution it is typically suppressed and the posterior is written

$$p(\theta \mid Y) \propto p(Y \mid \theta)\, p(\theta)$$

7.3.1 Conjugate families

It is straightforward to sample from the posterior distribution when its kernel (the portion of the density function or probability mass function that depends on the parameters of interest) is readily recognized. For a number of prior distributions (and likelihood functions), the posterior distribution is readily recognized as a standard distribution. This is referred to as *conjugacy* and the matching prior distribution is called the *conjugate prior*. A formal definition follows.

Definition 7.4 *If $\mathcal{F}$ is a class of sampling distributions $p(Y \mid \theta)$ and $\wp$ is a class of prior distributions for θ, then class $\wp$ is conjugate to $\mathcal{F}$ class if $p(\theta \mid Y) \in \wp$ for all $p(\cdot \mid \theta) \in \mathcal{F}$ and $p(\cdot) \in \wp$.*

For example, a binomial likelihood

$$\ell(\theta \mid s; n) = \begin{pmatrix} n \\ s \end{pmatrix} \theta^s (1-\theta)^{n-s}$$

$$s = \textstyle\sum_{i=1}^n y_i, \quad y_i = \{0, 1\}$$

combines with a beta$(\theta; \alpha, \beta)$ prior

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

to yield

$$\begin{aligned} p(\theta \mid y) &\propto \theta^s (1-\theta)^{n-s} \theta^{\alpha-1} (1-\theta)^{\beta-1} \\ &= \theta^{s+\alpha-1} (1-\theta)^{n-s+\beta-1} \end{aligned}$$

which is the kernel of a beta$(\theta \mid y; \alpha+s, \beta+n-s)$ distribution.

Also, a single draw from a Gaussian likelihood with known standard deviation, σ

$$\ell(\theta \mid y, \sigma) \propto \exp\left[-\frac{1}{2}\frac{(y-\theta)^2}{\sigma^2}\right]$$

combines with a Gaussian or normal prior

$$p(\theta \mid \mu_0, \tau_0) \propto \exp\left[-\frac{1}{2}\frac{(\theta-\mu_0)^2}{\tau_0^2}\right]$$

to yield[6]

$$p(\theta \mid y, \sigma, \mu_0, \tau_0) \propto \exp\left[-\frac{1}{2}\frac{(\theta-\mu_1)^2}{\tau_1^2}\right]$$

where $\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2}+\frac{1}{\sigma^2}}$ and $\tau_1^2 = \frac{1}{\frac{1}{\tau_0^2}+\frac{1}{\sigma^2}}$. The posterior distribution of the mean given the data and priors is Gaussian. And, for a sample of n exchangeable draws, the likelihood is

$$\ell(\theta \mid y, \sigma) \propto \prod_{i=1}^n \exp\left[-\frac{1}{2}\frac{(y_i-\theta)^2}{\sigma^2}\right]$$

[6] The product gives

$$\exp\left[-\frac{1}{2}\left(\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right)\right]$$

Then, expand the exponent and complete the square. Any constants are ignored in the identification of the kernel as they're absorbed through normalization of the posterior kernel.

combined with the above prior yields

$$p\left(\theta \mid y, \sigma, \mu_0, \tau_0\right) \propto \exp\left[-\frac{1}{2}\frac{\left(\theta - \mu_n\right)^2}{\tau_n^2}\right]$$

where $\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\overline{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$, $\overline{y}$ is the sample mean, and $\tau_1^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$. The posterior distribution of the mean given the data and priors is again Gaussian.

These and some other well-known and widely used conjugate family distributions are summarized in tables 7.1, 7.2, 7.3, and 7.4 (see Bernardo and Smith [1994] and Gelman et al [2003]).

Table 7.1: Conjugate families for univariate discrete distributions

likelihood $p\left(Y \mid \theta\right)$	**conjugate prior** $p\left(\theta\right)$	**posterior** $p\left(\theta \mid Y\right)$
Binomial $\left(s \mid n, \theta\right)$ where $s = \sum_{i=1}^{n} y_i, y_i \in \{0, 1\}$	Beta $\left(\theta; \alpha, \beta\right)$ $\propto \theta^{\alpha-1}\left(1-\theta\right)^{\beta-1}$	Beta $\left(\theta \mid \alpha + s, \beta + n - s\right)$
Poisson $\left(s \mid n\lambda\right)$ where $s = \sum_{i=1}^{n} y_i, y_i = 0, 1, 2, \ldots$	Gamma $\left(\theta; \alpha, \beta\right)$ $\propto \theta^{\alpha-1}e^{-\beta\theta}$	Gamma $\left(\theta \mid \alpha + s, \beta + n\right)$
Exponential $\left(t \mid n, \theta\right)$ where $t = \sum_{i=1}^{n} y_i, y_i = 0, 1, 2, \ldots$	Gamma $\left(\theta; \alpha, \beta\right)$ $\propto \theta^{\alpha-1}e^{-\beta\theta}$	Gamma $\left(\theta \mid \alpha + n, \beta + t\right)$
Negative-binomial $\left(s \mid \theta, nr\right)$ where $s = \sum_{i=1}^{n} y_i, y_i = 0, 1, 2, \ldots$	Beta $\left(\theta; \alpha, \beta\right)$ $\propto \theta^{\alpha-1}\left(1-\theta\right)^{\beta-1}$	Beta $\left(\theta \mid \alpha + nr, \beta + s\right)$
Beta and gamma are continuous distributions		

A few words regarding the multi-parameter Gaussian case with unknown mean and variance seem appropriate. The joint prior combines a Gaussian prior for the mean conditional on the variance and an inverse-gamma or inverse-chi square prior for the variance.[7] The joint posterior distribution is the same form as the prior

[7] The inverse-gamma(α, β) distribution

$$p\left(\sigma^2; \alpha, \beta\right) \propto \left(\sigma^2\right)^{-(\alpha+1)} \exp\left[-\frac{\beta}{\sigma^2}\right]$$

Table 7.2: Conjugate families for univariate continuous distributions

likelihood $p(Y \mid \theta)$	**conjugate prior** $p(\theta)$	**marginal posterior** $p(\theta \mid Y)$
Uniform $(Y_i \mid 0, \theta)$ where $0 < Y_i < \theta$, $t = \max\{Y_1, \ldots, Y_n\}$	Pareto $(\theta; \alpha, \beta)$ $\propto \theta^{-(\alpha+1)}$	Pareto $(\theta; \alpha + n, \max\{\beta, t\})$
Normal $\left(Y \mid \theta, \sigma^2\right)$ variance known	Normal $\left(\theta \mid \sigma^2; \theta_0, \tau_0^2\right)$ $\propto \tau_0^{-1} e^{-\frac{(y-\theta_0)^2}{2\tau_0^2}}$	$Normal$ $\left(\mu \mid \sigma^2; \begin{matrix} \frac{\frac{\theta_0}{\tau_0^2} + \frac{n\overline{Y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \\ \frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \end{matrix}\right)$
Normal $(Y \mid \mu, \theta)$ mean known, $\sigma^2 = \theta$	Inverse- gamma $(\theta; \alpha, \beta)$ $\propto \theta^{-(\alpha+1)} e^{-\beta/\theta}$	Inverse-gamma $\left(\theta; \frac{n+2\alpha}{2}, \beta + \frac{1}{2}t\right)$ where $t = \sum_{i=1}^{n} (Y_i - \mu)^2$
Normal $\left(Y \mid \theta, \sigma^2\right)$ both unknown	Normal$\left(\theta \mid \sigma^2; \theta_0, n_0\right)$ $*Inverse-$ $gamma\left(\sigma^2; \alpha, \beta\right)$	Student t $(\theta; \theta_n, \gamma, 2\alpha + n)$; Inverse-gamma $\left(\sigma^2; \alpha + \frac{1}{2}n, \beta_n\right)$
For the normal-inverse gamma posterior the parameters are $\theta_n = (n_0 + n)^{-1}\left(n_0\theta_0 + n\overline{Y}\right)$ $\gamma = (n + n_0)\left(\alpha + \frac{1}{2}n\right)\beta_n^{-1}$ $\beta_n = \beta + \frac{1}{2}(n-1)s^2 + \frac{1}{2}(n_0+n)^{-1} n_0 n \left(\theta_0 - \overline{Y}\right)^2$ $s^2 = (n-1)^{-1} \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$		

— Gaussian$\left(\theta \mid \sigma^2; \theta_n, \sigma_n^2\right)$ ×inverse-gamma$\left(\sigma^2 \mid \alpha + \frac{1}{2}n, \beta_n\right)$. Hence, the conditional distribution for the mean given the variance is Gaussian$\left(\theta \mid \sigma^2; \theta_n, \sigma_n^2\right)$ where $\sigma_n^2 = \frac{\sigma^2}{n_0+n}$. On integrating out the variance from the joint posterior the marginal posterior for the mean is noncentral, scaled Student t$(\theta \mid \theta_n, \gamma, \nu)$ distributed.

A scaled Student t$\left(X \mid \mu, \lambda = \frac{1}{\sigma^2}, \nu\right)$ is symmetric with mean μ, variance $\frac{1}{\lambda}\frac{\nu}{\nu-2}$ $= \sigma^2 \frac{\nu}{\nu-2}$, ν degrees of freedom, and the density function kernel is

$$\left[1 + \nu^{-1}\lambda (X - \mu)^2\right]^{-(\nu+1)/2} = \left[1 + \nu^{-1}\left(\frac{X - \mu}{\sigma}\right)^2\right]^{-(\nu+1)/2}$$

can be reparameterized as an inverse-χ^2 distribution$\left(\nu, \sigma_0^2\right)$

$$p\left(\sigma^2; \nu, \sigma_0^2\right) \propto \left(\sigma^2\right)^{-(\nu/2+1)} \exp\left[-\frac{\nu\sigma_0^2}{2\sigma^2}\right]$$

(see Gelman et al [2003], p. 50). Hence, $\alpha = \frac{\nu}{2}$ or $\nu = 2\alpha$ and $\beta = \frac{\nu\sigma_0^2}{2}$ or $\nu\sigma_0^2 = 2\beta$.

Hence, the *standard* t distribution is Student t$(Z \mid 0, 1, \nu)$ where $Z = \frac{X-\mu}{\sigma}$. Marginalization of the mean follows Gelman et al [2003] p. 76. For uninformative priors, $p\left(\theta, \sigma^2\right) \propto \sigma^{-2}$

$$\begin{aligned} p\left(\theta \mid y\right) &= \int_0^\infty p\left(\theta, \sigma^2 \mid y\right) d\sigma^2 \\ &= \int_0^\infty \sigma^{-n-2} \exp\left[-\frac{A}{2\sigma^2}\right] d\sigma^2 \end{aligned}$$

where $A = \left(n-1\right)s^2 + n\left(\theta - \overline{y}\right)^2$. Let $z = \frac{A}{2\sigma^2}$, then transformation of variables yields

$$p\left(\theta \mid y\right) \propto A^{-n/2} \int_0^\infty z^{(n-2)/2} \exp\left[-z\right] dz$$

Since the integral involves the kernel for a gamma, it integrates to a constant and can be ignored for identifying the marginal posterior kernel. Hence, we recognize

$$\begin{aligned} p\left(\theta \mid y\right) &\propto A^{-n/2} = \left[\left(n-1\right)s^2 + n\left(\theta - \overline{y}\right)^2\right]^{-\frac{n}{2}} \\ &\propto \left[1 + \frac{n\left(\theta - \overline{y}\right)^2}{\left(n-1\right)s^2}\right]^{-\frac{n}{2}} \end{aligned}$$

is the kernel for a noncentral, scaled Student t$\left(\theta; \overline{y}, \frac{s^2}{n}, n-1\right)$. Marginalization with informed conjugate priors works in analogous fashion.

Table 7.3: Conjugate families for multivariate discrete distributions

likelihood $p\left(Y \mid \theta\right)$	**conjugate prior** $p\left(\theta\right)$	**posterior** $p\left(\theta \mid Y\right)$
Multinomial$_k$ $(r; \theta, n)$ where $r_i = 0, 1, 2, \ldots$	Dirichlet$_k$ $(\theta; \alpha)$ where $\alpha = \{\alpha_1, \ldots, \alpha_{k+1}\}$	$Dirichlet_k$ $\left(\theta; \begin{array}{c} \alpha_1 + r_1, \ldots, \\ \alpha_{k+1} + r_{k+1} \end{array}\right)$

The Dirichlet distribution is a multivariate analog to the beta distribution and has continuous support where $r_{k+1} = n - \sum_{\ell=1}^{k} r_\ell$. Ferguson [1973] proposed the Dirichlet process as a Bayesian nonparametric approach. Some properties of the Dirichlet distribution include

$$E\left[\theta_i \mid \alpha\right] = \frac{\alpha_i}{\alpha_0}$$

$$Var\left[\theta_i \mid \alpha\right] = \frac{\alpha_i\left(\alpha_0 - \alpha_i\right)}{\alpha_0^2\left(\alpha_0 + 1\right)}$$

$$Cov\left[\theta_i, \theta_j \mid \alpha\right] = \frac{-\alpha_i \alpha_j}{\alpha_0^2 \left(\alpha_0 + 1\right)}$$

where $\alpha_0 = \sum_{i=1}^{k+1} \alpha_i$

Table 7.4: Conjugate families for multivariate continuous distributions

likelihood $p\left(Y \mid \theta\right)$	**conjugate prior** $p\left(\theta\right)$	**marginal posterior** $p\left(\theta \mid Y\right)$
Normal $\left(Y \mid \theta, \Sigma\right)$ parameters unknown	Normal$\left(\theta \mid \Sigma; \theta_0, n_0\right)$ $*$Inverse-Wishart $\left(\Sigma; \alpha, \beta\right)$	Student t$_k$ $\left(\theta; \theta_n, \Gamma, 2\alpha_n\right)$; Inverse-Wishart $\left(\Sigma; \alpha + \frac{1}{2}n, \beta_n\right)$
Linear regression Normal$\left(Y \mid X\theta, \sigma^2\right)$ parameters unknown	Normal$\left(\theta \mid \sigma^2; \theta_0, n_0^{-1}\sigma^2\right)$ $*$Inverse-gamma $\left(\sigma^2; \alpha, \beta\right)$	Student t$_k$ $\left(\theta; \theta_n, \Gamma, 2\alpha + n\right)$; Inverse-gamma $\left(\sigma^2; \alpha + \frac{1}{2}n, \beta_n\right)$

The multivariate Student t$_k$ $\left(X \mid \mu, \Gamma, \nu\right)$ is analogous to the univariate Student t$(X \mid \mu, \gamma, \nu)$ as it is symmetric with mean vector (length k) μ, $k \times k$ symmetric, positive definite variance matrix $\Gamma^{-1}\frac{\nu}{\nu-2}$, and ν degrees of freedom. For the Student t and inverse-Wishart marginal posteriors associated with multivariate normal likelihood function, the parameters are

$$\theta_n = \left(n_0 + n\right)^{-1}\left(n_0\theta_0 + n\overline{Y}\right)$$

$$\Gamma = \left(n + n_0\right)\alpha_n \beta_n^{-1}$$

$$\beta_n = \beta + \frac{1}{2}S + \frac{1}{2}\left(n_0 + n\right)^{-1} n_0 n \left(\theta_0 - \overline{Y}\right)\left(\theta_0 - \overline{Y}\right)^T$$

$$S = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)\left(Y_i - \overline{Y}\right)^T$$

$$\alpha_n = \alpha + \frac{1}{2}n - \frac{1}{2}\left(k - 1\right)$$

For the Student t and inverse-gamma marginal posteriors associated with linear regression, the parameters are[8]

$$\theta_n = \left(n_0 + X^T X\right)^{-1}\left(n_0\theta_0 + X^T Y\right)$$

$$n_0 = X_0^T X_0$$

$$\Gamma = \left(n_0 + X^T X\right)\left(\alpha + \frac{1}{2}n\right)\beta_n^{-1}$$

[8] Notice, linear regression subsumes the univariate, multi-parameter Gaussian case. If we let $X = \iota$ (a vector of ones), then linear regression becomes the univariate Gaussian case.

$$\beta_n = \beta + \frac{1}{2}\left(Y - X\theta_n\right)^T Y + \frac{1}{2}\left(\theta_0 - \theta_n\right)^T n_0\theta_0$$

Bayesian regression with conjugate priors works as if we have data from a prior period $\{Y_0, X_0\}$ and the current period $\{Y, X\}$ from which to estimate θ_n. Applying *OLS* to the stack of equations $\begin{bmatrix} Y_0 \\ Y \end{bmatrix} = \begin{bmatrix} X_0 \\ X \end{bmatrix} \theta_n + \begin{bmatrix} \varepsilon_0 \\ \varepsilon \end{bmatrix}$ yields[9]

$$\begin{aligned} \theta_n &= \left(X_0^T X_0 + X^T X\right)^{-1}\left(X_0^T Y_0 + X^T Y\right) \\ &= \left(n_0 + X^T X\right)^{-1}\left(n_0\theta_0 + X^T Y\right) \end{aligned}$$

The inverse-Wishart and multivariate Student t distributions are multivariate analogs to the inverse-gamma and (noncentral, scaled) univariate Student t distributions, respectively.

7.3.2 McMC simulations

Markov chain Monte Carlo (*McMC*) simulations are employed when the marginal posterior distributions cannot be derived or are extremely cumbersome to derive. *McMC* approaches draw from the set of conditional posterior distributions instead of the marginal posterior distributions. The Hammersley-Clifford theorem (Hammersley and Clifford [1971] and Besag [1974]) provides regulatory conditions

[9] This perspective of Bayesian regression is consistent with recursive least squares where the previous estimate θ_{t-1} based on data $\{Y_{t-1}, X_{t-1}\}$ is updated for data $\{Y_t, X_t\}$ as $\theta_t = \theta_{t-1} + \Im_t^{-1} X_t^T \left(Y_t - X_t\theta_{t-1}\right)$, where $\theta_{t-1} = \left(X_{t-1}^T X_{t-1}\right)^{-1} X_{t-1}^T Y_{t-1}$ and the information matrix is updated as $\Im_t = \Im_{t-1} + X_t^T X_t$. To see this, note

$$\theta_t = \Im_t^{-1} X_t^T Y_t + \left(I - \Im_t^{-1} X_t^T X_t\right)\theta_{t-1}$$

but

$$\begin{aligned} \left(I - \Im_t^{-1} X_t^T X_t\right)\theta_{t-1} &= \Im_t^{-1}\left(\Im_t^{-1} - X_t^T X_t\right)\theta_{t-1} \\ & \quad \Im_t^{-1}\left(X_{t-1}^T X_{t-1} + X_t^T X_t - X_t^T X_t\right)\theta_{t-1} \\ &= \Im_t^{-1} X_{t-1}^T X_{t-1}\theta_{t-1} \\ &= \Im_t^{-1} X_{t-1}^T Y_{t-1} \end{aligned}$$

since $\theta_{t-1} = \left(X_{t-1}^T X_{t-1}\right)^{-1} X_{t-1}^T Y_{t-1}$. Hence,

$$\begin{aligned} \theta_t &= \Im_t^{-1} X_t^T Y_t + \Im_t^{-1} X_{t-1}^T Y_{t-1} \\ &= \Im_t^{-1}\left(X_t^T Y_t + X_{t-1}^T Y_{t-1}\right) \\ &= \left(X_{t-1}^T X_{t-1} + X_t^T X_t\right)^{-1}\left(X_{t-1}^T Y_{t-1} + X_t^T Y_t\right) \end{aligned}$$

or, in the notation above

$$\theta = \left(X_0^T X_0 + X^T X\right)^{-1}\left(X_0^T Y_0 + X^T Y\right)$$

as indicated above.

for when a set of conditional distributions characterizes a unique joint distribution. The regulatory conditions are essentially that every point in the marginal and conditional distributions have positive mass. Common *McMC* approaches (Gibbs sampler and Metropolis-Hastings algorithm) are supported by the Hammersley-Clifford theorem. The utility of *McMC* simulation has evolved along with the **R** Foundation for Statistical Computing.

Gibbs sampler

Suppose we cannot derive $p(\theta \mid Y)$ in closed form (it does not have a standard probability distribution) but we can identify the conditional posterior distributions. We can utilize the full conditional posterior distributions to draw dependent samples for parameters of interest via *McMC* simulation.

For full conditional posterior distributions

$$\begin{gathered} p(\theta_1 \mid \theta_{-1}, Y) \\ \vdots \\ p(\theta_k \mid \theta_{-k}, Y) \end{gathered}$$

draws are made for θ_1 conditional on starting values for parameters other than θ_1, that is θ_{-1}. Then, θ_2 is drawn conditional on the θ_1 draw and the starting value for the remaining θ. Next, θ_3 is drawn conditional on the draws for θ_1 and θ_2 and the remaining θ. This continues until all θ have been sampled. Then the sampling is repeated for a large number of draws with parameters updated each iteration by the most recent draw.

The samples are dependent. Not all samples will be from the posterior; only after a finite (but unknown) number of iterations are draws from the marginal posterior distribution (see Gelfand and Smith [1990]). (Note, in general, $p(\theta_1, \theta_2 \mid Y) \neq p(\theta_1 \mid \theta_2, Y)\, p(\theta_1 \mid \theta_2, Y)$.) Convergence is usually checked using trace plots, burn-in iterations, and other convergence diagnostics. Model specification includes convergence checks, sensitivity to starting values and possibly prior distribution and likelihood assignments, comparison of draws from the posterior predictive distribution with the observed sample, and various goodness of fit statistics.

Albert and Chib's Gibbs sampler Bayes' probit

The challenge with discrete choice models (like probit) is that latent utility is unobservable, rather the analyst observes only discrete (usually binary) choices (see chapter 5). Albert & Chib [1993] employ Bayesian data augmentation to "supply" the latent variable. Hence, parameters of a probit model are estimated via normal Bayesian regression (see earlier discussion in this chapter). Consider the latent utility model

$$U_D = W\theta - V$$

The conditional posterior distribution for θ is

$$p(\theta | D, W, U_D) \sim N\left(b_1, \left(Q^{-1} + W^T W\right)^{-1}\right)$$

where

$$b_1 = \left(Q^{-1} + W^T W\right)^{-1} \left(Q^{-1} b_0 + W^T W b\right)$$
$$b = \left(W^T W\right)^{-1} W^T U_D$$

$b_0 =$ prior means for θ and $Q = \left(W_0^T W_0\right)^{-1}$ is the prior for the covariance. The conditional posterior distribution for the latent variables are

$$p\left(U_D | D = 1, W, \theta\right) \sim N\left(W\theta, I | U_D > 0\right) \text{ or } TN_{(0,\infty)}\left(W\theta, I\right)$$
$$p\left(U_D | D = 0, W, \theta\right) \sim N\left(W\theta, I | U_D \leq 0\right) \text{ or } TN_{(-\infty,0)}\left(W\theta, I\right)$$

where $TN\left(\cdot\right)$ refers to random draws from a truncated normal (truncated below for the first and truncated above for the second). Iterative draws for $\left(U_D | D, W, \theta\right)$ and $\left(\theta | D, W, U_D\right)$ form the Gibbs sampler. Interval estimates of θ are supplied by post-convergence draws of $\left(\theta | D, W, U_D\right)$. For simulated normal draws of the unobservable portion of utility, V, this Bayes' augmented data probit produces remarkably similar inferences to *MLE*.[10]

Metropolis-Hastings algorithm

If neither some conditional posterior, $p\left(\theta_j \mid Y, \theta_{-j}\right)$, or the marginal posterior, $p\left(\theta \mid Y\right)$, is recognizable, then we can employ the Metropolis-Hastings (*MH*) algorithm. The Gibbs sampler is a special case of the *MH* algorithm. The random walk Metropolis algorithm is most common and outlined next.

The random walk Metropolis algorithm is as follows. We wish to draw from $p\left(\theta \mid \cdot\right)$ but we only know $p\left(\theta \mid \cdot\right)$ up to constant of proportionality, $p\left(\theta \mid \cdot\right) = cf\left(\theta \mid \cdot\right)$ where c is unknown.

- Let $\theta^{(k-1)}$ be a draw from $p\left(\theta \mid \cdot\right)$.[11]
- Draw θ^* from $N\left(\theta^{(k-1)}, s^2\right)$ where s^2 is fixed.

[10] An efficient algorithm for this Gibbs sampler probit, rbprobitGibbs, is available in the bayesm package of R (http://www.r-project.org/), the open source statistical computing project. Bayesm is a package written to complement Rossi, Allenby, and McCulloch [2005].

[11] The procedure describes the algorithm for a single parameter. A general K parameter algorithm works similarly (see Train [2002], p. 305):

(a) Start with a value β_n^0.

(b) Draw K independent values from a standard normal density, and stack the draws into a vector labeled η^1.

(c) Create a trial value of $\beta_n^1 = \beta_n^0 + \sigma\Gamma\eta^1$ where σ is the researcher-chosen jump size parameter, Γ is the Cholesky factor of W such that $\Gamma\Gamma^T = W$. Note the proposal distribution is specified to be normal with zero mean and variance $\sigma^2 W$.

(d) Draw a standard uniform variable μ_1.

(e) Calculate the ratio $F = \frac{L\left(y_n | \beta_n^1\right)\phi\left(\beta_n^1 | b, W\right)}{L\left(y_n | \beta_n^0\right)\phi\left(\beta_n^0 | b, W\right)}$ where $L\left(y_n \mid \beta_n^1\right)$ is a product of logits, and $\phi\left(\beta_n^1 \mid b, W\right)$ is the normal density.

(f) If $\mu_1 \leq F$, accept β_n^1; if $\mu_1 > F$, reject β_n^1 and let $\beta_n^1 = \beta_n^0$.

(g) Repeat the process many times. For sufficiently large t, β_n^t is a draw from the marginal posterior.

- Let $\alpha = min\left\{1, \frac{p(\theta^*|\cdot)}{p\left(\theta^{(k-1)}|\cdot\right)} = \frac{cf(\theta^*|\cdot)}{cf\left(\theta^{(k-1)}|\cdot\right)}\right\}$.
- Draw z^* from $U\left(0,1\right)$.
- If $z^* < \alpha$ then $\theta^{(k)} = \theta^*$, otherwise $\theta^{(k)} = \theta^{(k-1)}$. In other words, with probability α set $\theta^{(k)} = \theta^*$, and otherwise set $\theta^{(k)} = \theta^{(k-1)}$.[12]

These draws converge to random draws from the marginal posterior distribution after a burn-in interval if properly tuned.

Tuning the Metropolis algorithm involves selecting s^2 (jump size) so that the parameter space is explored appropriately (see Halton sequences discussion below). Usually, smaller jump size results in more accepts and larger jump size results in fewer accepts. If s^2 is too small, the Markov chain will not converge quickly, has more serial correlation in the draws, and may get stuck at a local mode (multi-modality can be a problem). If s^2 is too large, the Markov chain will move around too much and not be able to thoroughly explore areas of high posterior probability. Of course, we desire concentrated samples from the posterior distribution. A commonly-employed rule of thumb is to target an acceptance rate for θ^* around 30% $(20 - 80\%$ is usually considered "reasonable").[13]

Some other *McMC* methods

Other acceptance sampling procedures such as WinBUGs (see Spiegelhalter, et al. [2003]) are self-tuned. That is, the algorithm adaptively tunes the jump size in generating random post convergence joint posterior draws. A difficulty with WinBUGs is that it can mysteriously crash with little diagnostic aid.

Halton sequences

Random sampling can be slow to provide good coverage and hence prove to be a costly way to simulate data. An alternative that provides better coverage with fewer draws involves *Halton sequences* (see Train [2002], ch. 9, p. 224-238). Unlike other methods discussed above, Halton draws tend to be negatively correlated. Importantly, Bhat [2001] finds that 100 Halton draws provided lower simulation error for his mixed logit than $1,000$ random draws, for discrete choice models. Further, the error rate with 125 Halton draws was half as large as with $1,000$ random draws and somewhat smaller than with $2,000$ random draws.

A Halton sequence builds around a pre-determined number k (usually a prime number). The Halton sequence is

$$s_{t+1} = \left\{s_t, s_t + \frac{1}{k^t}, s_t + \frac{2}{k^t}, \ldots, s_t + \frac{k-1}{k^t}\right\}$$

[12] A modification of the *RW* Metropolis algorithm sets $\theta^{(k)} = \theta^*$ with $log(\alpha)$ probability where $\alpha = min\{0, log[f(\theta^*|\cdot)] - log[f(\theta^{(k-1)}|\cdot)]\}$.

[13] Gelman, et al [2004] report the optimal acceptance rate is 0.44 when the number of parameters $K = 1$ and drops toward 0.23 as K increases.

starting with $s_0 = 0$ (even though zero is ignored). An example helps to fix ideas.

Example 7.1 *Consider the prime $k = 3$. The sequence through two iterations is*

$$\left\{ \begin{array}{c} 0 + 1/3 = 1/3, 0 + 2/3 = 2/3, \\ 0 + 1/9 = 1/9, 1/3 + 1/9 = 4/9, 2/3 + 1/9 = 7/9, \\ 0 + 2/9 = 2/9, 1/3 + 2/9 = 5/9, 2/3 + 2/9 = 8/9, \ldots \end{array} \right\}$$

This procedure describes uniform Halton draws. Other distributions are accommodated in the usual way — by inverse distribution functions.

Example 7.2 *For example, normal draws are found by $\Phi^{-1}(s_t)$. Continuing with the above Halton sequence, standard normal draws are*

$$\left\{ \begin{array}{c} \Phi^{-1}(1/3) \approx -0.43, \Phi^{-1}(2/3) \approx 0.43, \\ \Phi^{-1}(1/9) \approx -1.22, \Phi^{-1}(4/9) \approx -0.14, \Phi^{-1}(7/9) \approx 0.76, \\ \Phi^{-1}(2/9) \approx -0.76, \Phi^{-1}(5/9) \approx 0.14, \Phi^{-1}(8/9) \approx 1.22, \ldots \end{array} \right\}$$

Example 7.3 *For two independent standard normal unobservables we create Halton sequences for each from different primes and transform. Suppose we use $k = 2$ and $k = 3$. The first few draws are*

$$\left\{ \begin{array}{c} \varepsilon_1 = \left\langle \Phi^{-1}\left(\frac{1}{2}\right) = 0, \Phi^{-1}\left(\frac{1}{3}\right) = -0.43 \right\rangle, \\ \varepsilon_2 = \left\langle \Phi^{-1}\left(\frac{1}{4}\right) = -.67, \Phi^{-1}\left(\frac{2}{3}\right) = 0.43 \right\rangle, \\ \varepsilon_3 = \left\langle \Phi^{-1}\left(\frac{3}{4}\right) = 0.67, \Phi^{-1}\left(\frac{1}{9}\right) = -1.22 \right\rangle, \\ \varepsilon_4 = \left\langle \Phi^{-1}\left(\frac{1}{8}\right) = -1.15, \Phi^{-1}\left(\frac{4}{9}\right) = -0.14 \right\rangle, \\ \varepsilon_5 = \left\langle \Phi^{-1}\left(\frac{5}{8}\right) = 0.32, \Phi^{-1}\left(\frac{7}{9}\right) = 0.76 \right\rangle, \\ \varepsilon_6 = \left\langle \Phi^{-1}\left(\frac{3}{8}\right) = -0.32, \Phi^{-1}\left(\frac{2}{9}\right) = -0.76 \right\rangle, \\ \varepsilon_7 = \left\langle \Phi^{-1}\left(\frac{7}{8}\right) = 1.15, \Phi^{-1}\left(\frac{5}{9}\right) = 0.14 \right\rangle, \ldots \end{array} \right\}$$

As the initial cycle of elements (from near zero to near one) for multiple dimension sequences are highly correlated, the initial elements are usually discarded (treated as burn-in). The number of elements discarded is at least as large as the largest prime used in creating the sequences. Since primes cycle at different rates after the first cycle, primes are more effective bases (they have smaller correlation) for Halton sequences.

Randomized Halton draws

Halton sequences are systematic, not random, while asymptotic properties of estimators assume random (or at least pseudo-random) draws of unobservables. Halton sequences can be transformed in a way that makes draws pseudo-random (as is the case for all computer-based randomizations). Bhat [2003] suggests the following procedure:

1. Take a draw μ from a standard uniform distribution.
2. Add μ to each element of the Halton sequence. If the resulting element exceeds one, subtract 1 from it. That is, $s_n = \mathrm{mod}(s_0 + \mu)$ where s_0 (s_n) is the original (transformed) element of the Halton sequence and mod$(\cdot)$ returns the fractional

part of the argument.
Suppose $\mu = 0.4$ for the above Halton sequence (again through two iterations), the pseudo-random sequence is

$$\{0.4, 0.733, 0.067, 0.511, 0.844, 0.178, 0.622, 0.956, 0.289, \ldots\}$$

The spacing remains the same so we achieve the same coverage but draws are random. In a sense, this "blocking" approach is similar to bootstrapping regressions with heteroskedastic and/or correlated errors. A different draw for μ is taken for each unobservable.

Bhat [2003] also proposes scrambled Halton draws to deal with high dimension issues. Halton sequences for high dimension problems utilize larger prime numbers. For large prime numbers, correlation in the sequences may persist for much longer than the first cycle as discussed above. Bhat proposes scrambling the sequence so that if we think of the above sequence as BC then the sequence is reversed to be CB where $B = \frac{1}{3}$ and $C = \frac{2}{3}$. Different permutations are employed for different primes. Continuing with the above Halton sequence for $k = 3$, the original and scrambled sequences are tabulated below.

Original	Scrambled
$1/3$	$2/3$
$2/3$	$1/3$
$1/9$	$2/9$
$4/9$	$8/9$
$7/9$	$5/9$
$2/9$	$1/9$
$5/9$	$7/9$
$8/9$	$4/9$

7.4 Additional reading

Kreps [1988, ch. 11] and McCall [1991] discuss exchangeability and de Finetti's theorem as well as implications for economics. Davidson and MacKinnon [2003], MacKinnon [2002], and Cameron and Trivedi [2005] discuss bootstrapping, pivotal statistics, etc., and Horowitz [2001] provides an extensive discussion of bootstrapping. Casella and George [1992] and Chib and Hamilton [1995] offer basic introductions to the Gibbs sampler and Metropolis-Hastings algorithm, respectively. Tanner and Wong [1987] discuss calculating posterior distributions by data augmentation. Train [2002, ch. 9] discusses various Halton sequence approaches and other remaining open questions associated with this relatively new, but promising technique.