

Lecture Notes based on Koop (2003) “Bayesian Econometrics”

A. Colin Cameron
University of California - Davis

November 15, 2005

1. CH.1: Introduction

The concepts below are the essential concepts used throughout the book.

1.1. Theory: Posterior density

The key is the **posterior density** - the probability density of parameters given data.

Apply Bayes Rule that $\Pr[B|A] = \{\Pr[A|B] \times \Pr[B]\} / \Pr[A]$ to the probability of parameter vector $\boldsymbol{\theta}$ (event B) given data vector \mathbf{y} (event A):

1. Likelihood function (for data given parameters) $p(\mathbf{y}|\boldsymbol{\theta})$
2. Prior density (for parameters) $p(\boldsymbol{\theta})$
3. Posterior density (for parameters given data) $p(\boldsymbol{\theta}|\mathbf{y}) = p(\boldsymbol{\theta}) \times p(\mathbf{y}|\boldsymbol{\theta}) / p(\mathbf{y})$.

The user inputs 1. (as in ML) and 2. (special to Bayesian) and gets out the **posterior density**

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta}) \times p(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})}. \quad (1.1)$$

The term $p(\mathbf{y})$ in (1.1) is a constant that does not depend on θ . For some aspects of Bayesian analysis we can ignore the constant $p(\mathbf{y})$ and simply say

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta}) \times p(\mathbf{y}|\boldsymbol{\theta}). \quad (1.2)$$

[Aside: In general if density $f(x) = kg(x)$ where k is a constant that does not depend on x then $g(x)$ is called the **kernel** of the density. Much Bayesian analysis works with just the density kernel].

The **posterior density** $p(\boldsymbol{\theta}|\mathbf{y})$ is the starting point for further analysis. Many studies report

- **Posterior mean:** $E[\boldsymbol{\theta}] = \int \boldsymbol{\theta}p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$.
- **Posterior standard deviation:** $\sigma_{\boldsymbol{\theta}}$ where $\sigma_{\boldsymbol{\theta}}^2 = E[(\boldsymbol{\theta}-E[\boldsymbol{\theta}])^2] = \int (\boldsymbol{\theta}-E[\boldsymbol{\theta}])^2p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$.
- **95% highest posterior density interval (HPDI):** The Bayesian equivalent of a 95% confidence interval (see p.44) which is usually asymmetric.

1.2. Theory: Posterior Odds

See presentation in chapter 4 below.

1.3. Theory: Predictive Density

Consider prediction of new observation(s) \mathbf{y}^* given data \mathbf{y} . We want the density $p(\mathbf{y}^*|\mathbf{y})$. This can be done by conditioning on $\boldsymbol{\theta}$ and then integrating out over the posterior distribution for $\boldsymbol{\theta}$, giving the **predictive density**

$$p(\mathbf{y}^*|\mathbf{y}) = \int p(\mathbf{y}^*|\mathbf{y}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}. \quad (1.3)$$

Often \mathbf{y} and \mathbf{y}^* are independent of each other, in which case $p(\mathbf{y}^*|\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y}^*|\boldsymbol{\theta})$.

1.4. Computation: Monte Carlo Integration

Computation is used extensively in Bayesian analysis.

Suppose we know the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$ but not how to calculate a posterior moment $E[g(\boldsymbol{\theta})]$ such as the posterior mean. To estimate $E[g(\boldsymbol{\theta})] = \int g(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$ make many draws from $p(\boldsymbol{\theta}|\mathbf{y})$ and for each draw of $\boldsymbol{\theta}$ calculate $g(\boldsymbol{\theta})$ and average. With S draws **Monte Carlo integration** yields the estimate

$$\hat{g}^S = \hat{E}[g(\boldsymbol{\theta})] = \frac{1}{S} \sum_{s=1}^S g(\boldsymbol{\theta}^{(s)}). \quad (1.4)$$

The precision of this estimate is measured using the **numerical standard error** σ_g/\sqrt{S} where

$$\hat{\sigma}_g^2 = \frac{1}{S-1} \sum_{s=1}^S (g(\boldsymbol{\theta}^{(s)}) - \hat{g}^S)^2. \quad (1.5)$$

Koop begins with this example: calculations using the posterior usually involve an integral. A second reason for computational methods is when there is no closed form from the posterior but analysis is still possible using modern MCMC methods. This is not introduced until chapter 4 but is the basis for the Bayesian revival.

There are several ways in which we use Monte Carlo methods:

1. Monte Carlo integration - when can draw from $f()$.
2. Importance sampling integration - when have decent approximation for $f()$.
3. Gibbs sampler - draws when full conditional known.
4. Metropolis-Hastings when full conditional not known.
5. Data augmentation - when observeables depend on latent variables.

2. CH.2: Single Regressor, Natural Conjugate Prior

Here $y_i \sim \mathcal{N}[\beta x_i, \sigma^2]$, so scalar regressor and no need for matrix algebra.

This is rewritten as $y_i \sim \mathcal{N}[\beta x_i, 1/h]$ where $h = 1/\sigma^2$ is called the **precision parameter**.

The chapter assumes a “normal-gamma conjugate prior” for β and h .

2.1. Priors

A **conjugate prior** is one for which combining prior and likelihood leads to a posterior in the same class of distributions.

A **natural conjugate prior** is one where the **prior density has the same functional form as the likelihood**. This has two advantages

1. Tractability - we have a closed form solution for the posterior.
2. Interpretability - the prior can be interpreted as data (see e.g. pp.21-22.)

A **relatively noninformative prior** is a prior with large variances.

A **noninformative prior** is the limit as variances go to infinity and is often a uniform prior which is often **improper** as it does not integrate to one.

2.2. Posterior

This just gives results that are repeated in chapter 3 using matrix algebra.

Underscores are used for **prior** means and variances e.g. $\underline{\beta}$.

Overscores are used for **posterior** means and variances e.g. $\overline{\beta}$.

3. CH.3: Many Regressors, Natural Conjugate Prior

Here $y_i \sim \mathcal{N}[\mathbf{x}_i\boldsymbol{\beta}, \sigma^2]$ or equivalently $\mathbf{y} \sim \mathcal{N}[\mathbf{X}\boldsymbol{\beta}, h^{-1}\mathbf{I}_k]$ and use matrix algebra.

The chapter assumes a “normal-gamma conjugate prior” for $\boldsymbol{\beta}$ and h .

3.1. Priors

It is assumed that $p(\boldsymbol{\beta}, h) = p(\boldsymbol{\beta}|h) \times p(h)$ where

$$\begin{aligned}\boldsymbol{\beta}|h &\sim \mathcal{N}[\underline{\boldsymbol{\beta}}, h^{-1}\underline{\mathbf{V}}] \\ h &\sim \mathcal{G}[\underline{s}^{-2}, \underline{v}].\end{aligned}$$

This is strange in that it is more natural to specify a prior for $\boldsymbol{\beta}$ than for $\boldsymbol{\beta}|h$. But it gives a joint prior for $\boldsymbol{\beta}$ and h that is called the **normal-gamma prior**

$$\boldsymbol{\beta}, h \sim \mathcal{NG}[\underline{\boldsymbol{\beta}}, \underline{\mathbf{V}}, \underline{s}^{-2}, \underline{v}].$$

This is the natural-conjugate prior to the normal regression model (which can be written as a normal-gamma). It leads to a tractable **normal-gamma joint posterior**:

$$\boldsymbol{\beta}, h|\mathbf{y} \sim \mathcal{NG}[\bar{\boldsymbol{\beta}}, \bar{\mathbf{V}}, \bar{s}^{-2}, \bar{v}].$$

Denote the OLS (or MLE) estimates

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ s^2 &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/v \\ v &= (N - k).\end{aligned}$$

Then the posterior means and variances are

$$\begin{aligned}\bar{\mathbf{V}} &= [\underline{\mathbf{V}}^{-1} + \mathbf{X}'\mathbf{X}]^{-1} = [\underline{\mathbf{V}}^{-1} + ([\mathbf{X}'\mathbf{X}]^{-1})^{-1}]^{-1} \\ \bar{\boldsymbol{\beta}} &= \bar{\mathbf{V}}[\underline{\mathbf{V}}^{-1}\underline{\boldsymbol{\beta}} + ([\mathbf{X}'\mathbf{X}]^{-1})^{-1}\hat{\boldsymbol{\beta}}] \\ \bar{v} &= \underline{v} + N \\ \bar{v}s^2 &= \underline{v}\underline{s}^2 + vs^2 + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'[\underline{\mathbf{V}} + [\mathbf{X}'\mathbf{X}]^{-1}]^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).\end{aligned}$$

Note that

1. The matrix $\bar{\mathbf{V}}$ is a weighted average of the inverses of prior precision $h^{-1}\underline{\mathbf{V}}$ and data (OLS) precision $h^{-1}[\mathbf{X}'\mathbf{X}]^{-1}$.

2. The posterior mean $\bar{\beta}$ is a matrix-weighted average of prior mean and data (OLS) mean.
3. The posterior degrees of freedom are sum of prior and data degrees of freedom.
4. $\bar{v}s^2$ is posterior sum of squares which equals sum of prior and data sum of squares plus an extra term.

In addition to the joint posterior of β and h we can obtain the separate **marginal posteriors** of each. These are

$$\begin{aligned}\beta|\mathbf{y} &\sim t[\bar{\beta}, \bar{s}^2\bar{\mathbf{V}}, \bar{v}] \text{ with mean } \bar{\beta}, \text{ variance } \left(\frac{\bar{v}s^{-2}}{\bar{v}-2}\right)\bar{\mathbf{V}} \\ h|\mathbf{y} &\sim \mathcal{G}[\bar{s}^{-2}, \bar{v}] \text{ with mean } \bar{s}^{-2}, \text{ variance } 2\bar{s}^{-2}/\bar{v}.\end{aligned}$$

3.2. Model Comparison: Restrictions, HPDIs

Inequality restrictions $M_1 : \mathbf{R}\beta > \mathbf{r}$ versus $M_1 : \mathbf{R}\beta \not> \mathbf{r}$ easily compared using posterior odds ratio. Inequality restrictions easily imposed using importance sampling (see later chapter 4.3).

3.3. Data Example

This is very insightful for first look at applied methods.

Data are $N = 546$ houses sold in Windsor Canada in 1987. Model is

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i.$$

How are the priors determined?

First consider h . Say $\sigma = 5,000$ is best guess (as OLS yields s between \$3,000 and \$10,000 in typical application). So $s^2 = 5000^2$ and $s^{-2} = 4 \times 10^{-8}$. And set $\bar{v} = 5$ as the $\bar{v} = \underline{v} + N = 5 + 546$ so giving roughly 1% prior weight compared to data weight.

Second consider β . For slope β if β^* is best guess than let $\text{SD}[\beta] = \beta^*/2$ as then β lies between 0 and $2\beta^*$ with probability 0.95. For intercept more of a guess.

For slope more of a guess. For covariances assume zero. Then

	$\underline{\beta}$	$V[\underline{\beta}] = \left(\frac{vs^2}{v-2}\right) \underline{\mathbf{V}}$	$\underline{\mathbf{V}}$
intercept	0	$10,000^2$	2.40
lot size	10	5^2	6.0×10^{-7}
bedrooms	5,000	$2,500^2$	0.15
bathrooms	10,000	$5,000^2$	0.60
storeys	10,000	$5,000^2$	0.60

Note that the conditional prior variance $V[\underline{\beta}|h] = h^{-1}\underline{\mathbf{V}}$ but we need the unconditional prior variance $V[\underline{\beta}|h] = \left(\frac{vs^2}{v-2}\right) \underline{\mathbf{V}}$.

For this example:

- Posterior means of $\underline{\beta}$ are much closer to OLS (noninformative prior) than to prior.
- Posterior standard deviations of $\underline{\beta}$ are smaller with informative prior than with noninformative prior.
- 95% HPDI for e.g. $\beta_5 = (5686, 9596) = 7641 \pm 2.58 \times 997.02$ so multivariate t critical value of 2.58 is much larger than usual 1.96.
- Posterior odds = Bayes factor = 0.39 for β_3 means that $M_1 : \beta_3 = 0$ is 28% likely and $M_2 : \beta_3 \neq 0$ is 72% likely ($0.28/0.72 = 0.39$).

4. CH.4: Many Regressors, Other Priors

This chapter considers two separate complications and solution methods.

1. Independent normal/gamma prior. This is no longer conjugate. Solve by using the Gibbs sampler.
2. Normal/gamma prior (so conjugate) but inequality restrictions. Solve by using importance sampling.

It also presents ways to compute posterior odds.

4.1. Gibbs Sampler for Independent Normal-Gamma Priors

It is assumed that $p(\boldsymbol{\beta}, h) = p(\boldsymbol{\beta}) \times p(h)$, so priors independent, with

$$\begin{aligned}\boldsymbol{\beta} &\sim \mathcal{N}[\underline{\boldsymbol{\beta}}, \underline{\mathbf{V}}] \\ h &\sim \mathcal{G}[\underline{s}^{-2}, \underline{v}].\end{aligned}$$

Note that h no longer appears in the prior for $\boldsymbol{\beta}$.

Then the **conditional posteriors** are easy:

$$\begin{aligned}\boldsymbol{\beta}|\mathbf{y}, h &\sim \mathcal{N}[\bar{\boldsymbol{\beta}}, \bar{\mathbf{V}}] \\ h|\mathbf{y}, \boldsymbol{\beta} &\sim \mathcal{G}[\bar{s}^{-2}, \bar{v}],\end{aligned}$$

where

$$\begin{aligned}\bar{\mathbf{V}} &= [\underline{\mathbf{V}}^{-1} + (h^{-1}[\mathbf{X}'\mathbf{X}]^{-1})^{-1}]^{-1} \\ \bar{\boldsymbol{\beta}} &= \bar{\mathbf{V}}[\underline{\mathbf{V}}^{-1}\underline{\boldsymbol{\beta}} + (h^{-1}[\mathbf{X}'\mathbf{X}]^{-1})^{-1}\hat{\boldsymbol{\beta}}] \\ \bar{s}^{-2} &= [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \underline{v}s^2]/\underline{v} \\ \bar{v} &= N + \underline{v}.\end{aligned}$$

But we don't know the **joint posterior** $p(\boldsymbol{\beta}, h|\mathbf{y})$ or the **marginal posteriors** $p(\boldsymbol{\beta}|\mathbf{y})$ and $p(h|\mathbf{y})$. So can't take usual closed-form solution approach.

Instead use **Gibbs sampler** (see separate notes). At s^{th} of S rounds:

- Given $h^{(s-1)}$ and \mathbf{y}, \mathbf{X} compute $\bar{\boldsymbol{\beta}}^{(s-1)}$ and $\bar{\mathbf{V}}^{(s-1)}$ and draw $\boldsymbol{\beta}^{(s)}$ from $\mathcal{N}[\bar{\boldsymbol{\beta}}^{(s-1)}, \bar{\mathbf{V}}^{(s-1)}]$.
- Given $\boldsymbol{\beta}^{(s-1)}$ and \mathbf{y}, \mathbf{X} compute $\bar{s}^{-2,(s)}$ and $\bar{v}^{(s)}$ and draw $h^{(s)}$ from $\mathcal{G}[\bar{s}^{-2,(s)}, \bar{v}^{(s)}]$.

If this MCMC method has converged it gives correlated draws of $\boldsymbol{\beta}^{(s)}$ and $h^{(s)}$ from the analytically unknown posterior $p(\boldsymbol{\beta}, h|y)$. Some details:

- Throw away first S_0 draws (**burn-in**) and keep next S_1 draws.
- **Numerical standard error** $\hat{\sigma}_g/\sqrt{S_1}$ needs to compute $\hat{\sigma}_g$ allowing for correlation of $g(\boldsymbol{\theta}^{(s)})$ over draws (4.13) which has typo).
- To ensure chain has converged use **MCMC diagnostics**. Koop emphasizes
 - CD statistic (compare \hat{g} for subsamples of draws): should be < 1.96 in absolute value
 - \hat{R} statistic (compare \hat{g} for different Gibbs starting values): should be ≤ 1.2 .

The empirical illustration is same data as chapter 3. Set $S_0 = 1,000$ and $S_1 = 10,000$. Results are closer to prior than in chapter 3.9. CD is computed separately for each β_j and is always ≤ 1.22 in absolute value. Numerical standard error is small. Posterior odds change more than do posterior means of parameters.

4.2. Linear Model with Inequality Constraints

The inequality constrains are $\mathbf{1}(\boldsymbol{\beta} \in \mathbf{A})$.

The marginal posterior is $p(\boldsymbol{\beta}|y) \times \mathbf{1}(\boldsymbol{\beta} \in \mathbf{A})$, the same as without restrictions except multiplied by $\mathbf{1}(\boldsymbol{\beta} \in \mathbf{A})$.

For simplicity consider a tractable model with natural conjugate prior or non-informative prior.

Then without constraints can just draw from the joint posterior $p(\boldsymbol{\beta}, h|y)$.

With constraints we use importance sampling where we draw from $p(\boldsymbol{\beta}, h|y)$, keep only those draws with $\boldsymbol{\beta} \in \mathbf{A}$, so that $\mathbf{1}(\boldsymbol{\beta} \in \mathbf{A}) = 1$, and keep track of the number of draws for which $\mathbf{1}(\boldsymbol{\beta} \in \mathbf{A}) = 1$.

More generally **importance sampling** estimates $E[g(\boldsymbol{\theta})|\mathbf{y}] = \int g(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$ by instead drawing $\boldsymbol{\theta}^{(s)}$ from the **importance function** $q(\boldsymbol{\theta})$ and forming the weighted average

$$\hat{g}^S = \hat{E}[g(\boldsymbol{\theta})] = \frac{\sum_{s=1}^S w(\boldsymbol{\theta}^{(s)})g(\boldsymbol{\theta}^{(s)})}{\sum_{s=1}^S w(\boldsymbol{\theta}^{(s)})},$$

where

$$w(\boldsymbol{\theta}^{(s)}) = \frac{p(\boldsymbol{\theta} = \boldsymbol{\theta}^{(s)}|\mathbf{y})}{q(\boldsymbol{\theta} = \boldsymbol{\theta}^{(s)})} \text{ are weights.}$$

Here $w(\boldsymbol{\theta}^{(s)}) = \mathbf{1}(\boldsymbol{\beta}^{(s)} \in \mathbf{A})$.

The empirical illustration is same data as chapter 3.9 and 4.2. Here $\mathbf{1}(\boldsymbol{\beta} \in \mathbf{A})$ imposes $\beta_2 > 5$, $\beta_3 > 2500$, $\beta_4 > 5,000$ and $\beta_5 > 5,000$. There are 10,000 reps with importance sampling. Numerical standard error about 50% larger than in Gibbs sampler. β_3 and β_5 change quite a bit. Posterior odds change quite a bit.

4.3. Posterior Odds Computation

Posterior odds are the Bayesian analog of a likelihood ratio. Here we combine discussion in sections 1.1, 4.2.5, 4.3.4, 5.7 and 7.5. We give definitions and various ways to compute.

First, a frequentist analysis. For two models M_1 and M_2 let $p(\mathbf{y}, \boldsymbol{\theta}|M_1)$ and $p(\mathbf{y}, \boldsymbol{\theta}|M_2)$ denote their fitted likelihoods. Then the likelihood ratio $L_{12} = p(\mathbf{y}, \boldsymbol{\theta}|M_1)/p(\mathbf{y}, \boldsymbol{\theta}|M_2)$ is the basis for a likelihood ratio test. If M_2 is nested in M_1 and imposes one restriction on $\boldsymbol{\theta}$ then an asymptotic likelihood ratio test will reject M_2 in favor of M_1 if $2 \ln L_{12}$ exceeds $\chi_{.05}^2(1) = 3.84$, or equivalently if L_{12} exceeds $\exp(1.92) = 6.82$.

4.3.1. Posterior Odds and Bayes Factor (section 1.1)

Now the Bayesian analysis. The starting point is the unconditional probability of observing our data \mathbf{y} , i.e. after integrating out that parameters $\boldsymbol{\theta}$. This is called the **marginal likelihood**, and for model M_j is given by

$$p(\mathbf{y}|M_j) = \int p(\mathbf{y}|\boldsymbol{\theta}^j, M_j) \times p(\boldsymbol{\theta}^j|M_j) d\boldsymbol{\theta}^j. \quad (4.1)$$

This is the normalizing constant in the posterior density (1.1) and can be difficult to compute.

Bayes factor is simply the ratio of the marginal likelihoods

$$BF_{12} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}. \quad (4.2)$$

For example $BF_{12} = 2$ then model one is twice as likely as model 2 or the probability of model 1 being appropriate is 0.67 while for model 2 this probability is 0.33.

More generally we can weight by prior model probabilities $p(M_j)$ giving the **posterior odds ratio**

$$PO_{12} = \frac{p(\mathbf{y}|M_1) \times p(M_1)}{p(\mathbf{y}|M_2) \times p(M_2)} = \frac{p(M_1|\mathbf{y})}{p(M_2|\mathbf{y})}, \quad (4.3)$$

where the **posterior model probability**

$$p(M_j|\mathbf{y}) = \frac{p(\mathbf{y}|M_j) \times p(M_j)}{p(\mathbf{y})}$$

This reduces to BF_{12} if the prior model probabilities are equal, and what is reported as the posterior odds ratio in studies is usually the Bayes factor.

Note that care is needed in using posterior odds when an improper prior (such as uniform) is used. See page 42.

Several methods exist for computing posterior odds.

4.3.2. Savage-Dickey Density Ratio (section 4.2.5)

Suppose M_1 restricted is nested within M_2 and has equality restrictions, that a subcomponent of θ takes a particular value. Thus M_1 has (θ_1^0, θ_2) and M_2 has (θ_1, θ_2) . And suppose imposing M_1 on θ_1 does not effect prior on θ_2 , so $p(\theta_2|\theta_1 = \theta_1^0, M_1) = p(\theta_2|M_2)$. Then can show

$$BF_{12} = \frac{p(y|M_1)}{p(y|M_2)} = \frac{p(\theta_1 = \theta_1^0|y, M_2)}{p(\theta_1 = \theta_1^0|M_2)}.$$

So the **Bayes factor** can be computed as the marginal posterior for θ_1 in the unrestricted model divided by the marginal prior for θ_1 in the unrestricted model, where both are evaluated at $\theta_1 = \theta_1^0$. This ratio is called the **Savage-Dickey density ratio**.

This is convenient. Only need work with the unrestricted model. And there is no need to find the marginal density. Just need the marginal posterior (e.g. via Gibbs) and the marginal prior.

4.4. Inequality Constraints (section 4.3.4)

Suppose M_1 is $\beta \in \mathbf{A}$ and M_2 is $\beta \notin \mathbf{A}$. Use importance sampling to impose M_1 . Then $p(M_1|\mathbf{y})$ is simply the number of draws kept in importance sampling (i.e.

with $\beta \in \mathbf{A}$) and $p(M_2|\mathbf{y}) = 1 - p(M_1|\mathbf{y})$. So can easily compute Bayes factor BF_{12} and posterior odds PO_{12} .

Suppose additionally one model has an extra constraint. Say M_1 is $\beta \in \mathbf{A}$ plus $\beta = \beta_0$ and M_2 is $\beta \notin \mathbf{A}$. Then can show

$$BF_{12} = \frac{p(y|M_1)}{p(y|M_2)} = \frac{\bar{c} \times (\text{posterior kernel with } \beta = \beta^0)}{\underline{c} \times (\text{posterior kernel with } \beta = \beta^0)},$$

where \bar{c}^{-1} is the number of posterior draws with $\beta \in \mathbf{A}$ and \underline{c}^{-1} is the number of prior draws with $\beta \in \mathbf{A}$. Only need posterior kernels!

4.4.1. Gelfand-Day Method (section 5.7)

Previous methods are special cases where can compute BF and sometimes PO without computing marginal likelihood $p(\mathbf{y}|M_j)$. For any density $f(\theta)$ with support contained in Θ

$$\mathbb{E} \left[\frac{f(\theta)}{p(\theta|M_j)p(\mathbf{y}|\theta, M_j)} \mid \mathbf{y}, M_j \right] = \frac{1}{p(\mathbf{y}|M_j)}.$$

The left-hand side can be computed by simulated draws from $f(\theta)$, provided we know the complete functional form for the prior $p(\theta|M_j)$ and likelihood $p(\mathbf{y}|\theta, M_j)$, and not just the kernels. Taking the reciprocal gives an estimate of $p(\mathbf{y}|M_j)$.

4.4.2. Chib Method of Marginal Likelihood Computation (section 7.5)

Useful when large number of parameters. Uses Gibbs sampling and just requires a few extra lines of code. Key is that for any θ^*

$$p(\mathbf{y}) = \frac{p(\mathbf{y}|\theta^*)p(\theta^*)}{p(\theta^*|\mathbf{y})},$$

and $p(\theta^*|\mathbf{y}) = p(\theta_1^*, \theta_2^*|\mathbf{y}) \simeq \widehat{p(\theta_2^*|\mathbf{y})} \times \widehat{p(\theta_1^*|\mathbf{y})}$.

5. CH.5: Nonlinear Regression Model

Now $\mathbf{y} = \mathbf{f}(\mathbf{X}, \boldsymbol{\gamma}) + \boldsymbol{\varepsilon}$ rather than $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

With prior density $p(\boldsymbol{\gamma}, h)$ the posterior

$$p(\boldsymbol{\gamma}, h|\mathbf{y}) \propto p(\boldsymbol{\gamma}, h) \times \frac{h^{N/2}}{(2\pi)^{N/2}} \times \exp\left(-\frac{h}{2}(\mathbf{y} - \mathbf{f}(\mathbf{X}, \boldsymbol{\gamma}))'(\mathbf{y} - \mathbf{f}(\mathbf{X}, \boldsymbol{\gamma}))\right).$$

This is handled using Metropolis-Hastings algorithm. This draws $\boldsymbol{\gamma}$ from a candidate generating density but accepts only some of the draws and otherwise stays with the preceding value of $\boldsymbol{\gamma}$. One example is to draw $\boldsymbol{\gamma}$ from $t(\widehat{\boldsymbol{\gamma}}_{ML}, \widehat{\mathbf{V}}[\widehat{\boldsymbol{\gamma}}_{ML}], 10)$.

For general $\boldsymbol{\theta}$ the algorithm to draw from $p(\boldsymbol{\theta}|\mathbf{y})$ is:

- (0) start with $\boldsymbol{\theta}^{(0)}$
- (1) draw candidate value $\boldsymbol{\theta}^*$ from candidate density $q(\boldsymbol{\theta}^{(s-1)}; \boldsymbol{\theta}^*)$
- (2) calculate acceptance probability $\alpha(\boldsymbol{\theta}^{(s-1)}; \boldsymbol{\theta}^*)$ - see below
- (3) let $\boldsymbol{\theta}^{(s)} = \boldsymbol{\theta}^*$ with probability $\alpha(\boldsymbol{\theta}^{(s-1)}; \boldsymbol{\theta}^*)$ and otherwise equal $\boldsymbol{\theta}^{(s-1)}$
- (4) repeat (1)-(3) S times.

This MCMC method gives S correlated draws from $p(\boldsymbol{\theta}|\mathbf{y})$.

The acceptance probability varies with $q(\boldsymbol{\theta}^{(s-1)}; \boldsymbol{\theta}^*)$. If $q(\boldsymbol{\theta}^{(s-1)}; \boldsymbol{\theta}^*) = q^*(\boldsymbol{\theta}^*)$ so it does not depend on $\boldsymbol{\theta}^{(s-1)}$ and furthermore is symmetric then we have Metropolis with

$$\alpha(\boldsymbol{\theta}^{(s-1)}; \boldsymbol{\theta}^*) = \min\left(\frac{p(\boldsymbol{\theta} = \boldsymbol{\theta}^*|\mathbf{y})|\mathbf{y}}{p(\boldsymbol{\theta} = \boldsymbol{\theta}^{(s-1)}|\mathbf{y})}, 1\right).$$

Metropolis-Hastings allows more complicated $q(\boldsymbol{\theta}^{(s-1)}; \boldsymbol{\theta}^*)$ with more complicated $q(\boldsymbol{\theta}^{(s-1)}; \boldsymbol{\theta}^*)$. See Koop.

The empirical illustration has CES example with $N = 123$ and one regressor. For noninformative prior it has $S = 25,000$ and acceptance rates of 7% for independence chain with candidate density $t(\widehat{\boldsymbol{\gamma}}_{ML}, \widehat{\mathbf{V}}[\widehat{\boldsymbol{\gamma}}_{ML}], 10)$ and 21% for draws from random walk chain $\mathcal{N}(\widehat{\boldsymbol{\gamma}}_{ML}, \widehat{\mathbf{V}}[\widehat{\boldsymbol{\gamma}}_{ML}])$.

For informative independent normal-gamma prior the Metropolis-Hastings is embedded within Gibbs.

6. CH.6: GLS for Linear Regression Model

Here $y_i \sim \mathcal{N}[\mathbf{x}_i\boldsymbol{\beta}, h^{-1}\boldsymbol{\Omega}]$ where now $\boldsymbol{\Omega} \neq \mathbf{I}$.

6.1. Priors

Independent priors are assumed with $p(\boldsymbol{\beta}, h) = p(\boldsymbol{\beta}) \times p(h) \times p(\boldsymbol{\Omega})$ where

$$\begin{aligned}\boldsymbol{\beta} &\sim \mathcal{N}[\boldsymbol{\beta}, h^{-1}\mathbf{V}] \\ h &\sim \mathcal{G}[\underline{s}^{-2}, \underline{v}] \\ \boldsymbol{\Omega} &\sim \text{varies with setting.}\end{aligned}$$

Then the full conditionals for the posterior are

$$\begin{aligned}\boldsymbol{\beta}|\mathbf{y}, h, \boldsymbol{\Omega} &\sim \mathcal{N}[\bar{\boldsymbol{\beta}}, \bar{\mathbf{V}}], \quad \bar{\mathbf{V}} = (\mathbf{V}^{-1} + h\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1} \\ h|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\Omega} &\sim \mathcal{G}[\bar{s}^{-2}, \bar{v}], \quad \bar{v} = N + \underline{v} \\ p(\boldsymbol{\Omega}|\mathbf{y}, \boldsymbol{\beta}, h) &\propto p(\boldsymbol{\Omega})f(\mathbf{y}|\boldsymbol{\beta}, h, \boldsymbol{\Omega})\end{aligned}$$

Thus can use Gibbs sampling, where sometimes $\boldsymbol{\Omega}|\mathbf{y}, \boldsymbol{\beta}, h$ can be directly drawn and sometimes Metropolis-Hastings is used.

For **heteroskedasticity of known form**,

$$V[\varepsilon_i] = h^{-1}(1 + \alpha_1 z_{1i} + \alpha_2 z_{2i} + \cdots + \alpha_p z_{pi})^2$$

(though Koop says can consider other functions). Then $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\alpha})$ and an uninformative prior is used so $p(\boldsymbol{\alpha}) = 1$. Then need to use Metropolis-Hastings (Koop uses with random walk chain).

For **heteroskedasticity of unknown form**, a **hierarchical prior** is used with

$$\begin{aligned}\varepsilon_i &\sim \mathcal{N}[0, h^{-1}\lambda_i^{-1}] \\ \lambda_i &\sim \mathcal{G}[1, v_\lambda] \\ v_\lambda &\sim \mathcal{G}[\underline{v}_\lambda, 2].\end{aligned}$$

Koop says this allows heteroskedastic errors but I think they do not depend on regressors so it just gives fatter tail errors than normal, but still homoskedastic. Koop uses M-H with random walk chain.

For **autocorrelated errors**, Koop considers AR(p) errors, so

$$\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + \cdots + \rho_p\varepsilon_{t-p} + u_t,$$

where u_t is white noise. Then convert to $y_t^* = \mathbf{x}_t' \boldsymbol{\beta} + u_t$. For prior for $\boldsymbol{\rho} = (\rho_1, \dots, \rho_p)$ use truncated multivariate normal, with $p(\boldsymbol{\rho}) = f_{normal}(\boldsymbol{\rho}, \underline{\boldsymbol{\rho}}, \mathbf{V}_{\boldsymbol{\rho}}) \times \mathbf{1}(\boldsymbol{\rho} \in \text{unit sphere})$. Can use just Gibbs.

For **seemingly unrelated regressions**, consider case where errors within equation are iid but across equations they are correlated. Then back to chapter 4, except scalar h is replaced by matrix \mathbf{H} and gamma prior on h^{-1} is replaced by **Wishart prior** on \mathbf{H} . Can do Gibbs, drawing from $p(\boldsymbol{\beta} | \mathbf{y}, \mathbf{H})$ and $p(\mathbf{H} | \mathbf{y}, \boldsymbol{\beta})$.

7. CH.7: Panel Data and Linear Regression Model

Notation is data $(y_{it}, \mathbf{x}_{it})$, $i = 1, \dots, N$, $t = 1, \dots, T$

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} \quad \text{etc. as usual}$$

With intercept we have $\mathbf{X}'_i \boldsymbol{\beta}$ and without intercept have $\tilde{\mathbf{X}}'_i \tilde{\boldsymbol{\beta}}$.

The **pooled model** sets $\boldsymbol{\beta}$ same for all i and t . So

$$y_{it} = \mathbf{x}'_{it} \boldsymbol{\beta} + \varepsilon_{it}.$$

Then just OLS if ε_{it} iid, and SUR if $V[\boldsymbol{\varepsilon}_i] = h^{-1} \boldsymbol{\Omega}$.

The **individual effects model** lets intercept vary over i . So

$$\begin{aligned} y_{it} &= \alpha_i + \tilde{\mathbf{x}}'_{it} \tilde{\boldsymbol{\beta}} + \varepsilon_{it} \\ \mathbf{y}_i &= \alpha_i \mathbf{i}_T + \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}_i \\ \mathbf{y} &= \mathbf{I}_{NT} \boldsymbol{\alpha} + \tilde{\mathbf{X}} \tilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}. \end{aligned}$$

A **nonhierarchical prior** treats $\boldsymbol{\alpha}$ like $\tilde{\boldsymbol{\beta}}$ and imposes a normal prior on $(\boldsymbol{\alpha}, \tilde{\boldsymbol{\beta}})$. This is like fixed effects, and is same as chapter 4. A **hierarchical prior** posits $\alpha_i \sim \mathcal{N}[\mu_\alpha, v_\alpha]$, where $\mu_\alpha \sim \mathcal{N}[\underline{\mu}_\alpha, \underline{\sigma}_\alpha^2]$ and $v_\alpha \sim \mathcal{G}[\underline{v}_\alpha^{-1}, \underline{v}_\alpha^{-1}]$, while as usual $\tilde{\boldsymbol{\beta}}$ has a normal prior. This is like **random effects**, and needs the Gibbs sampler.

A **random coefficients model** allows all regression parameters to vary over i . Then

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i,$$

where a hierarchical prior for $\boldsymbol{\beta}_i$ is used with $\boldsymbol{\beta}_i \sim \mathcal{N}[\mu_\beta, \mathbf{V}_\beta]$, $\mu_\beta \sim \mathcal{N}[\underline{\mu}_\beta, \underline{\Sigma}_\beta]$, and $\mathbf{V}_\beta^{-1} \sim \text{Wishart}[\underline{v}_\beta, \underline{\mathbf{V}}_\beta^{-1}]$.

8. CH.9: Data Augmentation

Discrete data and censored models (limited dependent variable models) can be motivated as having dependent variable y that is a partially observed **latent variable** y^* . Thus for probit

$$y^* = \mathbf{x}'\boldsymbol{\beta} + u, \quad u \sim \mathcal{N}[0, 1]$$

$$y = \begin{cases} 1 & \text{if } y^* > 0 \\ 0 & \text{if } y^* \leq 0, \end{cases}$$

and for Tobit

$$y^* = \mathbf{x}'\boldsymbol{\beta} + u, \quad u \sim \mathcal{N}[0, \sigma^2]$$

$$y = \begin{cases} y^* & \text{if } y^* > 0 \\ 0 & \text{if } y^* \leq 0. \end{cases}$$

Analysis would be much simpler in both cases if y^* was observed - we would just be able to use chapter 3 results for normal linear regression, and the probit case would be especially simple as $\sigma^2 = 1$ so we need only analyze $\boldsymbol{\beta}$.

For models with latent variable y^* , Tanner and Wong (JASA, 1987) proposed the data augmentation method. Mathematically note that in general

$$p(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta}, \mathbf{y}^*) d\mathbf{y}^* = \int p(\boldsymbol{\theta}|\mathbf{y}^*) p(\mathbf{y}^*) d\mathbf{y}^*.$$

Now additionally condition on \mathbf{y} throughout. Then the **posterior** $p(\boldsymbol{\theta}|\mathbf{y})$

$$p(\boldsymbol{\theta}|\mathbf{y}) = \int p(\boldsymbol{\theta}|\mathbf{y}^*, \mathbf{y}) p(\mathbf{y}^*|\mathbf{y}) d\mathbf{y}^*$$

$$\simeq \frac{1}{M} \sum_{m=1}^M p(\boldsymbol{\theta}|\mathbf{y}^{*(m)}),$$

where we use $p(\boldsymbol{\theta}|\mathbf{y}^*, \mathbf{y}) = p(\boldsymbol{\theta}|\mathbf{y}^*)$, as \mathbf{y} adds no information beyond \mathbf{y}^* , and $\mathbf{y}^{*(m)}$ are M draws from $p(\mathbf{y}^{*(s)}|\mathbf{y})$, called the predictive density.

It is not possible to make draws from $p(\mathbf{y}^*|\mathbf{y})$ as $p(\mathbf{y}^*|\mathbf{y}) = \int p(\mathbf{y}^*|\boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$ and $p(\boldsymbol{\theta}|\mathbf{y})$ is the unknown desired posterior. Instead the **data augmentation method** uses the following iterative process where at the s^{th} round

1. **imputation step:** (a) draw $\boldsymbol{\theta}^{(s)}$ from the current estimate of $p(\boldsymbol{\theta}|\mathbf{y})$, and then (b) impute M values of \mathbf{y}^* by making M draws from $p(\mathbf{y}^*|\boldsymbol{\theta}^{(s)}, \mathbf{y})$ (this is the data augmentation)

2. **posterior step:** given the M draws of \mathbf{y}^* estimate $p(\boldsymbol{\theta}^{(s)}|\mathbf{y}) = \frac{1}{M} \sum_{m=1}^M p(\boldsymbol{\theta}^{(s)}|\mathbf{y}^{*(m)})$.

At convergence $p(\boldsymbol{\theta}^{(\infty)}|\mathbf{y})$ is the approximation for $p(\boldsymbol{\theta}|\mathbf{y})$.

The method is actually often implemented differently. It is enough to have $M = 1$. Then the algorithm reduces to:

- a. draw $\boldsymbol{\theta}^{(s)}$ from $p(\boldsymbol{\theta}^{(s)}|\mathbf{y}^{*(s-1)})$
- b. draw $\mathbf{y}^{*(s)}$ from $p(\mathbf{y}^*|\boldsymbol{\theta}^{(s)}, \mathbf{y})$
- c. return to a.

Similar to Gibbs sampling, at each round we get a value of $\boldsymbol{\theta}^{(s)}$, and provided the method has converged (i.e. after burn-in) the S values of $\boldsymbol{\theta}^{(s)}$ approximate $p(\boldsymbol{\theta}|\mathbf{y})$.

Note that given \mathbf{y}^* there is assumed to be no problem in computing $p(\boldsymbol{\theta}|\mathbf{y}^*, \mathbf{y})$. For these latent variable models \mathbf{y} contains less information than \mathbf{y}^* so $p(\boldsymbol{\theta}|\mathbf{y}^*, \mathbf{y}) = p(\boldsymbol{\theta}|\mathbf{y}^*)$. If we work with natural conjugate priors then $p(\boldsymbol{\theta}|\mathbf{y}^*) \propto p(\boldsymbol{\theta}) \times p(\mathbf{y}^*|\boldsymbol{\theta})$ is readily obtained.

For the probit model the only parameters are $\boldsymbol{\beta}$ (as $\sigma^2 = 1$). With an uninformative prior on $\boldsymbol{\beta}$, the posterior $p(\boldsymbol{\beta}|\mathbf{y}^*)$ is simply $\mathcal{N}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*, (\mathbf{X}'\mathbf{X})^{-1}]$. And the draws from $p(\mathbf{y}^*|\boldsymbol{\beta}, \mathbf{y})$ are draws from $\mathcal{N}[\mathbf{X}\boldsymbol{\beta}, \mathbf{I}]$ with left truncation at $\mathbf{0}$ if $y = 1$ and right truncation at $\mathbf{0}$ if $y = 0$. The algorithm is

- a. draw $\boldsymbol{\beta}^{(s)}$ from $\mathcal{N}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^{*(s-1)}, (\mathbf{X}'\mathbf{X})^{-1}]$
- b. for $i = 1, \dots, N$ draw $y_i^{*(s)}$ from $\mathcal{N}[\mathbf{x}'\boldsymbol{\beta}^{(s)}, 1]$, left truncated at 0 if $y_i = 1$ and right truncated at 0 if $y_i = 0$. This gives a draw of $\mathbf{y}^{*(s)}$.
- c. return to a.
- d. stop after S iterations and throw away first S_1 (burn-in).

There are various ways to draw from truncated normal. For left-truncated at zero we can just draw from normal until get a draw greater than zero, but this is computationally expensive if $\mathbf{x}'\boldsymbol{\beta}^{(s)}$ is large and negative as there will be a low probability that the draw exceeds zero. Better is to use the inverse-transformation method applied to the truncated distribution.

The book also considers Tobit, ordered probit and multinomial probit models.