Chapter 7 Simultaneity

> *Using this Chapter*
>
> This chapter discusses the problem of Bayesian inference for models in which both the response variable and some of marketing mix variables are jointly determined. We can no longer focus only on the model of the distribution of the response variable conditional on the marketing mix variables. We must build a model which (at least implicitly) specifies the joint distribution of these variables conditional on a set of driving or "exogenous" variables. Section 7.1 provides a Bayesian treatment of the linear "instrumental" variables problem, a problem for which standard classical asymptotic methods have proved inadequate. Section 7.2 considers a system of supply and demand where the demand system is built up by aggregating consumer level choice models. Section 7.3 considers the situation in which simultaneity is present in a hierarchical model.

At the base of all the models considered so far is the distribution of a dependent variable (or vector) conditional on a set of independent factors. The classic marketing example is a sales response model in which quantity demanded is modeled conditional on marketing mix variables which typically include price and advertising measures. However, we should recognize that firms may set these marketing mix variables in a strategic fashion. For example, firms may consider strategic considerations or the response of other competitors in setting price. Firms may also set the levels of marketing mix variables by optimizing a profit function that involves the sales response parameters. These considerations lead toward a joint or "simultaneous" model of the entire vector of sales responses and marketing mix variables. In this chapter, we will consider three approaches to this problem. We will first consider a Bayesian version of the instrumental variables or "limited information" approach. We will then consider a joint or simultaneous approach. Finally, we will consider the implications of optimization in the selection of the marketing mix on the estimation of conditional sales response models.

7. 1 A Bayesian Approach to Instrumental Variables

Instrumental variables techniques are widely used in economics to avoid the "endogeneity" bias of including an independent variable that is correlated with the error term. It is probably best to start with a simple example[1].

(7.1.1) $$x = \delta z + \varepsilon_1$$

(7.1.2) $$y = \beta x + \varepsilon_2$$

If $\varepsilon_1, \varepsilon_2$ are independent, then both (7.1.2) and (7.1.1) are valid regression equations and form what is usually termed a recursive system. From our perspective, we can analyse each equation separately using the standard Bayesian treatment of regression. However, if $\varepsilon_1, \varepsilon_2$ are dependent, then (7.1.2) is no longer a regression equation in the sense that the conditional distribution of ε_2 given x depends on x. In this case, x is often referred to as an "endogenous" variable. If (7.1.1) is still a valid regression equation, econometricians call z an "instrument." z is a variable related to x but independent of ε_2.

The system in (7.1.1)-(7.1.2) can be motivated by the example of a sales response model. If y is sales volume in units and x is price, then there are situations in which x can depend on the value of ε_2. One situation (c.f. Villas-Boas and Winer (1999)) has a shock which affects the demand of all consumers and which is known to the firm setting prices. For example, a manufacturer coupon will be dropped in a market. The drop of a blanket coupon will presumably increase retail demand for the product for all consumers. If retailers know about the coupon drop, they may adjust retail price (x) in order to take advantage of this knowledge. This would create a dependence or correlation between x and ε_2. Another classic example, in this same vein, is the "unobserved" characteristics argument of Berry, Levinson and Pakes (1995). In this example, there are characteristics of a product which drive demand but are not observed to the econometrician. These unobserved characteristics also influence price. If we are looking across markets, then these characteristics could be market-specific demand differences. The data represent a time series then these characteristics would have to vary across time. In both cases, this motivates a dependence between x and ε_2 in the same manner as the common demand shock argument.

These examples provide an "omitted" variables interpretation of the "endogenity" problem. If we were able to observe the omitted variable, then we would have a standard multivariate regression system.

$$x = \delta z + \alpha_x w + u_1 \tag{7.1.3}$$

$$y = \beta x + \alpha_y w + u_2 \tag{7.1.4}$$

If w is unobserved, this induces a dependence or correlation between x and $\varepsilon_2 = \alpha_y w + u_2$. w would be interpreted as the common demand shock or unobserved characteristic (s) in the examples mentioned above. The consequence of ignoring the dependence between x and ε_2 and analyzing (7.1.2) using standard conditional models is a so-called "endogneity" bias. Only part of the movement in x helps us to identify β. The part of the movement in x which is correlated with ε_2 should not be used in inferring about β. This means that if the correlation between ε_1 and ε_2 is positive, there will be a positive "endogneiety" bias.

To illustrate this situation, figure 7.1 plots x vs z and y vs x using a "brushing" technique. The data is simulated from a model in which errors are bivariate normal with correlation = .8. Points of the same color are linked between the two scatterplots. In the

[1] See also Lancaster (2004), chapter 8 for an excellent introduction to instrumental variables as well as a lucid

graph of x vs z, we hold the variation in the residuals of x on z relatively constant. Thus, within each color group we see variation in x that has been created by movements in the instrument z. The dotted lines in the y vs x graph show a least squares estimate of β for each brushing group. The solid black line shows the biased least squares estimate of β from a simple regression of y on x while the solid blue line shows the true value of β. The role of the instrument is to inject variation into x which is independent of ε_2.

It is easy to see that if the instrument is "weak" in the sense that it induces little variation in x, then there will be little information in the sample regarding β. What is less obvious is that, as instruments become weaker and weaker, we approach an unidentified case. To see this, we need to write down the likelihood for this model. The likelihood is the joint distribution of (x,y) given z. This is often called the "reduced" form. In order to proceed further, we will need to make assumptions regarding the joint distribution of $\varepsilon_1, \varepsilon_2$. Given the regression specification, a natural starting point would be to assume that $\varepsilon_1, \varepsilon_2$ are bivariate normal.

The joint distribution of (x,y) given z can be written down by substituting (7.1.1) into (7.1.2).

$$\begin{aligned} x &= \delta z + \varepsilon_1 \\ y &= \beta\delta z + \left(\beta\varepsilon_1 + \varepsilon_2\right) \end{aligned} \qquad (7.1.5)$$

or

$$\begin{aligned} x &= \pi_x z + v_1 \\ y &= \pi_y z + v_2 \end{aligned} \qquad (7.1.6)$$

discussion of the classic returns to education endogeneity problem.

Here $\beta = {}^{\pi_y}\!/_{\pi_x}$ and we can think of the model as a multivariate regression with a restriction on the coefficient matrix. This is the approach taken by Lancaster (2004). However, we must recognize that the covariance of v_1, v_2 depends on β and elements of the covariance matrix of the "structural" equation errors, ε_1 and ε_2. This is also the source of potential identification problems in this model. Consider the case with $\delta = 0$. Here

(7.1.7)
$$\begin{aligned} x &= \varepsilon_1 \\ y &= \beta\varepsilon_1 + \varepsilon_2 \end{aligned}$$

Variances and covariances are identified via observation of x and y. (7.1.7) implies that

(7.1.8)
$$\frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)} = \beta + \frac{\sigma_{12}}{\sigma_{11}} \quad \text{with } \sigma_{12} = \operatorname{cov}(\varepsilon_1, \varepsilon_2)$$

The identified quantity on the left hand side of (7.1.8) can be achieved by many different combinations of β and $\frac{\sigma_{12}}{\sigma_{11}}$. Thus, the model, in the limiting case of weak instruments, is not identified. In the case where variation in δz is a small fraction of the total variation in x, there will be a "ridge" in the likelihood, reflecting the trade-off between β and $\frac{\sigma_{12}}{\sigma_{11}}$[2].

Consider a more general version of (7.1.1) - (7.1.2),

(7.1.9)
$$x = z'\delta + \varepsilon_1$$

(7.1.10)
$$y = \beta x + w'\gamma + \varepsilon_2$$

[2] Note that this is different from the intuition in Lancaster (2004). Lancaster parameterizes the covariance of the reduced form errors, ignoring the relationship between the structural errors and the reduced form errors. The dependence between δ and β given the correlation in reduced form errors shown in Lancaster is, in fact, due to the correlation between β and $\frac{\sigma_{12}}{\sigma_{11}}$ in our parameterization.

(7.1.11) $$\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\end{pmatrix} \sim N(0,\Sigma)$$

This system shows the case a structural equation (7.1.10) with one "endogenous" variable, multiple instruments, and an arbitrary number of other regressors. We will put standard conditionally conjugate priors on these parameters.

(7.1.12) $$\begin{gathered}\Sigma \sim IW(\upsilon, V)\\ \delta \sim N\left(\bar{\delta}, A_{\delta}^{-1}\right)\\ \begin{pmatrix}\beta\\ \gamma\end{pmatrix} \sim N\left(\begin{pmatrix}\bar{\beta}\\ \bar{\gamma}\end{pmatrix}, A_{\beta\gamma}^{-1}\right)\end{gathered}$$

We believe that it is appropriate to put a prior on the covariance matrix of the structural errors. Lancaster (2004) puts an IW prior on the covariance matrix of the reduced form errors and makes this prior independent of the prior on β. The reduced form errors are related to the structural errors by the transformation:

(7.1.13) $$\begin{pmatrix}v_1\\ v_2\end{pmatrix} = \begin{bmatrix}1 & 0\\ \beta & 1\end{bmatrix}\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\end{pmatrix}$$

Given the relationship in (7.1.13), we do not feel that a prior on the covariance of the reduced form errors should be independent of the prior on β. In our prior, we induce a dependence via (7.1.13).

We can easily develop a Gibbs Sampler for the model in (7.1.9) - (7.1.12)[3]. The basic three sets of conditionals are given by

(7.1.14) $$\beta, \gamma \,\middle|\, \delta, \Sigma, x, y, w$$

(7.1.15) $$\delta \,\middle|\, \beta, \gamma, \Sigma, x, y, w$$

[3] Geweke (1996) considers a model similar to ours but with a "shrinkage" prior that specifies that each of the regression coefficients are independent with zero mean and the same variance.

(7.1.16) $$\Sigma \,\middle|\, \beta,\gamma,\delta,x,y,w$$

The first conditional in (7.1.14) can easily be accomplished by a standard Bayesian regression analysis. The key insight is to recognize that given δ, we can "observe" ε_1. We can then condition our analysis of (7.1.10) on ε_1.

(7.1.17) $$\begin{aligned} y &= \beta\left(z'\delta+\varepsilon_1\right)+w'\gamma+\varepsilon_2\,|\varepsilon_1 \\ &= \beta\left(z'\delta+\varepsilon_1\right)+w'\gamma+\frac{\sigma_{12}}{\sigma_{11}}\varepsilon_1+v_{2|1} \end{aligned}$$

Since $\operatorname{var}\left(v_{2|1}\right)\equiv\sigma_{2|1}^2=\sigma_{22}-\frac{\sigma_{12}^2}{\sigma_{11}}$, we can rewrite (7.1.17) so that we can use standard Bayes regression with a unit variance error term.

(7.1.18) $$\left(y-\tfrac{\sigma_{12}}{\sigma_{11}}\varepsilon_1\right)\Big/\sigma_{2|1}=\beta\left(z'\delta+\varepsilon_1\right)+w'\gamma+\zeta \qquad \zeta\sim N(0,1)$$

The second conditional in (7.1.15) can be handled by transforming to the reduced form which can be written as a regression model with "double" the number of observations.

(7.1.19) $$\begin{aligned} x &= z'\delta+\varepsilon_1 \\ \tilde{y} &= \left(\frac{y-w'\gamma}{\beta}\right)=z'\delta+\left(\varepsilon_1+\varepsilon_2/\beta\right) \end{aligned}$$

We can transform the system above to an uncorrelated set of regressions by computing the covariance matrix of the vector of errors in (7.1.19).

(7.1.20) $$\begin{aligned} &\operatorname{Var}\begin{pmatrix}\varepsilon_1 \\ \varepsilon_1+\frac{1}{\beta}\varepsilon_2\end{pmatrix}=A\Sigma A'=\Omega=LL' \\ &A=\begin{bmatrix}1 & 0 \\ 1 & \frac{1}{\beta}\end{bmatrix} \end{aligned}$$

Pre-multiplying (7.1.19) with L^{-1} reduces the system to a bivariate system with unit covariance matrix and we can simply stack it up and perform a Bayes regression analysis with unit variance. That is,

$$L^{-1}\begin{pmatrix} x \\ \tilde{y} \end{pmatrix} = L^{-1}\begin{bmatrix} z' \\ z' \end{bmatrix}\delta + u \qquad \mathrm{Var}(u) = I_2 \tag{7.1.21}$$

The draw of Σ given the other parameters can be accomplished by computing the matrix of residuals and doing a standard IW draw.

$$\begin{aligned} &\Sigma \mid \delta, \beta, \gamma, x, y, w \sim IW(\upsilon + n, S + V) \\ &S = \textstyle\sum_{i=1}^{n} \varepsilon_i \varepsilon_i' \\ &\varepsilon_i = \begin{pmatrix} \varepsilon_{1,i} \\ \varepsilon_{2,i} \end{pmatrix} \end{aligned} \tag{7.1.22}$$

The Gibbs sampler defined by (7.1.14) - (7.1.16) is implemented in the function `rivGibbs` in our R package, *bayesm*. In the case of weak instruments and high correlation between the two structural errors (high "endogeneity"), we might expect that the Gibbs Sampler will exhibit the highest autocorrelation due to the "ridge" in the likelihood between β and $\frac{\sigma_{12}}{\sigma_{11}}$. The "hem-stitching" behavior of the Gibbs sampler might induce slower navigation and, hence, autocorrelation in this case.

To illustrate the functioning of the sampler and to gain insight into this model, we consider a simulated example.

$$\begin{aligned} &x = \lambda_x + \delta z + \varepsilon_1 \\ &y = \lambda_y + \beta x + \varepsilon_2 \\ &\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim N\left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right) \end{aligned} \tag{7.1.23}$$

Z is a vector of 200 unif(0,1) draws and $\lambda_x = \lambda_y = 1$. We use relatively diffuse priors for coefficients, A = .04 I, and for the covariance matrix, $\Sigma \sim \mathrm{IW}\left(3, 3I_3\right)$. We first consider the case of "strong instruments" and a high degree of "endogeneity." This is achieved by simulating data with δ=4 and ρ=.8. In this situation, much of the variation in x is "exogenous" due to variation in δz and we would expect to see good performance of the sampler. Figure 7.2 shows plots of the posterior draws for each pair of the three key parmeters, $\left(\beta, \delta, \frac{\sigma_{12}}{\sigma_{11}}\right)$. In addition, we display the sequence of the beta draws on the bottom right. 5000 draws were used with every 5th draw plotted. The dotted lines in each of the scatterplots represent the true values used to simulate the data. The sampler performs very well with a very short "burn-in" period and high numerical efficiency relative to an iid sampler (sqrt(f) = 1.6). As expected, there is a negative correlation between β and $\frac{\sigma_{12}}{\sigma_{11}}$ but it is not too high.

It is well known in the classical econometrics literature that "weak" instruments present problems for the standard asymptotic approximations to the distribution of instrumental variables estimators. This problem is caused by the fact that the IV estimator is a ratio of sample moments. In a likelihood-based procedure such as ours, the weak instruments case creates a situation of near non-identification as indicated in equation (7.1.8). This means there will be a ridge in the likelihood function which will create problems for asymptotic approximations. Since our methods do not rely on asymptotic approximations of any kind, we should obtain more accurate inferences. However, we should point out that since there is a near non-identification problem, the role of the prior will be critical in determining the posterior. In addition, our intuition suggests that the MCMC draws may become more autocorrelated due to the ridge in the likelihood function.

In the simulated example, the weak instruments case corresponds to situations in which the variation in δz is small relative to the total variation in x. If we set $\delta=.1$, we create a situation with extremely weak instruments (the population "R-squared" of the instrument regression is .01). Figure 7.3 shows the distribution of the posterior draws for $\delta=.1$ and $\rho = .8$ which we dub the case of weak instruments and high endogeneity. As expected, there is a very high negative correlation between the draws of β and $\frac{\sigma_{12}}{\sigma_{11}}$. This creates a sampler with higher autocorrelation (sqrt(f) = 10), or a relative numerical efficiency of $1/10^{th}$ of an iid sampler. However, more than 1 million draws of the sampler can be achieved in less than one hour of computing time so that we do not think this is an important issue (figure 7.2 is based on 10,000 draws with every 10 draw plotted).

It is important to note the influence of the prior for the weak instrument case. We note that the posterior distributions of β and $\frac{\sigma_{12}}{\sigma_{11}}$ is centered away from the true values. This is due to the prior on Σ. Even though it is set to barely proper values, the prior is centered on the identity matrix. This means that $\frac{\sigma_{12}}{\sigma_{11}}$ will be "shrunk" toward zero. There is so little information in the likelihood that this prior "shows" through. However, it is important to realize that the posterior of both quantities is very diffuse, revealing the near lack of identification. Figure 7.4 shows the same situation with an improper prior on Σ. Now the posterior is centered closer to the true values but with huge variability. Figure 7.5 shows the marginal posterior distributions for a longer run of 100,000 draws. We can see that the marginal posteriors for β and $\frac{\sigma_{12}}{\sigma_{11}}$ have a mode at the true value and a huge "shoulder" representing high uncertainty.

One might argue that the case of weak instruments and a modest or low degree of endogeneity is more representative of the applications of instrumental variable models. For this reason, we considered the case δ=.1 and ρ=.1. Figure 7.6 shows a plot of the posterior draws for these parameter settings, using the proper but diffuse prior on Σ. There is still a high degree of dependence between β and $\frac{\sigma_{12}}{\sigma_{11}}$, reflecting the low amount of information in the weak instrument case. The sampler is slightly less autocorrelated than the weak instruments/high endogeneity case with a relative numerical efficiency of 1/8th of an iid sample. The posterior distributions are very spread out, properly reflecting the small amount of sample information. In addition, the bivariate posteriors have a decidedly non-elliptically symmetric shape, reminiscent of the shape of a double exponential distribution.

7.2 Structural Models and Endogeneity/Simultaneity

The basic "instrumental variable" model can also be interpreted as resulting from a structural model with attention "limited" to one structural equation. The classic example is the example of supply and demand. If we observed marketing clearing quantities and prices along with a "demand" shifter or "exogeneous" variable, then we can write down a model that is identical to our simple example which we started section 7.1 with.

$$\begin{aligned} q_d &= \alpha_0 + \alpha_1 p + \alpha_2 z + \varepsilon_d \\ q_s &= \beta_0 + \beta_1 p + \varepsilon_s \end{aligned} \tag{7.2.1}$$

Here z is a variable such as advertising or promotion which shifts demand but does not alter supply conditions. If we impose the market clearing condition that $q_d = q_s$, then we can rewrite the first equation as a regression and obtain the "limited" information or instrumental variables model.

(7.2.2)
$$\begin{aligned} p &= \frac{\alpha_0 - \beta_0}{\beta_1 - \alpha_1} + \frac{\alpha_2}{\beta_1 - \alpha_1} z + \frac{\varepsilon_d - \varepsilon_s}{\beta_1 - \alpha_1} \\ q &= \beta_0 + \beta_1 p + \varepsilon_s \end{aligned}$$

Thus, one possible justification of "instrumental" variables approach is that the model arises from some joint or simultaneous model. The joint model imposes the condition of market equilibrium and makes an assumption about the joint distribution of $(\varepsilon_s, \varepsilon_d)$. This allows us to formulate the likelihood which is the joint distribution of (q, p) given z via the Jacobian of the transformation from $(\varepsilon_s, \varepsilon_d)$ to (q, p). This idea can be extended to more "realistic" models of demand in which the demand equation at the aggregate level is the sum of the demands over heterogeneous consumers. Yang et al (2003) tackle this problem with a demand model aggregated from heterogenous logits.

The estimation of simultaneous demand and supply with household heterogeneity is challenging. In a full information analysis, where specific assumptions are made about the nature of competition among firms, the supply-side equation is a complex function of household-level parameters and common error terms, or shocks. To date, likelihood-based approaches have not been developed for models with unobserved heterogeneity specified as random-effects. The reason is that a frequentist (i.e., non-Bayesian) approach to analysis does not view the shocks and random-effects as parameters, and analysis proceeds by first integrating them out of the likelihood function. This integration is computationally demanding when the shocks and random-effects are jointly present, and researchers have instead controlled for consumer heterogeneity by incorporating past purchases, or other exogenous (e.g., demographic) variables, into the model specification. In the context of demand analysis using aggregate level data, it has been shown that it is important to control for both price endogeneity and heterogeneity to avoid potential biases in demand side parameter estimates.

We first set notation by developing the demand and supply models used in our study of consumer brand choice and retailer pricing. Consumers are assumed to make brand choice decisions according to a standard discrete choice model. On the supply side, we develop specifications derived from profit-maximizing assumptions made about manufacturer and retailer behavior.

Demand Model

The disaggregate demand function is specified as a logistic normal regression model. Suppose we observe purchase incidences and choices (y) for a group of individuals (i = 1, …, I) for J brands (j = 0, …, J) in a product category over T time periods (t = 1, …, T). The utility of consumer i for brand j at time t is specified as:

(7.2.3) $$u_{ijt} = \beta_i{}'x_{jt} + \alpha_i p_{jt} + \xi_{jt} + \varepsilon_{ijt}$$

and in the case of no purchase from the J available brands, we denote j = 0 and the associated utility function as:

(7.2.4) $$u_{i0t} = \varepsilon_{i0t}$$

where x_{jt} is a vector with observed product characteristics including brand intercepts, feature and display variables, p_{jt} is the unit price for brand j at time t, β_i and α_i are individual-level response coefficients, ξ_{jt} is an unobserved demand shock for brand j at time period t, and ε_{ijt} is the unobserved error term that is assumed to be uncorrelated with price. We make assumptions on the error terms and response coefficients as the following:

$\varepsilon_{ijt} \sim$ Extreme Value (0,1)

(7.2.5) $\xi_t \sim \text{MVN}(0, \Sigma_d)$

$\theta_i = (\alpha_i, \beta_i')' \sim \text{MVN}(\bar{\theta}, \Sigma_\theta)$

The type I extreme value specification of ε_{ijt} leads to a standard logit choice probability for person i choosing brand j at time t,

(7.2.6) $$\Pr(y_{ijt} = 1) = s_{ijt} = \frac{\exp(V_{ijt})}{1 + \sum_k \exp(V_{ikt})}$$

where $V_{ijt} = \beta_i ' x_{jt} + \alpha_i p_{jt} + \xi_{jt}$

Assuming the sample is representative of the market and households do not make multiple purchases, we obtain market share for brand j at time t as,

(7.2.7) $$s_{jt} = \sum_i s_{ijt} / I$$

If firms use expected demand (s_{jt}) to set price, then price is not exogenously determined. Price and the demand-side error, ξ_t in equation (7.2.3), will be correlated because expected demand, used to set prices, is a function of both. That is, it is not possible to write the joint distribution of prices and demand as a conditional demand distribution and a marginal price distribution. Demand and prices are both functions of consumer price sensitivity and other model parameters. Not accounting for the endogenous nature of price will produce biased estimates of model parameters, including household price sensitivity.

Supply Model – Profit Maximizing Prices

We illustrate the use of Bayesian methods to estimate simultaneous demand and supply models using a simple supply-side model. More complicated models of manufacturer and retailer behavior can be found in Yang, Chen and Allenby (2003). We assume that each manufacturer produces only one product and maximizes the following objective function

(7.2.8) $$\text{Max}_{w_i} \quad \pi_i = Ms_i(w_i - c_{mi})$$

where M is the potential market size, w is the wholesale price, and c_{mi} is manufacturer i's marginal cost. The first order condition for the manufacturers implies,

(7.2.9) $$w - c_m = (HQ)^{-1}(-s)$$

$$\text{where} \quad H_{ik} = \frac{\partial s_i}{\partial p_k} \quad \text{and} \quad Q_{ik} = \frac{\partial p_i}{\partial w_k} \quad (i = 1, \ldots, J, \text{ and } k=1, \ldots, J).$$

Next, we turn to the retailer's pricing strategy. For the purpose of illustration, we only model a single retailer's pricing behavior even though competition among multiple retailers is possible. One simple rule the retailer can use is to simply charge a fixed markup over wholesale price for each brand, resulting the following specification,

(7.2.10) $$p_i = w_i + m_i$$

where m stands for the fixed markup. This pricing strategy implies that $\frac{\partial p_i}{\partial w_i} = 1$ and $\frac{\partial p_i}{\partial w_j} = 0$. Substituting those two conditions into equation (7.2.9), we obtain the following pricing equation,

(7.2.11) $$p = c_m + m - (H)^{-1} s$$

Finally, we can specify the manufacturer and retailer cost c_m as a brand-specific linear function of cost shifters Z, that is

(7.2.12) $$c_{mt} = Z_t{}'\delta_j + \eta_t$$

where η_t is the supply side error which we assume a multivariate normal distribution, that is, $\eta_t \sim MVN(0, \Sigma_s)$, or

(7.2.13) $$p_t = m + Z_t{}'\delta_j - H_t^{-1} s_t + \eta_t$$

The distribution of observed prices is obtained from the distribution of the supply side error by using change-of-variable calculus. This distribution is of non-standard form because equation (7.2.13) is implicit in price – i.e., price appears on both the left and right side of the equal sign in the terms H and s. The distribution of observed prices is obtained by defining a new variable $r = p + H^{-1}s - m - Z\delta$ that is distributed normal with mean 0 and covariance Σ_s. The likelihood for price is obtained in the standard way as the likelihood for r multiplied by the determinant of the Jacobian ($J = \{\partial r_i/\partial p_j\}$).

Bayesian Estimation

Data augmentation is used to facilitate estimation of the model. We introduce household specific coefficients $\{\theta_i\}$ and supply shock realizations $\{\xi_t\}$ as augmented, latent variables, and use them as conditioning arguments in the model hierarchy. The dependent variables are choice (y_{it}) and prices (p_t), and the model can be written in hierarchical form:

$y_{it} \mid p_t, \theta_i, \xi_t, \varepsilon_{it}$	Observed demand
$p_t \mid \{\theta_i\}, \{\xi_t\}, \delta, \eta_t$	Observed prices
$\theta_i \mid \bar{\theta}, \Sigma_\theta$	Heterogeneity
$\xi_t \mid \Sigma_d$	Demand shock
$\eta_t \mid \Sigma_s$	Supply shock
ε_{it}	Extreme value (logit) error

where $\theta_i = (\alpha_i', \beta_i')'$. Observed demand for the i^{th} household is dependent on the household's coefficients (θ_i), the demand shock (ξ_t), the unobserved error (ε_{it}) and the explanatory variables, including prices (equation 6). Observed prices are determined by the set of household coefficients $\{\theta_i\}$, cost shifter coefficients (δ) and the supply shock (η_t), and are set in response to the expected demand across the heterogeneous households (7.2.13).

Household coefficients are specified as random-effects, and the demand and supply shocks specified as normally distributed.

Given the household coefficients $\{\theta_i\}$ and demand shocks $\{\xi_t\}$, the joint distribution of demand and prices is obtained by multiplying the conditional (on prices) demand density by the marginal price density. The marginal price density, given $\{\theta_i\}$ and $\{\xi_t\}$, is derived from the supply-side error term, η_t. The conditional demand density, given $\{\theta_i\}$, $\{\xi_t\}$ and prices, is multinomial with logit probabilities (7.2.6). The joint density of all model parameters is then:

(7.2.14)

$$f(\{\theta_i\}, \{\xi_t\}, \delta, \bar{\theta}, \Sigma_d, \Sigma_s, \Sigma_\theta \,|\, \{y_{it}\}, \{p_t\}) \propto \prod_{t=1}^{T}\prod_{i=1}^{I} \text{prob}(y_{it}\,|\,\theta_i, p_t, \xi_t)\,\pi_1(\xi_t\,|\,\Sigma_d)\,\pi_2(p_t\,|\,\{\theta_i\},\{\xi_t\},\delta,\Sigma_s)\,\pi_3(\theta_i\,|\,\bar{\theta},\Sigma_\theta)\,\pi_4(\delta,\bar{\theta},\Sigma_d,\Sigma_s,\Sigma_\theta)$$

where $\text{prob}(y_{it}|\ \theta_i, p_t, \xi_t)$ is the logit choice probability for household i at time t, π_1 is the density contribution of the demand error ξ_t, π_2 is the density contribution of the observed prices at time t that depend on consumer preferences and price sensitivities $\{\theta_i\}$, demand errors $\{\xi_t\}$, cost variables and coefficients (Z, δ) and the supply-side error (η_t), π_3 is the distribution of heterogeneity and π_4 is the prior distribution on the hyper-parameters.

Given these augmented variables, estimation proceeds using standard distributions for heterogeneity ($\bar{\theta}$ and Σ_θ) and error covariance matrices (Σ_d and Σ_s). Draws of the augmented variables are obtained from the full conditional distribution for ξ_t and θ_t:

(7.2.15)

$$[\xi_t\,|\,*] \propto \prod_i \Pr(\text{choice}_{it}) \times |J_t| \times \left(p_t + H_t^{-1}s_t - m - Z_t\delta \sim \text{MVN}(0,\Sigma_s)\right) \times \left(\xi_t \sim \text{MVN}(0,\Sigma_d)\right)$$

where

$$\Pr(\text{choice}_{it} \neq 0) = \prod_{j=1}^{J} \left(\frac{\exp(V_{ijt})}{1 + \sum_{k=1}^{J} \exp(V_{ikt})} \right)^{I(\text{choice}_{ijt}=1)} \qquad \text{(choose one of the brands)}$$

$$\Pr(\text{choice}_{it} = 0) = \frac{1}{1 + \exp(V_{it})} \qquad \text{(choose outside good)}$$

$$V_{ijt} = \beta_i ' x_{jt} + \alpha_i p_{jt} + \xi_{jt}$$

$$|J_t| \text{ is the Jacobian } = \begin{vmatrix} \frac{\partial r_{1t}}{\partial p_{1t}} & \cdots & \frac{\partial r_{1t}}{\partial p_{jt}} \\ \vdots & \ddots & \vdots \\ \frac{\partial r_{jt}}{\partial p_{1t}} & \cdots & \frac{\partial r_{jt}}{\partial p_{jt}} \end{vmatrix}$$

and $r_t = p_t + H_t^{-1} s_t - m - Z_t \delta$

$$H_t^{-1} s_t = \frac{\sum_{1}^{I} s_{ijt}}{\sum_{i=1}^{I} \alpha_i s_{ijt}(1 - s_{ijt})}$$

where

$$s_{ijt} = \frac{\exp(V_{ijt})}{1 + \sum_{k=1}^{J} \exp(V_{ikt})}$$

$$\begin{aligned} &[\theta_i \mid *] \propto \\ (7.2.16) \quad &\prod_{t=1}^{T} \Pr(\text{choice}_{it}) \times \prod_{t=1}^{T} \left| J_t \right| \times \prod_{t=1}^{T} \left(p_t + H_t^{-1} s_t - m - Z_t \delta \sim N(0, \Sigma_s) \right) \\ &\times \left(\theta_i \sim N(\bar{\theta}, \Sigma_\theta) \right) \end{aligned}$$

Bayesian analysis does not require the integration of the random-effects (θ_i) and demand shocks (ξ_t) to obtain the marginalized, or unconditional likelihood. Such marginalization is difficult to evaluate because the integral is of high dimension and involves highly non-linear functions of the model parameters, including the Jacobian needed to obtain the distribution of observed prices. An advantage of using a Bayesian MCMC estimator is that these variables can be used as conditioning arguments for generating other model parameters – resulting in significant simplification in model estimation.

7.3 Non-Random Marketing Mix Variables

Thus far, we have focused on examples where supply-side behavior has resulted in an "endogeniety" problem in which the marketing mix variables such as price can be functions of unobserved variables. Typically, these ideas are applied to a problem in which the marketing mix variable is being set on a uniform basis for a given market or retailer. However, if firms have access to information regarding the response parameters of a given customer, then the marketing mix variables can be customized to that specific account. This creates a related situation in which the levels of the marketing mix variables are set as a function of the response parameters. Manchanda et al (2004) explore this problem and offer a general approach as well as a specific application to the problem of allocating sales force in the pharmaceutical industry.

A General Framework

Consider the general setting in which the marketing mix or "x" variables are chosen strategically by managers. The basic contribution is to provide a framework for situations in which the mix variables are chosen with some knowledge of the response parameters of the sales response equation.

Sales response models can be thought of as particular specifications of the conditional distribution of sales (y) given the marketing mix x.

$$y_{it} | x_{it}, \beta_{it} \qquad (7.3.1)$$

where i represents the individual customer/account and t represents the time index.

For example, a standard model would be to use the log of sales or the logit of market share and specify a linear regression model,

$$\ln\left(y_{it}\right) = x_{it}^{'}\beta_i + \varepsilon_{it} \quad \varepsilon_{it} \sim \text{Normal}$$

Here the transform of y is specified as conditionally normal with sales response parameters, β_i. Analysis of (7.3.1) is usually conducted under the assumption that the marginal distribution of x is independent of the conditional distribution in (7.3.1). In this case the marginal distribution of x provides no information regarding β_i and the likelihood factors. If $x_{it}|\theta$ is the marginal distribution of x, then the likelihood factors as follows

$$\ell\left(\{\beta_i\},\theta\right) = \prod_{i,t} p\left(y_{it}|x_{it},\beta_i\right)p\left(x_{it}|\theta\right) = \prod_{i,t} p\left(y_{it}|x_{it},\beta_i\right)\prod_{i,t} p\left(x_{it}|\theta\right) \qquad (7.3.2)$$

This likelihood factorization does not occur once the model is changed to build dependence between the marginal distribution of x and the conditional distribution. There are many forms of dependence possible, but in the context of sales response modeling with marketing mix variables a particularly useful form is to make the marginal distribution of x dependent on the response parameters in the conditional model. Thus, our general approach can be summarized as follows

$$\begin{aligned} &y_{it}\,|\,x_{it},\beta_i \\ &x_{it}\,|\,\beta_i,\tau \end{aligned} \qquad (7.3.3)$$

(7.3.3) is a generalization of the models developed by Chamberlain (1980, 1984) and applied in a marketing context by Bronnenberg and Mahajan (2001). Chamberlain considers situations in which the x variables are correlated to random intercepts in a variety of standard linear and logit/probit models. Our random effects apply to all of the response model parameters and we can handle non-standard and non-linear models. However, the basic results of Chamberlain regarding consistency of the conditional modeling approach apply. Unless T grows, any likelihood-based estimator for the *conditional* model will be inconsistent. The severity of this asymptotic bias will depend on model, data and T. For small T, these biases have been documented to be very large.

The general data-augmentation and Metropolis Hasting MCMC approach is ideally suited to exploit the conditional structure of (16). That is, we can alternate between draws of $\beta_i|\tau$ (here we recognize that the { $\beta_{i\cdot}$ } are independent conditional on τ and $\tau\left|\{\beta_i\}\right.$. With some care in the choice of the proposal density, this MCMC approach can handle a very wide range of specific distributional models for both the conditional and marginal distributions in (7.3.3).

To further specify the model in (7.3.3), it is useful to think about the interpretation of the parameters in the β vector. We might postulate that in the marketing mix application, the important quantities are the level of sales given some "normal" settings of x (e.g. the baseline sales) and the derivative of sales wrt various marketing mix variables. In many situations, decision makers are setting marketing mix variables proportional to the baseline level of sales. More sophisticated decision makers might recognize that the effectiveness of the marketing mix is also important in allocation of marketing resources. This means that the specification of the marginal distribution of x should make the level of x a function of the baseline level of sales and the derivatives of sales with respect to the elements of x.

An Application to Detailing Allocation

Manchanda et al (2004) consider an application to the problem of allocation of sales force effort ("detailing") across "customers" (physicians). The data is based on sales calls (termed "details") made to physicians for the purpose of inducing them to prescribe a specific drug. In theory, sales managers should allocate detailing efforts across the many thousands of regularly prescribing physicians so as to equalize the marginal impact of a detail across doctors (assuming equal marginal cost which is a reasonable assumption according to industry sources).

The barrier to implementing optimal allocation of detailing effort is the availability of reliable estimates of the marginal impact of a detail. While individual physician level data is available on the writing of prescriptions from syndicated suppliers such as IMS and Scott-Levin, practitioners do not fit individual physician models due to the paucity of detailing data and extremely noisy coefficient estimates obtained from this data. Instead, practitioners pool data across physicians in various groups, usually on the basis of total drug category volume. Detailing targets are announced for each group. Generally speaking, higher volume physicians receive greater detailing attention. Even if detailing had no effect on prescription behavior, volume-based setting of the detailing independent variable would create a spurious detailing effect in pooled data. In addition to general rules which specify that detailing levels are related to volume, it is clear that individual sales force managers adjust the level of detailing given informal sources of knowledge regarding the physician. This has the net effect of making the levels of detailing a function of baseline volume and, possibly, detailing responsiveness. Thus, the independent variable in our analysis has a level which is related to parameters of the sales response function.

Given the need for physician-specific detailing effects, it might seem natural to apply Bayesian hierarchical models to this problem. Bayesian hierarchical models "solve" the problem of unreliable estimates from individual physician models by a form of "shrinkage" or partial pooling in which information is shared across models. A Bayesian hierarchical model can be viewed as a particular implementation of a random coefficients model.

If detailing levels are functions of sales response parameters, then standard Bayesian hierarchical models will be both biased and inefficient. The inefficiency, which can be very substantial, comes from the fact that the *level* of the independent variable has information about the response coefficients. This information is simply not used by the standard approach. We supplement the sales response function by an explicit model for the distribution of detailing which has a mean related to response coefficients. In our application, given that sales (prescriptions) and detailing are count data, we use an NBD regression as the sales response function and a Poisson distribution for detailing. We demonstrate that this joint model provides much more precise estimates of the effects of detailing and improved predictive performance. Rather than imposing optimality conditions on our model, we estimate the detailing policy function used by the sales managers.

Conditional Modeling Approach

A conditional model for the distribution of prescriptions written given detailing and sampling is the starting point for our analysis. Our data are count data with most observations at less than 10 prescriptions in a given month. Manchanda et al provide some evidence that the distribution of the dependent variable is over-dispersed relative to the Poisson distribution. For this reason, we will adopt the Negative Binomial (NBD) as the base model for the conditional distribution and couple this model with a model of the distribution of coefficients over physicians. The NBD model is flexible in the sense that it can exhibit a wide range of degrees of over-dispersion, allowing the data to resolve this issue. An NBD distribution with mean λ_{it} and over-dispersion parameter α is given by

$$(7.3.4) \qquad \Pr\left(y_{it} = k \middle| \lambda_{it}\right) = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)\Gamma(k+1)}\left(\frac{\alpha}{\alpha + \lambda_{it}}\right)^{\alpha}\left(\frac{\lambda_{it}}{\alpha + \lambda_{it}}\right)^{k}$$

y_{it} is the number of new prescriptions written by physician i in month t. As α goes to infinity, the NBD distribution approaches the popular Poisson distribution.

We adopt the standard log-link function and specify that the log of the mean of the conditional distribution is linear in the parameters.

$$\text{(7.3.5)} \qquad \lambda_{it} = \mathrm{E}\left[y_{it} \middle| x_{it} \right] = \exp\left(x_{it}^{'}\beta_i \right)$$

$$\text{(7.3.6)} \qquad \ln\left(\lambda_{it}\right) = \beta_{0,i} + \beta_{1,i}\mathrm{Det}_{it} + \beta_{2,i}\ln\left(y_{it-1} + d\right)$$

The lagged log-prescriptions term, $\ln\left(y_{it-1} + d\right)$, in equation (7.3.6) allows the effect of detailing to be felt not only in the current period, but in subsequent periods. We add d to the lagged level of prescriptions to remove problems with zeroes in the prescription data. The smaller the number added, the more accurate the Koyck solution is as an approximation. The problem here is that the log of small numbers can be a very large in magnitude which would have the effect of giving the zeroes in the data undue influence on the carry-over coefficients. We choose d=1 as the smallest number which will not create large outliers in the distribution of$\ln\left(y_{it} + d\right)$.

To complete the conditional model, we specify a distribution of coefficients across physicians. This follows a standard hierarchical formulation.

$$\beta_i = \Delta z_i + \nu_i$$

$$\nu_i \sim N\left(0, V_\beta\right)$$

The z vector includes information on the nature of the physicians practice and level of sampling (note that we use PCPs as the base physician type):

$$z^{'} = \left(1, \mathrm{SPE}, \mathrm{OTH}, \mathrm{SAMP}\right)$$

SAMP is the mean (per physician) number of monthly samples divided by 10. This specification of z and the model in (7.3.4) allows there to be a main effect and an interaction for both physician specialty type and sampling. We might expect that physicians with a specialty directly relevant to the therapeutic class of drug X will have a different level of prescription writing. In addition, detailing may be more or less effective depending on the physician's specialty. Sales calls may include the provision of free drug X samples. The effect of sampling is widely debated in the pharmaceutical industry with some arguing that it enhances sales and others arguing for cannibalization as a major effect. Most believe that sampling is of secondary importance to detailing. Sampling is conditional on detailing in the sense that sampling cannot occur without a detail visit. For this reason, we include the average sampling variable in the mean of the hierarchy which creates an interaction term between detailing and sampling.

Beyond the Conditional Model

The company producing drug X does not set detailing levels randomly. The company contracts with a consultant to help optimize the allocation of their national sales force. The consultant recognizes that detailing targets for the sales force should be set at the physician level and not at some higher level such as the sales territory. According to the consultant, detailing is set primarily on the basis of the physician decile computed by IMS for the quarter prior to the annual planning period. IMS assigns each physician to a decile based on the physician's total prescription writing for all drugs in the therapeutic class. These annual targets are then adjusted quarterly based on previous quarter deciles These quarterly adjustments tend to be minor.

Conditional modeling approaches rely on the assumption that the marginal distribution of the independent variables does not depend on the parameters of the conditional distribution specified in (7.3.6). If total category volume is correlated with the parameters of the conditional response model, then this assumption will be violated. We think it is highly likely that physicians who write a large volume of drug X prescriptions regardless of detailing levels (e.g., have high value of the intercept in equation (7.3.6)) will also have higher than average category volume. This means that marginal distribution of detailing will depend at the minimum on the intercept parameter in equation (7.3.6). This dependence is the origin of the spurious correlation that can occur if higher volume physicians are detailed more.

It is also clear that, although detailing targets are set on an annual basis (and revised quarterly), there is much month to month variation in detailing due to factors outside of the control of the sales force managers. In addition, even though detailing targets are set at a high level in the firm, sales force district or territory managers may change the actual level of detailing on the basis of their own specialized knowledge regarding specific physicians. If sales force managers had full knowledge of the functional form and parameters of the detailing response function, then detailing would be allocated so as to equalize the marginal effects across physicians. Given that the current industry practice is not to compute individual physician estimates, it is unreasonable to assume that firms are using a full information optimal allocation approach.

We adopt a specification of the detailing distribution that allows for some partial knowledge of detailing response parameters. A simple but flexible approach would be to assume that detailing is iid with mean set as a function of the long-run response parameters from equation (7.3.6). Note that the average first order autocorrelation for detailing is less than 0.3. Monthly detailing is a count variable with rarely more than 5 details per month. Detailing is modeled as an iid draw from a Poisson distribution with a mean that is a function of baseline sales and the long-run response to detailing.

The iid model of detailing is as given by the Poisson distribution:

(7.3.7) $$\Pr\left(\mathrm{Det}_{it} = m \middle| \eta_i\right) = \frac{\eta_i^m \exp\left(-\eta_i\right)}{m!}$$

The mean of this Poisson distribution is a function of the (approximate long-run) coefficients as:

(7.3.8) $$\ln\left(\eta_i\right) = \gamma_0 + \gamma_1\left(\beta_{0i} \big/ \left(1-\beta_{2i}\right)\right) + \gamma_2\left(\beta_{1i} \big/ \left(1-\beta_{2i}\right)\right)$$

The specification in (7.3.8) allows for a variety of different possibilities. If detailing is set with no knowledge of responsiveness to detailing, then we should expect γ_2 to be zero. On the other hand, if detailing is set with some knowledge of responsiveness to detailing, then we should expect γ_1 and γ_2 to have posteriors massed away from zero. There are a variety of different functional forms for the relationship between the mean level of detailing and the response parameters. We regard our specification as exploratory and as a general linear approximation to some general function of long-run effects.

To summarize, our approach is to enlarge the conditional model by specifying a model for the marginal distribution of detailing. The marginal distribution of detailing depends on conditional response parameters. Using the standard notation for conditional distributions in hierarchical models, the new model can be expressed as follows:

(7.3.9) $y_{it} \left| \mathrm{Det}_{it}, y_{i,t-1}, \beta_i, \alpha \right.$ NBD Regression

(7.3.10) $\mathrm{Det}_{it} \left| \beta_i, \gamma \right.$ Poisson Marginal

This dependence of marginal distribution on the response parameters alters the standard conditional inference structure of hierarchical models. In the standard conditional model given only by (7.3.9), inference about the response parameters, β_i, is based on time-series variation in detailing for the same physician and via similarities between physicians as expressed by the random effects or first-stage prior. However, when (7.3.10) is added to the model, inferences about β_i will change as new information is available from the level of detailing. The marginal model in (7.3.10) implies that the level of detailing is informative about responsiveness and this information is incorporated into the final posterior on β_i. For example, suppose $\gamma_2 < 0$ in equation (7.3.8), then detailing is set so that less responsive physicians are detailed at higher levels and this provides an additional source of information which will be reflected in the β_i estimates. Thus, the full model consisting of equations (7.3.9) and (7.3.10) can deliver improved estimates of physician level parameters by exploiting information in the levels of detailing. The model specified is conditional on β_i. We add the standard heterogeneity distribution on β_i.

Another way of appreciating this modeling approach is to observe that likelihood for β_i has two components – the NBD regression and the Poisson marginal model.

$$(7.3.11) \qquad \ell\left(\{\beta_i\}\right) = \prod_i \prod_t p_{NBD}\left(y_{it} \middle| Det_{it}, \beta_i, \alpha\right) p_{Poisson}\left(Det_{it} \middle| \beta_i, \gamma\right)$$

β_i is identified from both the NBD and the Poisson portions of the model. Examination of the form of the likelihood in (7.3.11) and the mean function in (7.3.8) indicate some potential problems for certain data configurations. In the Poisson portion of the model, elements of the γ vector and the collection of β_i values enter multiplicatively. In terms of the Poisson likelihood, $\left[\gamma_1, \{\beta_{1,i}\}\right]$ and $\left[-\gamma_1, -\{\beta_{1,i}\}\right]$, for example, are observationally equivalent. What identifies the signs of these parameters is the NBD regression. In other words, if the signs of the detailing coefficients are flipped, then the NBD regression fits will suffer, lowering the posterior at that mode. This suggests that in datasets where there is only weak evidence for the effects of detailing (or, in general, any independent variable), then there may exist two modes in the posterior of comparable height. Navigating between these modes can be difficult for Metropolis-Hastings algorithms. To gauge the magnitude of this problem, we simulated a number of different data sets with varying degrees of information regarding the effect of the independent variable. We found that for moderate amounts of information, similar to that encountered in our data, the multi-modality problem was not pronounced. However, for situations with little information regarding the $\{\beta_i\}$, there could potentially be two modes.

Manchanda et al (2004) provide extensive analysis of this model applied to detailing and sales (prescription) data for a specific drug category. They find ample evidence that detailing is set with at least partial knowledge of the sales response parameters. They confirm that detailing is set on the basis of expected volume of prescriptions but also on the responsiveness of sales to detailing activities. Finally, the information in the detailing levels improves the physician level parameter estimates substantially.

Figure 7.1
Illustration of Instrumental Variables Method

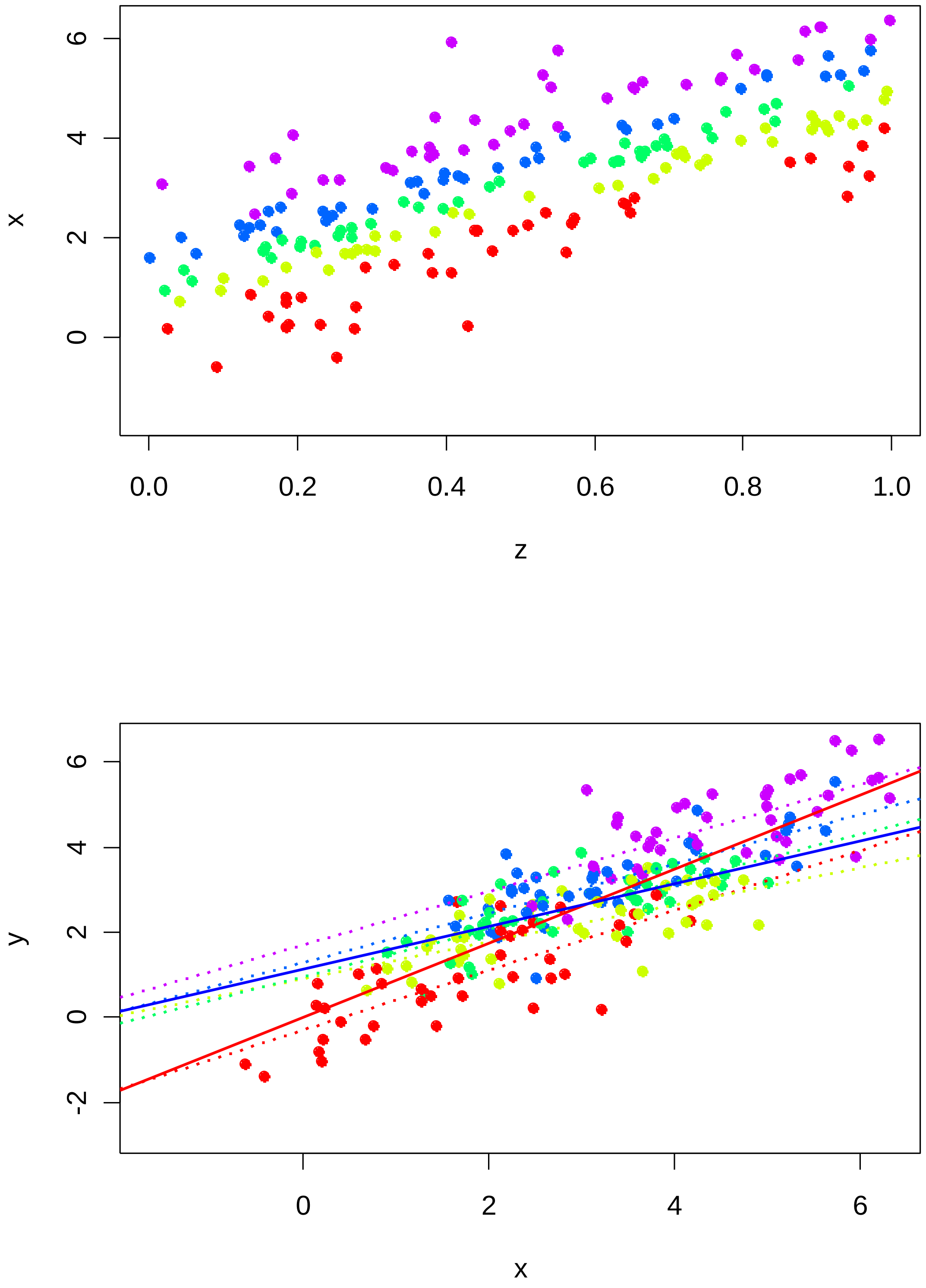


Figure 7.2
MCMC Draws for the Strong Instrument/High Endogneity Case

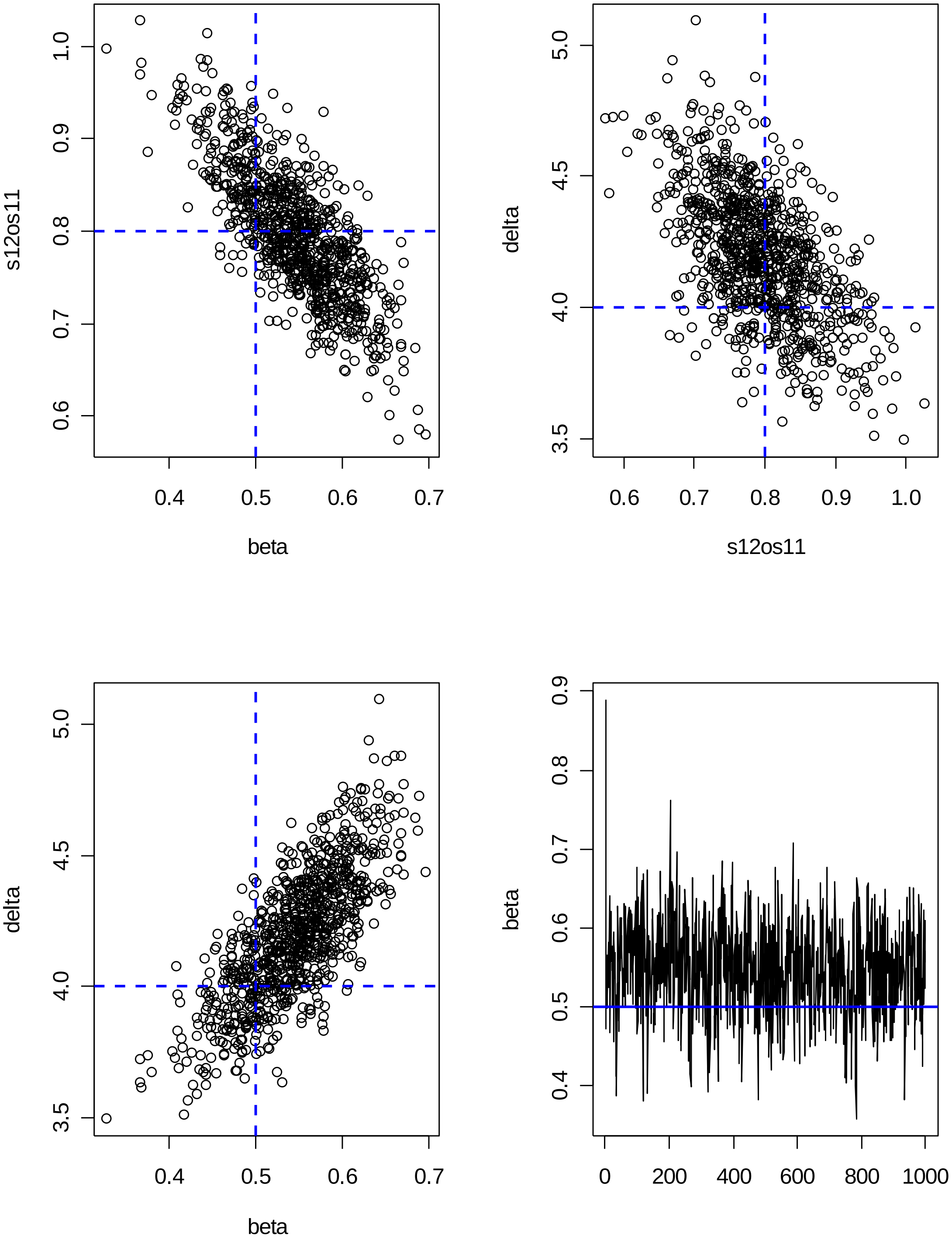


Figure 7.3
MCMC Draws for the Weak Instrument/High Endogeneity Case
Diffuse, but Proper Prior

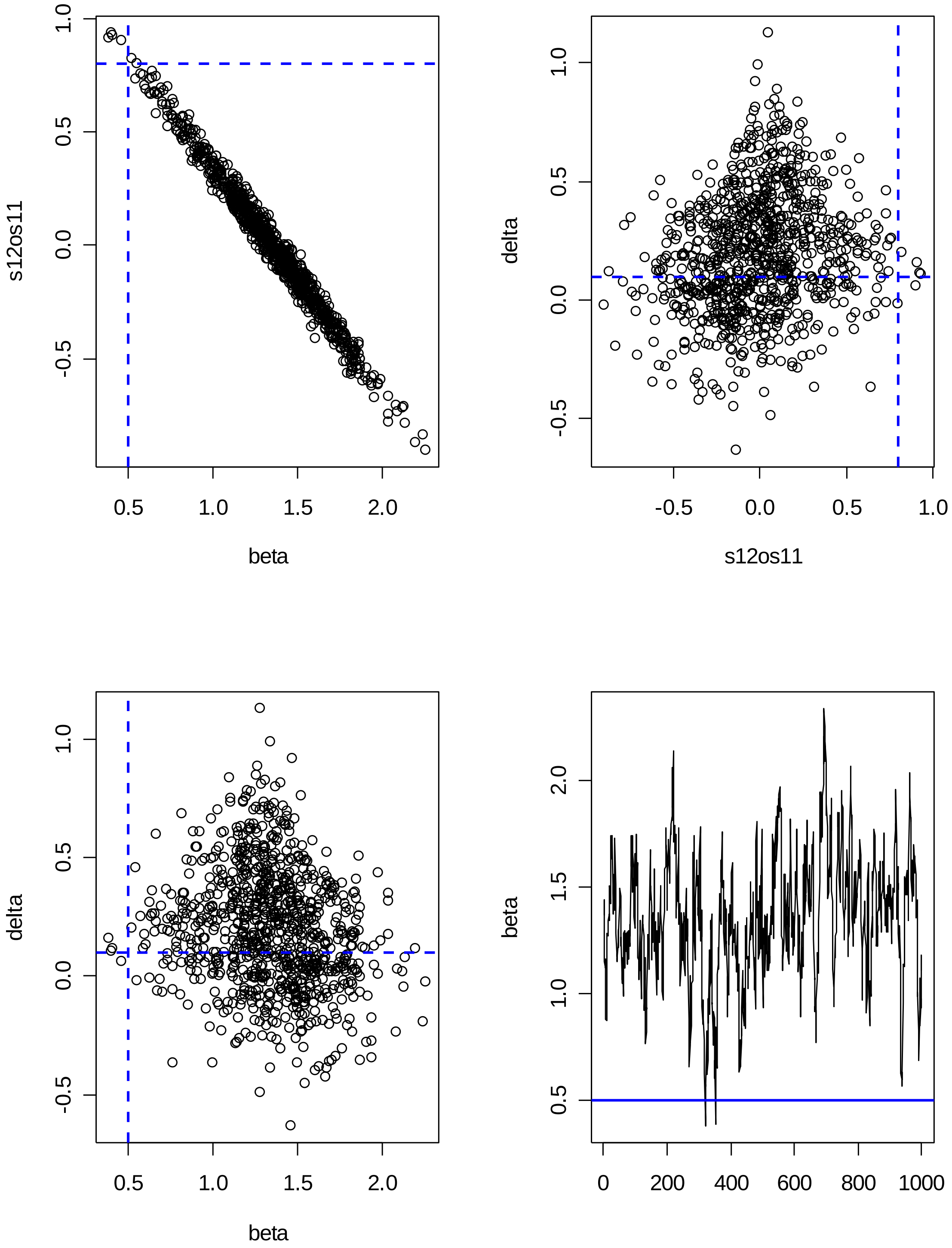


Figure 7.4
MCMC Draws for the Weak Instrument/High Endogneity Case
Improper Prior

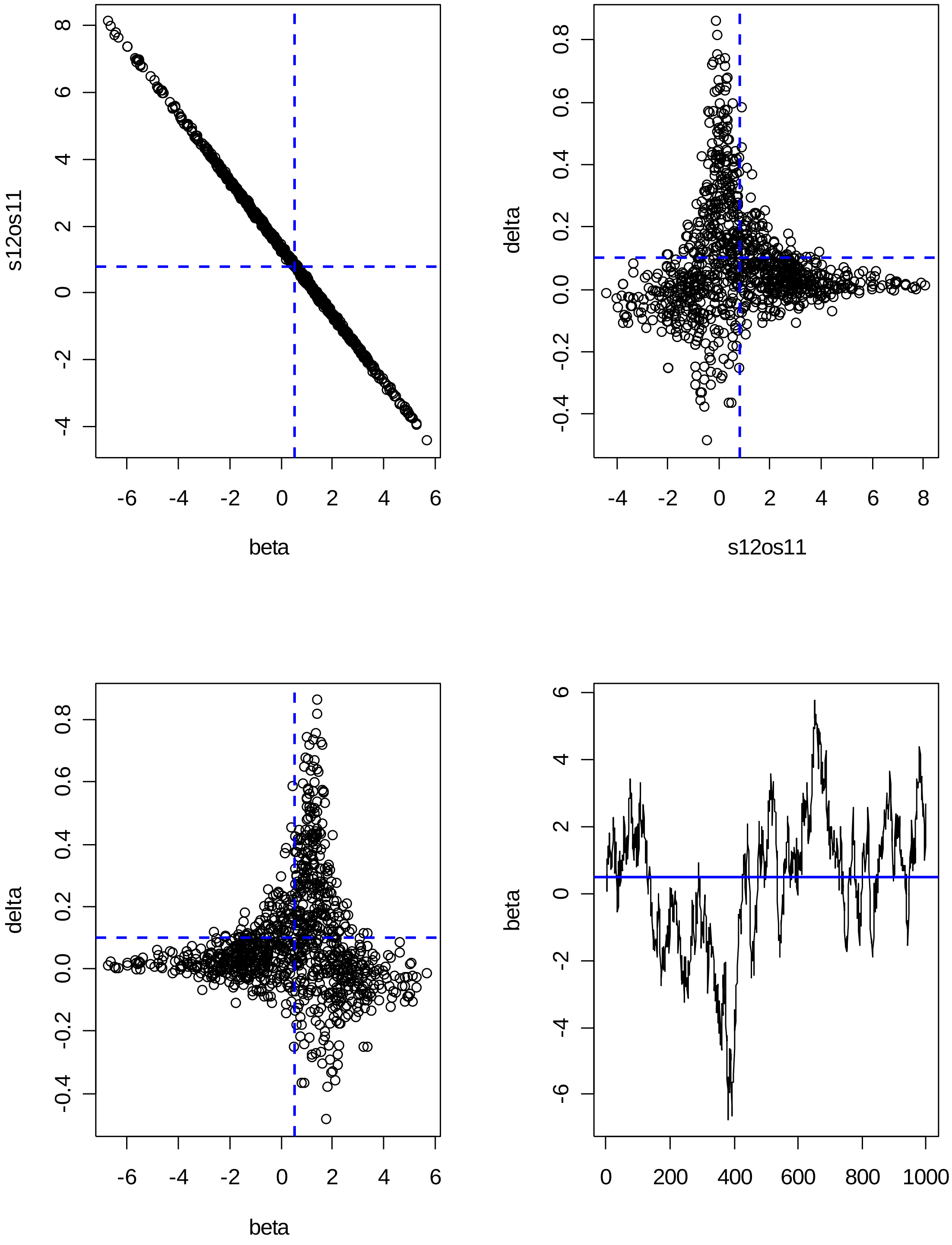


Figure 7.5
Marginal Posteriorrs for Weak Insturments/High Endogeneity Case
Improper Prior, 100,000 draws

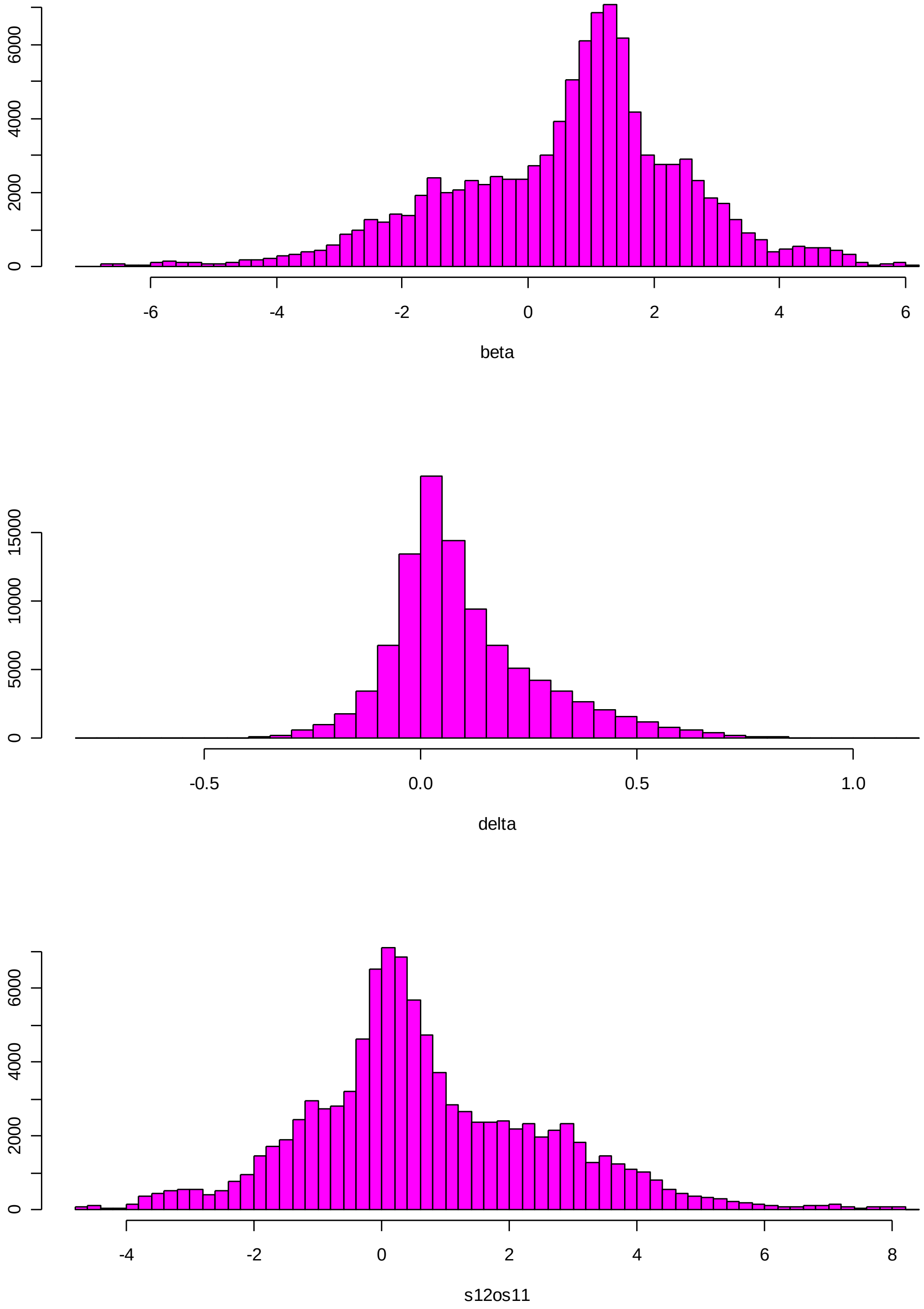
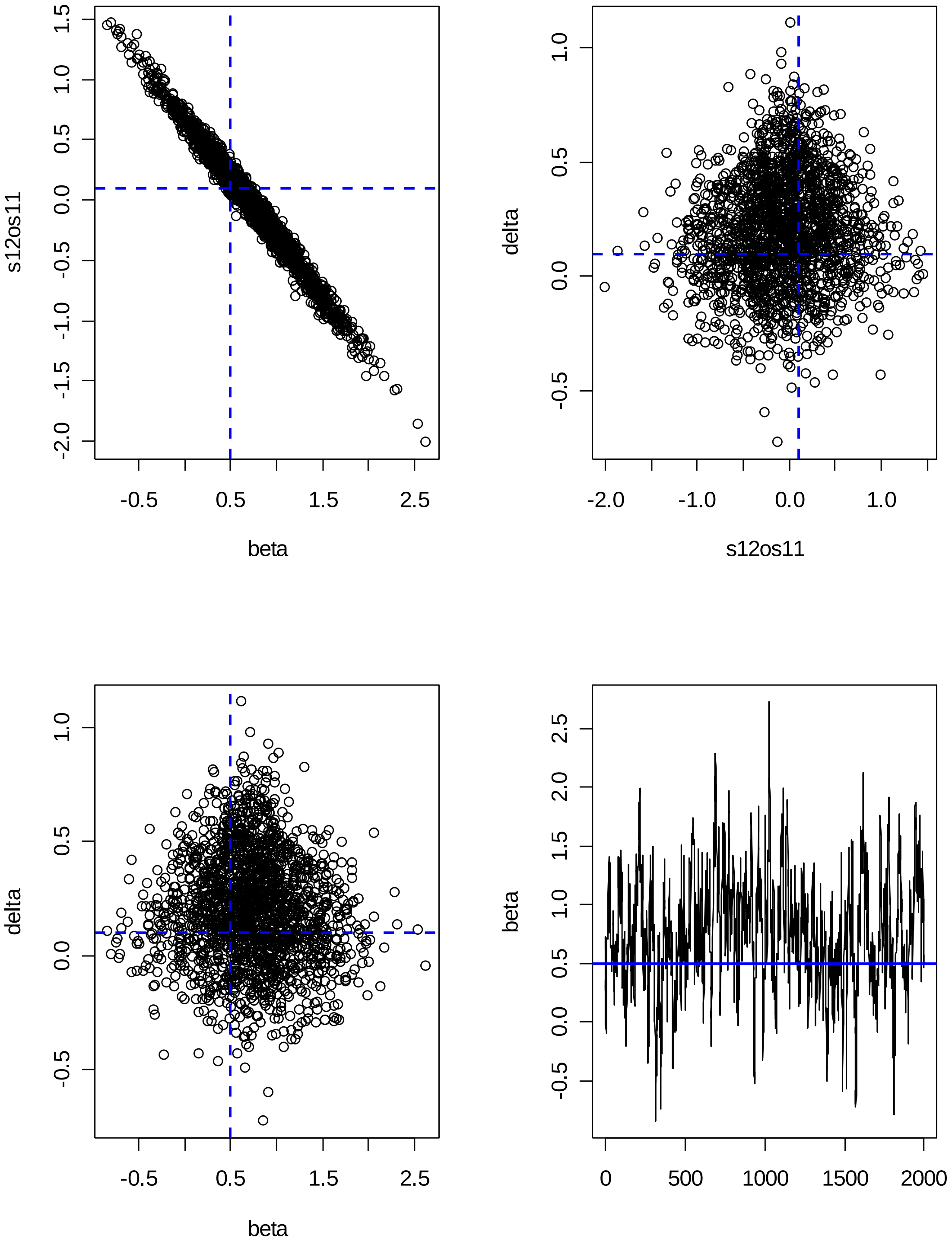


Figure 7.6
MCMC Draws for the Weak Instrument/Low Endogeneity Case


s12os11
beta
delta
s12os11
delta
beta
beta