2011

Modeling Endogenous Treatment Eects with Heterogeneity: A Bayesian Nonparametric Approach

Xuequn Hu
University of South Florida, stewarthu@gmail.com

Modeling Endogenous Treatment Effects with Heterogeneity: A Bayesian Nonparametric Approach

by

Xuequn Hu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Economics
College of Arts and Sciences
University of South Florida

Major Professor: Murat Munkin, Ph.D.
Gabriel Picone, Ph.D.
Yi Deng, Ph.D.
Wonkuk Kim , Ph.D.

Date of Approval:
October 19, 2011

Keywords: Dirichlet Process, DPM, Roy-type Model, Bridge Sampling, MCMC

Table of Contents

List of Tables

List of Figures

Abstract

This dissertation explores the estimation of endogenous treatment effects in the presence of heterogeneous responses. A Bayesian Nonparametric approach is taken to model the heterogeneity in treatment effects. Specifically, I adopt the Dirichlet Process Mixture (DPM) model to capture the heterogeneity and show that DPM often outperforms Finite Mixture Model (FMM) in providing more flexible function forms and thus better model fit. Rather than fixing the number of components in a mixture model, DPM allows the data and prior knowledge to determine the number of components in the data, thus providing an automatic mechanism for model selection.

Two DPM models are presented in this dissertation. The first DPM model is based on a two-equation selection model. A Dirichlet Process (DP) prior is specified on some or all the parameters of the structural equation, and marginal likelihoods are calculated to select the best DPM model. This model is used to study the incentive and selection effects of having prescription drug coverage on total drug expenditures among Medicare beneficiaries.

The second DPM model utilizes a three-equation Roy-type framework to model the observed heterogeneity that arises due to the treatment status, while the unobserved heterogeneity is handled by separate DPM models for the treated and untreated outcomes. This Roy-type DPM model is applied to a data set consisting of $33,081$ independent individuals from the Medical Expenditure Panel Survey (MEPS), and the treatment effects of having private medical insurance on the outpatient expenditures are estimated.

Key Words: Treatment Effects, Endogeneity, Heterogeneity, Finite Mixture Model, Dirichlet

Process Prior, Dirichlet Process Mixture, Roy-type Modeling, Importance Sampling, Bridge Sampling

1 Introduction

The central issue studied in this dissertation is the estimation of causal treatment effects. Treatment effects measure the effectiveness of treatment, or the difference between the outcomes with and without treatment. The treatment in question can be any form of binary intervention. For example, the decision to purchase a medical insurance policy is such a treatment, and we are interested in the difference in medical resource utilization between those with and without coverage. Another example would be to evaluate the effectiveness of a job training program, where we want to find out the difference in earnings between those who have participated in the training program and those who have not.

Formally, let Y_1 denote the outcome with treatment and Y_0 the outcome without treatment. Since an individual cannot be in both states, only one of the two outcomes is observed for that individual, with the other one being the counterfactual. Let D be the binary treatment variable. Define $D = 1$ if the individual is in the treatment group, and $D = 0$ if it is not. The realized outcome Y is thus defined as:

$$Y = \begin{cases} Y_0 & if\ D = 0 \\ Y_1 & if\ D = 1 \end{cases}$$

or

$$Y = DY_1 + (1 - D)Y_0 \tag{1.1}$$

and the treatment effect is defined as $Y_1 - Y_0$.

When evaluating treatment effects, economists often have to work with observational data where

a nonrandom sample is almost never the case, unlike drug trials where patients are randomly assigned into treatment and control groups. This lack of a random sample in observational data is one of the classic sources of endogeneity, and is often at the center of causal inference. A variable is said to be endogenous if it is correlated with the error term. Endogeneity can arise from issues such as the nonrandom samples mentioned above, measurement error, omitted variables, simultaneity, etc.

When estimating endogenous treatment effects, naive Ordinary Least Square (OLS) estimates will be biased. In the example of medical insurance coverage, the OLS estimates will be upwardly biased if those who have the coverage would have spent more than those who do not, even if they had not bought the coverage in the first place. The difference between OLS estimates and the actual treatment effects is called selection bias or selection effect.

To understand why the OLS estimates are biased in this case, define simple linear regression models for both outcomes:

$$Y_1 = X\beta_1 + U_1 \tag{1.2}$$

$$Y_0 = X\beta_0 + U_0 \tag{1.3}$$

where X are covariates, U_0 and U_0 are error terms after controlling X for each outcome, and β_0 and β_1 are the corresponding parameters on the covariates. Substituting equations (1.2) and (1.3) into (1.1), we have:

$$\begin{aligned} Y &= DY_1 + (1-D)Y_0 \\ &= DX\beta_1 + DU_1 + (1-D)X\beta_0 + (1-D)U_0 \\ &= X\beta_0 + D(X\beta_1 - X\beta_0 + U_1 - U_0) + U_0 \end{aligned}$$

Denoting $\alpha = X\beta_0$, $\gamma = X\beta_1 - X\beta_0 + U_1 - U_0$, and $\varepsilon = U_0$, we have a simple linear regression

model:

$$Y = \alpha + \gamma D + \varepsilon \tag{1.4}$$

where γ is the treatment effect estimated from OLS. If D is endogenous, i.e., the covariance between D and the error term ε, $Cov(D, \varepsilon)$, is not 0, then the OLS estimates are biased.

To address selection bias, or endogeneity in a general sense, the Instrumental Variable (IV) approach is often employed by econometricians. Suppose we have an instrument Z that satisfies the traditional IV definitions:

$$Cov(Z, D) \neq 0$$

$$Cov(Z, \varepsilon) = 0$$

then the IV method will produce a consistent estimator of the treatment effect γ:

$$plim\ \hat{\beta}_{iv} = \frac{Cov(Z, Y)}{Cov(Z, D)} = \gamma$$

Separating selection biases from treatment effects empirically, however, is not an easy task. While IV estimation is a powerful tool in economists' toolkit, it is, at the same time, a tool where great care is needed to make sure the instruments are used properly. The validity of instruments, by definition, hinges on the assumption that the instruments are not correlated with the disturbance term in the equation of interest – something that calls for great diligence and scrutiny. In practice, intuition and pretesting are usually employed by researchers to evaluate the validity of chosen instruments. "We can never entirely dispel the clouds of uncertain validity that hang over instrumental variable analyses, but we should chase away what clouds we can" (Murray, 2006).

A much bigger challenge of IV estimation – when used to infer causal treatment effects – comes from the situation where the responses are heterogeneous. Heckman et al. (2006) show that validity of interpreting IV estimates as meaningful treatment effects depends on an important assumption:

the homogeneity of responses after controlling for the observables. If the treatment effect, conditional on X, is heterogeneous, i.e., individuals self-select into treatment based on their idiosyncratic gains, the IV method breaks down in that it does not identify mean treatment effects. In such cases of selection on gains, the treatment is not exogenous because it is correlated with the gains. As a result, the average treatment effects are not identified even with a valid instrument.

To show why this is the case under heterogeneous responses, denote $\overline{\gamma}$ as the mean treatment effect, and η as the idiosyncratic gains:

$$\overline{\gamma} = X\beta_1 - X\beta_0 \tag{1.5}$$

$$\eta = U_1 - U_0 \tag{1.6}$$

Plugging (1.5) and (1.6) into (1.4) we have:

$$Y = \alpha + \overline{\gamma}D + (\varepsilon + \eta D) \tag{1.7}$$

In the case of selection on gains, η and D are not statistically independent, and as a result the conditional expectation of the second term of the error in (1.7), ηD, on the instrument Z, is not zero:

$$E(\eta D \mid Z) = E(\eta \mid D = 1, Z)Pr(D = 1 \mid Z) \tag{1.8}$$

The left hand side of (1.8) is not zero because the first term at the right hand side is not zero due to selection on gains. Heckman et al. (2006) refer to this situation as the presence of *essential heterogeneity*. They show that, under heterogeneous responses, the IV estimator in general does not identify any mean treatment effects. In such cases, different instruments identify different economic parameters, and there is no assurance that the IV estimator will be less biased than OLS.

Specifically, Heckman and Vytlacil (2005) show that the standard linear IV estimator, as well as the average treatment effect (ATE) and the average treatment effect for the treated (ATT), can

be interpreted as a weighted average of marginal treatment effect (MTE), which is defined as the average treatment effect for those that are indifferent to treatment given a fixed amount of unobservable utility of being treated. In order to be able to interpret the IV estimator as a meaningful average treatment effect, we have to ensure that the weights on the MTE are positive. In the presence of response heterogeneity, the weights are guaranteed to be nonnegative if the instrument is the propensity score itself, or a monotonic transformation of the propensity score – with the propensity score defined as the probability of receiving treatment. This underscores the central role that the propensity score plays in the IV estimation for treatment effects, and takes away the robustness to misspecification of the selection equation that is enjoyed by IV estimation under homogeneous responses where any valid instrument, as long as it satisfies the traditional definitions of IV, is able to identify the same underlying economic parameters of interest (Heckman et al., 2006).

It is thus of great importance that the heterogeneity in the data is modeled accurately. In this dissertation, I take a Bayesian Nonparametric approach to model the heterogeneity in treatment effects. Specifically, I adopt the Dirichlet Process Mixture (DPM) model to capture the heterogeneity and show that DPM often outperforms Finite Mixture Model (FMM) in providing more flexible functional forms and thus better model fit. Rather than fixing the number of components in a mixture model, DPM allows the data and prior knowledge to determine the number of components in the data, thus providing an automatic mechanism for model selection.

Two DPM models are presented in this dissertation. The first DPM model is based on a two-equation selection model. The motivation is to identify potential homogeneous components within the data. The homogeneity assumption in responses within each of those components is much more realistic and defensible, thus any instrument, as long as it satisfies the traditional definitions of IV, will be able to identify the same underlying economic parameters of interest. A Dirichlet Process

(DP) prior is specified on some or all the parameters of the structural equation, and marginal likelihoods are calculated to select the best DPM model. The application for this model is to estimate the endogenous treatment effects of having prescription drug coverage on drug expenditures. The sample consists of $2,309$ distinct Medicare beneficiaries who enrolled into Medigap plans from 2003 and 2005. The average treatment effect and average treatment effect for the treated are estimated as $\$1,132$ and $\$858$ respectively. This is consistent with the findings from Fang et al. (2008) where they find substantial advantageous selection effects in the Medigap insurance market.

The two-equation DPM selection model works well for data sets where the treatment group and the control group are not too far apart from each other. If this is not the case, then a model with Roy-type selection is warranted. Thus, in the second DPM model, I utilize a three-equation Roy-type framework to model the observed heterogeneity that arises due to the treatment status, while the unobserved heterogeneity is handled by separate DPM models for the treated and untreated outcomes. I apply the Roy-type DPM model to a data set derived from the 2002-08 Medical Expenditure Panel Survey (MEPS), the latest seven years of data that are available from the MEPS website, and estimate the treatment effects of having private medical insurance on outpatient expenditures. The outpatient expenditures are calculated as the total medical expenditures minus hospital expenditures. With the Roy-type DPM model, I find that there exist strong incentive effects, with the average treatment effect and average treatment effect for the treated estimated as $\$799$ and $\$764$ respectively. One interesting finding is that the selection effects are quite small after heterogeneity is adequately modeled.

The remainder of this dissertation is organized as follows: Chapter 2 provides a short introduction to the Markov Chain Monte Carlo (MCMC) algorithm. Different marginal likelihood calculation methods are also discussed. Chapter 3 reviews the literature of treatment effects, the Dirichlet

Process prior, and the Dirichlet Process Mixture models. Chapter 4 presents the two-equation DPM selection model and its application to the prescription drug expenditure data set. Chapter 5 presents the Roy-Type DPM model and its application to the data set of outpatient expenditures. Concluding remarks and proposals for future research are discussed in Chapter 6.

2 Bayesian Econometrics

2.1 MCMC Algorithm and Gibbs Sampler

In this section, I provide a short introduction to the most commonly used posterior sampling techniques in modern Bayesian inference, namely the MCMC algorithm, and the associated Gibbs sampler.

The basic idea behind the MCMC method is to construct a Markov chain on the space of the parameter of interest. A Markov chain specifies a method for generating a sequence of random variables $\{\theta_1, \theta_2, \cdots, \theta_r \cdots\}$ starting from some initial point θ_0. The sequence is generated by some transition probabilities for the chain to move from one state to another. The transition kernel is specified by choosing the conditional distribution $\theta_{r+1} \mid \theta_r$ or $\theta_r \mid \theta_{r+1}$, where r stands for the current state, and $r + 1$ the next state. The key requirement to construct such a chain is that one must choose the transitional kernel for the chain in such a manner that the stationary distribution of the kernel is indeed the target distribution. A desirable property of a stationary distribution is ergodicity, i.e., the average of an ergodic Markov chain converges to the expectation with regard to the stationary distribution.

The Gibbs sampler is a Markov chain generated by iterating through a set of conditional distributions of the joint posterior distribution. The idea is to break the joint posterior into groups of parameters, so that the conditional distribution of the parameters in each group is analytically

tractable. Denote the complete parameter vector as θ, and the posterior kernel is as follows:

$$p(\theta \mid y) \propto p(\theta)p(y \mid \theta)$$

where $p(\theta)$ is the prior and $p(y \mid \theta)$ is the likelihood. If this kernel density is not of any known analytical form, we cannot draw directly from it. But if we split θ into two blocks:

$$\theta' = (\theta_1, \theta_2)$$

and the following two conditionals are tractable:

$$p(\theta_1 \mid y, \theta_2)$$

$$p(\theta_2 \mid y, \theta_1)$$

then, under very mild conditions, the Gibbs sampler illustrated below will be assured to generate an ergodic chain:

- Start with some initial value for θ_1, denoted as θ_1^0.
- Draw value for θ_2^0 from $p(\theta_2 \mid y, \theta_1^0)$.
- Draw value for θ_1^1 from $p(\theta_1 \mid y, \theta_2^0)$.
- Repeat this process until enough draws are obtained.

2.2 Marginal Likelihood Calculation

It is well known that calculating the marginal likelihood for mixture models is not a trivial task. The Harmonic mean estimator – a natural byproduct of MCMC algorithms and thus the easiest one to calculate – is known to be unstable for mixture models.

Based on Bayes's rule, the marginal likelihood is the normalizing constant of the posterior distribution:

$$p(y \mid M_k) = \frac{p(y \mid \theta_k)p(\theta_k)}{p(\theta_k \mid y)} \tag{2.1}$$

Chib (1995) provides an approximation of equation (2.1):

$$\hat{p}(y \mid M_k) \approx \frac{p(y \mid \theta_k^*)p(\theta_k^*)}{p(\theta_k^* \mid y)} \tag{2.2}$$

where θ_k^* can be an arbitrary point of θ_k. For efficiency, θ_k^* is often selected in high density areas. Chib shows that the posterior ordinate $p(\theta_k^* \mid y)$ can be calculated without much additional programming or computation time within a Gibbs sampler. In the case of Metropolis-Hastings sampling, Chib and Jeliazkov (2001) provide a generalized estimator of the Chib (1995) method.

Chib's method, however, can be biased for calculating marginal likelihoods for mixture models. Since Chib's method estimates the marginal density from the mean of conditional densities, it is essential that the posterior draws mix well and the chain visits the whole parameter space. In the case of FMM, that means the MCMC chain needs to explore all $K!$ modal regions, which is not necessary the case for Gibbs samplers on which Chib's method is based. Indeed, Neal (1999) points out in a simulation that the bias of Chib's method is exactly $-log(K!)$.

Given this potential issue of Chib's method, Fruhwirth-Schnatter (2006) argues that, for mixture models, sampling-based techniques are particularly effective in estimating the marginal likelihood. The key issue is to construct an importance density that meets several criteria:

- The construction of this density should be somewhat automatic, without much manual tuning.
- It should be relatively easy to sample from this importance distribution.
- The importance density constructed should be a rough approximation to the posterior distribution.

Since the posterior distribution for FMM is usually multimodal, a multimodal importance density seems to be appropriate. Fruhwirth-Schnatter (2001) shows that a random subsequence of MCMC draws is a good candidate for importance density, and it can be fully automated and integrated into the algorithm of MCMC sampling.

With a good candidate for the importance density, there are at least three simulation-based estimators for marginal likelihood:

- The Importance Sampling (IS) estimator of the marginal likelihood is defined as follows:

$$p_{IS}(y \mid M_k) = \int \frac{p(y \mid \theta_k)p(\theta_k)}{q(\theta_k)} q(\theta_k) d\theta_k \tag{2.3}$$

  where $p(y \mid \theta_k)$ is the likelihood, $p(\theta_k)$ is the prior density, and $q(\theta_k)$ is the importance density. A sample drawn from the importance density, which itself is constructed as a random subsequence of MCMC draws, is as follows:

$$\theta_k^{(l)} \sim q(\theta_k), l = 1, \cdots, L$$

  The sample analogue of (2.3) can be computed as follows:

$$\hat{p}_{IS}(y \mid M_k) = \frac{1}{L} \sum_{1}^{L} \frac{p(y \mid \theta_k^{(l)})p(\theta_k^{(l)})}{q(\theta_k^{(l)})}$$

  Even with a good importance density, the IS estimator can be unstable: In order for the variance of the IS estimator to be finite, the importance density has to be fatter than the posterior density in the tails – something we cannot ensure for mixture models.

- The Reciprocal Importance Sampling (RI) estimator of the marginal likelihood is defined as follows:

$$p_{RI}(y \mid M_k) = \int \frac{q(\theta_k)}{p(y \mid \theta_k)p(\theta_k)} p(\theta_k \mid y, M_k) d\theta_k \tag{2.4}$$

where $p(\theta_k \mid y, M_k)$ is the normalized posterior density. A random sample from posterior MCMC draws is denoted as follows:

$$\theta_k^{(m)} \sim p(\theta_k \mid y, M_k), m = 1, \cdots, M$$

The sample analogue of (2.4) can be computed as follows:

$$\hat{p}_{RI}(y \mid M_k) = \left(\frac{1}{M} \sum_{1}^{M} \frac{q(\theta_k^{(m)})}{p(y \mid \theta_k^{(m)}) p(\theta_k^{(m)})} \right)^{-1}$$

Similarly, the RI estimator can be unstable due to the possible explosion of its variance – we need to make sure the importance density is thinner than the posterior density in the tails, which again is not something we can ensure for mixture models.

- In light of the potential problem of unbounded variances for IS and RI estimators, Fruhwirth-Schnatter (2006) proposes the Bridge Sampling (BS) estimator based on the bridge sampling method introduced by Meng and Wong (1996). In a sense BS is a generalization of IS in a way that the posterior MCMC draws are combined with the draws from importance sampling. Formally the BS estimator is defined as:

$$p_{BS}(y \mid M_k) = \frac{E_q(\alpha(\theta_k) p(y \mid \theta_k) p(\theta_k))}{E_p(\alpha(\theta_k) q(\theta_k))} \tag{2.5}$$

For the denominator, the expectation is taken with regard to the importance density; and for the numerator, the expectation is with regard to the posterior density.

The sample analogue of (2.5) then can be computed as follows:

$$\hat{p}_{BS}(y \mid M_k) = \frac{L^{-1} \sum_1^L (\alpha(\theta_k^{(l)}) p(y \mid \theta_k^{(l)}) p(\theta_k^{(l)}))}{M^{-1} \sum_1^M (\alpha(\theta_k^{(m)}) q(\theta_k^{(m)}))}$$

It can be seen that the BS estimator uses both importance sampling draws and posterior draws in calculating the marginal likelihood, and Fruhwirth-Schnatter (2004) shows that the variance of the BS estimator is always bounded, thus making the BS estimator a much more stable one.

3 Literature Review

3.1 Treatment Effects

The literature on treatment effects is vast, both in the theoretical and empirical fronts. In particular, the last two decades have seen a great body of work in this area. Sometimes this literature is also referred to as program evaluation, a term that originates from social studies where the effectiveness of some social program is of interest to the researchers. The central problem studied in this literature is to compare the outcomes of two groups of units or individuals, the control and treatment groups. For economists groups can be economic agents such as households, individuals, firms, or nations, and example treatments can be a housing subsidy program, purchase of an insurance policy, enforcement of one regulation, or adoption of some policy, respectively. Of particular interest is the situation where the treatment is binary – people choose to participate in a social program, or decide whether or not to buy insurance coverage. The differences in the outcomes between the treatment unit and the control unit – a measure of the effectiveness of a social program – is referred to as the treatment effect.

The treatment effect literature in statistics comes from randomized experiments, going back to Fisher and Neyman's work on the design of randomized experiments in 1920s. From the very beginning this literature takes the approach of potential outcomes, pioneered by Rubin. In a series of papers, Rubin formulated the approach of estimating the treatment effects in observational studies by comparing the potential outcomes of the same unit, rather than relying only on the observed

outcome (Imbens and Wooldridge, 2009). This model is often referred to as Rubin's model, or the Rubin Causal Model (RCM)(Holland, 1986).

For econometricians a more familiar term is Roy's model, or the switching regression model. Roy (1951) developed a model to explain occupational choice and its impact on earnings, hence the name Roy's model. Econometricians, in order to make some causal inference, are often more concerned with the issue of endogeneity. In labor economics, for example, the effectiveness of job training programs has been the most prominent application where the concern for self-selection is often the focal point of estimation (Ashenfelter, 1978; Ashenfelter and Card, 1985; Card, 1988; LaLonde, 1986).

In a general sense there are two categories of estimators for treatment effects. The first set hinges on the assumption of exogeneity. This assumption goes by many names in the literature: unconfoundedness, ignorability of treatment, or selection on observables. The idea behind this assumption is that, after conditioning on observed covariates X, the treatment does not depend on the outcomes:

$$Y_1, Y_0 \perp D \mid X \tag{3.1}$$

where $\perp$ stands for independence.

Under the assumption of unconfoundedness, there are many parametric and nonparametric estimators for treatment effects: regression, kernel matching, propensity score, or combinations of them, with nonparametric estimation methods being the research focus. Imbens and Wooldridge (2009) provide a complete review of estimating treatment effects under unconfoundedness, and Imbens (2004) reviews the nonparametric estimation of average treatment effects under exogeneity. A desirable feature of the estimators in the first category is that no distribution assumptions are needed for estimation. However, that flexibility comes with a huge burden – the assumption of ignorability,

which is often violated in economic analyses with observational data.

A slightly weaker assumption is the conditional constant effects, or homogeneous responsiveness:

$$(Y_1 - Y_0) \perp D \mid X \tag{3.2}$$

Under this assumption the outcomes are correlated with the treatment variable, but the correlation is assumed to be the same for both outcomes, thus the difference is independent of the treatment, conditional on covariates X. This assumption essentially rules out the Roy-type selection bias, and the traditional Instrumental Variable (IV) estimator is a consistent estimator for the mean treatment effect.

If we cannot make the unconfoundedness or conditional constant effect assumption – this is sometimes referred to as selection on unobservables – then the average treatment effects, in general, are not identified without some distributional assumptions on the error terms. Under the framework of latent utility, Heckman et al. (2001) provide a comprehensive exposition of various treatment effects with selection bias. The selection equation is specified as follows:

$$D^* = Z\theta + U_D \tag{3.3}$$

where D^* is the latent utility of receiving treatment, U_D is the error term, and Z is the instrument. The latent variable D^* determines the treatment decision based on the following rule:

$$D(Z) = I(D^* > 0) = I(Z\theta + U_D > 0)$$

where $I(A)$ is the indicator function that takes the value 1 if the event A is true.

The three equation framework – two outcome equations (1.2) and (1.3), defined in the introduction, and the selection equation defined above, provide a canonical model for defining the four common treatment effects of our interest:

- Average treatment effect (ATE). ATE is defined as the expected gain from treatment for a randomly selected individual:

$$ATE(x) = E(\Delta \mid X = x) = x(\beta_1 - \beta_0)$$
$$ATE = E(\Delta) = \int ATE(X)dF(X) \approx \frac{1}{n}\sum_{i=1}^{n} ATE(x_i) = \overline{x}(\beta_1 - \beta_0)$$

- Treatment effect for the treated (TT). TT is the average gain from treatment for those who receive the treatment:

$$\begin{aligned} TT(x, z, D[z] = 1) &= E(\Delta \mid X = x, Z = z, D[z] = 1) \\ &= x(\beta_1 - \beta_0) + E(U_1 - U_0 \mid U_D + z\theta > 0, X = x, Z = z) \\ &= x(\beta_1 - \beta_0) + E(U_1 - U_0 \mid U_D + z\theta > 0) \end{aligned}$$

$$\begin{aligned} TT &= E(\Delta \mid D[z] = 1) = \int TT(X, Z, D[Z] = 1)dF(X, Z|D[Z] = 1) \\ &\approx \frac{1}{n}\sum_{i=1}^{n} D_i TT(x_i, z_i, D[z_i] = 1) \end{aligned}$$

- Local Average Treatment Effect (LATE). LATE is a concept proposed by Imbens and Angrist (1994). It is defined as the gain from those who are induced to receive treatment with some change of the instrument variables, but otherwise would not. It is the average treatment effect of the "compliers".

$$\begin{aligned} LATE(x, D[z] = 0, D[z'] = 1) &= E(\Delta \mid X = x, D[z] = 0, D[z'] = 1) \\ &= x(\beta_1 - \beta_0) + E(U_1 - U_0 \mid -z'\theta < U_D < -z\theta, X = x) \\ &= x(\beta_1 - \beta_0) + E(U_1 - U_0 \mid -z'\theta < U_D < -z\theta) \end{aligned}$$

$$\begin{aligned} LATE &= E(\Delta \mid D[z] = 0, D[z'] = 1) = \int LATE(X, D[z] = 0, D[z'] = 1)dF(X) \\ &\approx \frac{1}{n}\sum_{i=1}^{n} LATE(x_i, D[z] = 0, D[z'] = 1) \end{aligned}$$

- Marginal Treatment Effect (MTE). MTE is defined as the gains from those who are just indifferent, given a fixed unobservable utility u_D, between receiving treatment or not. Intuitively, MTE can be interpreted as the LATE in the limit, or LATE as a discrete approximation of MTE.

$$\begin{aligned} MTE(x, u_D) &= E(\Delta \mid X = x, U_D = u_D) \\ &= x(\beta_1 - \beta_0) + E(U_1 - U_0 \mid U_D = u_D, X = x) \\ &= x(\beta_1 - \beta_0) + E(U_1 - U_0 \mid U_D = u_D) \end{aligned}$$

$$\begin{aligned} MTE(u_D) &= \int MTE(X, u_D) dF(X) \approx \frac{1}{n} \sum_{i=1}^{n} MTE(x_i, u_D) \\ &= \overline{x}(\beta_1 - \beta_0) + E(U_1 - U_0 \mid U_D = u_D) \end{aligned}$$

These four parameters generally estimate different average treatment effects, and the economic issue in question will dictate which parameter is the most relevant one.

In order to identify and estimate the four treatment effects defined above, some distributional assumptions are needed. If we impose a trivariate normal distribution among the three error terms as follows:

$$\begin{pmatrix} U_D \\ U_1 \\ U_0 \end{pmatrix} \sim \left(0, \begin{bmatrix} 1 & \sigma_{1D} & \sigma_{0D} \\ \sigma_{1D} & \sigma_1^2 & \sigma_{10} \\ \sigma_{0D} & \sigma_{10} & \sigma_0^2 \end{bmatrix} \right)$$

then there are closed form solutions for all four parameters:

$$\begin{aligned} &ATE(x, z, D[z] = 1) = x(\beta_1 - \beta_0) \\ &TT(x, z, D[z] = 1) = x(\beta_1 - \beta_0) + (\sigma_{1D} - \sigma_{0D}) \frac{\phi(z\theta)}{\Phi(z\theta)} \\ &LATE(x, D[z] = 0, D[z'] = 1) = x(\beta_1 - \beta_0) + (\sigma_{1D} - \sigma_{0D}) \frac{\phi(z'\theta) - \phi(z\theta)}{\Phi(z'\theta) - \Phi(z\theta)} \\ &MTE(x, z, D[z] = 1) = x(\beta_1 - \beta_0) + (\sigma_{1D} - \sigma_{0D}) u_D \end{aligned} \tag{3.4}$$

where $\phi(.)$ and $\Phi(.)$ are the probability density function and cumulative probability of the standard normal distribution. From (3.4) we can see that, under the assumptions of (3.1) or (3.2), all four parameters are identical.

This textbook solution is often too restrictive and unrealistic. With Bayesian methods, we can relax the trivariate normal distribution by allowing more flexible function forms, while still enjoying the computational tractability of MCMC algorithms. Especially interesting cases are those where some form of mixture models are utilized to model the joint distribution of the error terms. A few notable mixture models in the Bayesian treatment effect estimation are as follows:

- Albert and Chib (1993) propose a Roy-type model where a multivariate-t distribution is assumed for the joint distributions among error terms:

$$\begin{pmatrix} U_{iD} \\ U_{i1} \\ U_{i0} \end{pmatrix} \sim \begin{pmatrix} \\ 0, \lambda_i^{-1} \begin{bmatrix} 1 & \sigma_{1D} & \sigma_{0D} \\ \sigma_{1D} & \sigma_1^2 & \sigma_{10} \\ \sigma_{0D} & \sigma_{10} & \sigma_0^2 \end{bmatrix} \end{pmatrix}$$

  where

$$\lambda_i \sim Gamma(v/2, v/2)$$

  resulting in a marginal distribution of multivariate-t.

- Chib and Hamilton (2000) use FMM approach to model the joint distribution of outcomes and latent utility (D^*, Y_1, Y_0):

$$\begin{pmatrix} d_i^* \\ y_{i1} \\ y_{i0} \end{pmatrix} \sim \left(\begin{pmatrix} z_i'\theta_j \\ x_i'\beta_{j1} \\ x_i'\beta_{j0} \end{pmatrix}, \begin{bmatrix} 1 & \sigma_{1D} & \sigma_{0D} \\ \sigma_{1D} & \sigma_1^2 & \sigma_{10} \\ \sigma_{0D} & \sigma_{10} & \sigma_0^2 \end{bmatrix}_j \right)$$

  where

$$j = 1, \cdots, m$$

and the number of components m is selected based on the marginal likelihood.

- Munkin and Trivedi (2010) (henceforth MT2010) adopt a FMM approach to model the heterogeneity in treatment effects. That model is a mixture model with a MNP probability equation, in the spirit of Geweke and Keane (2007). There are three equations in the model:

  1. Selection equation:

  $$D_i = W_i\alpha + u_i \tag{3.5}$$

  such that

  $$d_i = I\{D_i \geqslant 0\},$$

  and the marginal distribution of u_i is assumed to be from the standard normal distribution:

  $$u_i \overset{iid}{\sim} N(0,1)$$

  the variance is set to 1 for identification since we do not observe D_i.

  2. Outcome equations:

  $$Y_{ij} = X_i\beta_j + \rho_j d_i + \delta_j u_i + \varepsilon_{ij} \tag{3.6}$$

  with

  $$\varepsilon_{ij} \sim N(0, \sigma_j^2)$$

  where Y_{ij} $(j = 1, 2)$ are two latent variables, and we only observe one of the two outcomes.

  3. Component probability equation:

  $$M_i = V_i\gamma + \xi_i \tag{3.7}$$

  with

  $$\xi_i \overset{iid}{\sim} N(0,1)$$

and the ith individual is assigned to the first component if and only if $M_i \geqslant 0$.

In studying the drug expenditures for elderly people, this model identifies two groups interpreted as the relatively healthy and relatively unhealthy within in the sample, and estimates the incentive and selection effects for each group.

3.2 Dirichlet Process Mixture Model

The Dirichlet Process (DP) prior is the cornerstone of Bayesian nonparametric econometrics. It was first introduced in the landmark paper Ferguson (1973). A DP prior is a prior defined on the space of distribution functions – sometimes loosely referred to as "distribution over distribution". It allows for the support of the prior distribution to include all distributions on the real line.

Formally, a probability measure G is called a Dirichlet Process if, for any measurable partition $(A_1, \cdots, A_k)$ of the parameter space A, we have:

$$(G(A_1), \cdots, G(A_k)) \sim Dirichlet(\alpha G_0(A_1), \cdots, \alpha G_0(A_k))$$

where Dirichlet stands for Dirichlet distribution, which is the multivariate generalization of the Beta distribution. We denote the Dirichlet Process defined above as:

$$G \sim DP(\alpha G_0)$$

where α is the precision parameter and G_0 is the base distribution. It defines the "location" of the DP prior. α controls the precision, or smoothness, of DP prior. A large value of α will produce a G with many "atoms", or clusters.

The benefits of the large support of the DP prior lead to a much wider range of shapes for the posterior distribution by allowing the actual prior on the model parameters to stochastically deviate from the baseline distribution. This is a very appealing feature in modeling unobserved

heterogeneity, as the DP prior provides the flexibility of specifying multimodal, skewed, or fat-tailed priors. As a result, the DP prior finds itself in wide use in many areas such as latent class models, mixture models, clustering, and classification.

3.2.1 Pólya Urn and Stick-breaking Representations of DP

Due to its infinite-dimensional nature, it is not easy to work directly with G. Instead it is often marginalized to get the conditional or joint distributions of the parameters generated from DP. The Pólya Urn Scheme is such a representation of DP.

Assume G_0 is a continuous distribution over colors, and each draw from G_0 is a unique color. Starting with an empty urn, we draw the first ball with a new color from G_0 and put into the urn, and the color of the subsequent ith draw, θ_i, is determined based on the following rules:

- With probability proportional to $i-1$ (i.e., the number of balls currently in the urn), randomly take a ball from the urn, record its color and put it back, and also put another ball with the recorded color into the urn.
- With probability proportional to α, draw a new color from G_0 and put a ball with that new color into the urn.

Blackwell and MacQueen (1973) show this representation results in the following prediction rule:

$$\theta_i \mid G \sim G$$

$$G \sim DP(\alpha G_0)$$

$$\theta_i \mid \theta_1, \cdots, \theta_{i-1} \sim \frac{1}{i-1+\alpha} \sum_{j=1}^{i-1} \delta(\theta_j) + \frac{\alpha}{i-1+\alpha} G_0$$

where δ_{θ_j} is a probability measure concentrated at θ_j.

It is important to note the clustering or grouping implied by the Pólya Urn Scheme. Denote θ_k^* as unique values of all θ_i, N_k the size of kth cluster, and K the number of total clusters. In the Pólya Urn representation above, θ_k^* corresponds to one unique color, θ_i the color of the ith ball, N_k the number of balls with the kth color, and K the total number of unique colors. In terms of clusters, the prediction rule can be rewritten as follows:

$$\theta_i \mid \theta_1, \cdots, \theta_{i-1} \sim \frac{1}{i-1+\alpha} \sum_{k=1}^{K} N_k \delta_{(\theta_k^*)} + \frac{\alpha}{i-1+\alpha} G_0$$

where $\delta_{\theta_k^*}$ is a probability measure concentrated at θ_k^*.

The clustering effect of DP is sometimes characterized as the Chinese Restaurant Process (CRP). Consider a restaurant with infinitely many tables, and a new customer coming to the restaurant prefers to join a table that already has many customers sitting there, but also with probability proportional to α that he or she might start a new table. Some researchers refer to this characterization as the "Collapsed" scheme since it deals with the clusters – or unique parameters θ_k^*, rather than each individual parameter θ_i drawn from DP.

Remarkably, Sethuraman (1994) shows DP can also be represented by an infinite stick-breaking process, which characterizes DP as an infinite mixture model:

$$\theta_i \mid G \sim G$$

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{(\theta_i)}$$

$$\pi_i = V_i \prod_{j=1}^{i-1} (1 - V_j)$$

$$V_i \sim Beta(1, \alpha)$$

where $\delta_{(\theta_i)}$ is a probability measure concentrated at θ_i. Visually, the stick-breaking procedure proceeds as follows:

1. Draw a new value, θ_i, from the baseline distribution G_0.
2. Break what is left of the stick (For the first draw, it would be whole length of the stick) based on Beta distribution $Beta(1, \alpha)$, and that length is the mixing weight of the newly drawn θ_i.
3. Repeat the previous two steps infinitely.

3.2.2 DPM and its Posterior Simulation

In practice, it is the clustering nature of DP that is used in most Bayesian nonparametric or semiparametric applications. A particularly popular model – in the sense of Bayesian hierarchical modeling – is the DP Mixture Model (DPM) proposed by Lo (1984) and Ferguson (1983) :

$$y_i \mid \theta_i, x_i \sim f(\theta_i, x_i)$$

$$\theta_i \mid G \sim G$$

$$G \sim DP(\alpha G_0)$$

where y_i is the dependent variable, x_i the covariates, f the conditional density of y_i on x_i, and θ_i the parameter to be estimated. The baseline distribution G_0 is often assumed to be from the Gaussian family.

Even though the theoretical foundations of DP were laid decades ago, its practical applications were very limited due to the difficulties in sampling from the posterior. With development of Markov chain methods, especially since the introduction of the Gibbs sampler (Gelfand and Smith, 1990), many previous intractable models have become computationally feasible, and DPM is a particular beneficiary of this development.

Broadly speaking, there are two categories of samplers for DPM: marginal sampler and conditional sampler, based on the Pólya urn scheme and the stick-breaking representation respectively. The marginal sampler does not work directly with G by marginalizing over it, thus it needs to calculate the marginal densities of the data, which is not a trivial task if the baseline prior G_0 is not conjugate with the likelihood. Escobar and West (1995) and Escobar and West (1998) present a practical algorithm for the Pólya urn based sampler, but only for conjugate cases. Neal (2000) gives eight algorithms for marginal samplers, especially addressing the nonconjugate cases:

- Algorithm #1 is a literal implementation of the Pólya urn scheme which is quite inefficient.
- Algorithm #2 is the "Collapsed" sampler for conjugate cases.
- Algorithm #5 is based on Metropolis-Hastings sampling.
- Algorithm #8 is a Gibbs sampler based on auxiliary variables, and it is similar to the "No Gap" algorithm by MacEachern and Muller (1998) when the number of auxiliary clusters is set to one. This algorithm is probably the most efficient one for nonconjugate cases. The widely used R package "DPPackage", for example, uses this algorithm as the default one. I implement this algorithm in C++ for the posterior simulation of the Roy-Type DPM model.

The conditional sampler works directly with G, and does not integrate out the mixing weights. Ishwaran and James (2001) present a conditional sampler – or blocked sampler – based on the stick-breaking presentation, and Ishwaran and James (2002) spell out the details of the algorithm of the blocked sampler. Since the blocked sampler does not integrate out the infinite-dimensional G, the issue for conjugacy does not apply for this type of sampler. However, the blocked sampler is based on an approximation of DP – a truncated DP, and it works with a finite version of the stick-breaking process as illustrated in section (3.2.1). To address the concern of truncation, Papaspiliopoulos and

Roberts (2008) propose a retrospective sampler that does not need truncation while keeping the desirable features of the blocked sampler. Separately, Walker (2007) gives a slice sampler for DPM models, addressing the same issue of truncation.

In the field of linear IV models, Conley et al. (2008) use DPM to model the joint distribution of the error terms from structural and selection equations. They show that the resulting semiparametric Bayesian estimator is more precise when compared to Bayesian estimators with normal prior, or to frequentist estimators such as Two-Stage Least Square (TSLS) and Limited Information Maximum Likelihood (LIML) estimators. Particularly, the DP estimator is found to excel in efficiency when the errors are log-normal or the instruments are not strong.

In the area of treatment effects estimation, Chib and Hamilton (2002) propose a semiparametric DPM with Roy-type selection. The joint distribution of the three error terms is specified as follows:

$$\begin{pmatrix} U_{iD} \\ U_{i1} \\ U_{i0} \end{pmatrix} \sim \begin{pmatrix} & \begin{bmatrix} 1 & \sigma_{1D} & \sigma_{0D} \\ \sigma_{1D} & \sigma_1^2 & \sigma_{10} \\ \sigma_{0D} & \sigma_{10} & \sigma_0^2 \end{bmatrix} \\ 0, \lambda_i^{-1} & \end{pmatrix}$$

where

$$\lambda_i \sim G$$

$$G \mid G_0 \sim DP(\alpha G_0)$$

$$G_0 \mid Gamma(v/2, v/2)$$

Here the DP prior is placed on the scale parameter of the joint distribution of the error terms. This model appears to be tractable and gives reasonable results for their chosen application of estimating the wage premium associated with union membership. However, since the DP prior is only placed on the scale parameter, the heterogeneity that is captured by this model specification can be quite limited.

In marketing, Burda et al. (2008) apply a DP prior on key parameters such as price elasticity in a discrete choice model. Their DPM is able to identify a complex multi-modal preference distribution, and provide useful insights into consumers' shopping behavior.

4 Modeling Heterogeneity in Treatment Effects: A Two-Equation DPM Selection Model

4.1 Model Specification

Consider a canonical linear model for estimating treatment effects that allows for possible endogeneity:

$$\begin{aligned} X_i^* &= Z_i'\delta + \eta_i \\ Y_i &= X_i\beta_i + W_i'\gamma_i + \xi_i \end{aligned} \tag{4.1}$$

Here Y_i, X_i, W_i and Z_i are observable variables from N independent individuals ($i = 1, \cdots, N$). Y_i is the dependent variable of our interest, and X_i is the binary variable indicating the status of treatment: $X_i = 1$ if the individual is in the treatment group, and $X_i = 0$ for those in the control group.

The selection model is specified as a Probit, same as in MT2010. The latent variable X_i^* is interpreted as the utility obtained from the treatment, and

$$X_i = \begin{cases} 1 & if\ X_i^* >= 0 \\ 0 & if\ X_i^* < 0 \end{cases}$$

W_i are exogenous covariates. Z_i are the instruments that include all the variables in W_i.

The usual way of modeling the endogeneity is to assume a bivariate normal distribution of η_i

and ξ_i:

$$\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix} \sim \left(0, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \right)$$

An alternative – and often computationally simpler – way of handling endogeneity is to reparameterize ξ_i as a linear function of η_i (Geweke, Gowrisankaran, and Robert, 2003):

$$\xi_i = \rho_i \eta_i + \varepsilon_i \tag{4.2}$$

where the covariance of η_i and ε_i, cov(η_i, ε_i), is assumed to be 0. Plugging equation (4.2) into equation (4.1) we have:

$$\begin{aligned} X_i^* &= Z_i'\delta + \eta_i \\ Y_i &= X_i\beta_i + W_i'\gamma_i + \rho_i\eta_i + \varepsilon_i \end{aligned} \tag{4.3}$$

η_i is assumed to be from normal distribution:

$$\eta_i \sim N(0, 1)$$

Since we do not observe X^*, we normalize the variance of η_i to be 1 for identification. ε_i is also assumed to be from a normal distribution:

$$\varepsilon_i \sim N(\mu_i, \sigma_i^2)$$

Note here the subscript i on β, γ, μ and σ^2 indicates that they are random effects parameters: Coefficients of different individuals can be drawn from different distributions. If panel data are available, this setup naturally leads to a hierarchical model in Bayesian inference. In the case of cross sectional data, some form of clustering is needed to be able to estimate individual parameters, resulting in a mixture model.

Denote $\theta_i = (\beta_i, \gamma_i, \rho_i, \mu_i, \sigma_i^2)$, and augment the data with indicators s_i:

$$\theta_i = \theta_k \text{ if and only if } s_i = k (k = 1, \cdots, K)$$

Here K is the number of components, or latent classes. We can rewrite (4.3) conditional on the indicators:

$$\begin{aligned} X_i^* &= Z_i'\delta + \eta_i \\ Y_i \mid s_i &= X_i\beta_{s_i} + W_i'\gamma_{s_i} + \rho_{s_i}\eta_i + \varepsilon_i \end{aligned} \tag{4.4}$$

and

$$\varepsilon_i \mid s_i \sim N(\mu_{s_i}, \sigma_{s_i}^2)$$

Denote the probability that one individual belongs to kth component as p_k, where $k = 1, \cdots, K$. The vector of probability is denoted as $P = (p_1, \cdots, p_K)$. If we assume the indicators s_i are drawn from the multinomial distribution given the prior P, and P is from the conjugate Dirichlet distribution:

$$\begin{aligned} s_i \mid P &\sim Mutli(p_1, \cdots, p_K) \\ P &\sim Dirichlet(\alpha/K, \cdots, \alpha/K) \end{aligned} \tag{4.5}$$

then this setup results in the commonly used FMM model if K is finite. Note this setup is different from MT2010 and Geweke and Keane (2007) in that no covariates are used to model explicitly the probability of one individual belonging to some latent class.

Interesting enough, Neal (2000) shows that when K goes to infinity this setup leads to a Dirichlet Process Mixture model (DPM), with α as the precision parameter of the Dirichlet Process. This connection demonstrates that DPM can well be an alternative to FMM, especially when we do not have good information on the probable number of components in the data.

In (4.4) a DP prior is put on all parameters of the structural equation. But in many situations it is appropriate to put a DP prior only on some parameters. There are a few benefits of this selectiveness. For one thing, the posterior simulation of DPM is computationally intensive, and if we know a priori that some parameters are indeed from the same distribution, we can save some computation time

by not using a DP prior on them. A more important concern is that the base distribution of a DP prior, G_0, has to be informative. And as the dimension of a DP prior goes up, it will become more difficult to find an appropriate prior for the base distribution. A third reason is that DPM is known to have difficulties dealing with not-well-separated components, thus it is necessary to reduce the dimensionality of a DP prior when appropriate (see Green and Richardson (2001) for a detailed discussion).

Formally the DPM is specified as follows:

$$\begin{aligned} X_i^* &= Z_i'\delta + \eta_i \\ Y_i &= U_i'\psi + V_i'\phi_i + \varepsilon_i \end{aligned} \tag{4.6}$$

where

$$\varepsilon_i \sim N(0, \sigma_i^2)$$

Note a DP prior is put on ϕ_i and ε_i, but not on ψ.

For the reasons mentioned above, I start with a model of putting a DP prior only on the error term of the structure equation. Progressively, a DP prior is placed on more parameters. This progression leads to four models as follows:

- $U = (X, W, \eta), V = 1\iota$. A DP prior is only placed on the error term. This scenario is referred to as *Error* in the model selection section.
- $U = (X, W), V = (1\iota, \eta)$. A DP prior is placed on the error term and the covariance term. This scenario is referred to as *Rho*.
- $U = W, V = (1\iota, X, \eta)$. A DP prior is placed on the error term, covariance term and the coefficient of the endogenous treatment variable. This scenario is referred to as *Both*.

- U is an empty set, and $V = (1\iota, X, \eta, W)$. A DP prior is placed on all parameters of structural equation. This scenario is referred to as *Mixed*.

4.2 MCMC Algorithm

Denote $\theta_i = \{\phi_i, \sigma_i\}, \tilde{Y}_i = Y_i - U_i'\psi = V_i'\phi_i + \varepsilon_i, \Delta_i = (U_i, V_i, Z_i, \psi)$. The estimation algorithm progresses in two steps:

1. Conditional on θ_i draw δ and ψ.

2. Conditional on δ and ψ, calculate $\tilde{Y}_i$, and draw θ_i taking $\tilde{Y}_i$ as "Data".

Step 1 can be implemented with a Gibbs sampler after we augment the data with latent variable X^*; step 2 is implemented with a marginal DP sampler taking $\tilde{Y}$ as given.

The joint density of observed data and latent data for the ith individual, conditional on θ_i, thus $\tilde{Y}_i$, is as follows:

$$\begin{aligned} f(X_i^*, X_i, Y_i|\tilde{Y}_i, \Delta_i) &= p(X_i^*|\tilde{Y}_i, \Delta_i) \\ &\quad \times p(X_i|X_i^*, \tilde{Y}_i, \Delta_i) \\ &\quad \times p(Y_i|X_i, X_i^*, \tilde{Y}_i, \Delta_i) \end{aligned}$$

For the purpose for model exposition, η is separated from either U or V. Denote U_i^-, V_i^-,ϕ_i^-, and ψ^- as follows:

$$U_i = (U_i^-, \eta_i), \psi = (\psi^-, \rho)$$

$$V_i = (V_i^-, \eta_i), \phi_i = (\phi_i^-, \rho_i)$$

If a DP prior is used on the ρ_i, then the likelihood is as follows:

$$\begin{aligned} f(X_i^*, X_i, Y_i|\tilde{Y}_i, \Delta_i) &= \frac{\exp\left(-0.5\left(X_i^* - Z_i\delta\right)^2\right)}{\sqrt{2\pi}}\left[X_i I_{\{X_i^* \geqslant 0\}} + (1 - X_i)\, I_{\{X_i^* < 0\}}\right] \\ &\quad \times \left[\frac{\exp(-0.5\sigma_i^{-2}(Y_i - U_i\psi - V_i^-\phi_i^- - \rho_i\left(X_i^* - Z_i\delta\right))^2}{\sqrt{2\pi\sigma_i^2}}\right] \end{aligned}$$

otherwise:

$$\begin{aligned} f(X_i^*, X_i, Y_i|\tilde{Y}_i, \Delta_i) &= \frac{\exp\left(-0.5\left(X_i^* - Z_i\delta\right)^2\right)}{\sqrt{2\pi}}\left[X_i I_{\{X_i^* \geqslant 0\}} + (1 - X_i)\, I_{\{X_i^* < 0\}}\right] \\ &\quad \times \left[\frac{\exp(-0.5\sigma_i^{-2}(Y_i - U^-{}_i\psi^- - V_i\phi_i - \rho\left(X_i^* - Z_i\delta\right))^2}{\sqrt{2\pi\sigma_i^2}}\right] \end{aligned}$$

The priors are specified as follows:

$$\begin{aligned} &\phi_i \mid G \overset{iid}{\sim} G \\ &G \sim DP(\alpha G_0) \\ &G_0 \equiv MVN(\underline{\phi}, \underline{H}_\phi^{-1}|h)Gamma(a, b) \\ &\underline{\phi} \sim MVN(\underline{\underline{\phi}}, \underline{\underline{H}}_\phi^{-1}) \\ &\psi \sim MVN(\underline{\psi}, \underline{H}_\psi^{-1}) \\ &\delta \sim MVN(\underline{\delta}, \underline{H}_\delta^{-1}) \end{aligned} \tag{4.7}$$

Priors are chosen to reflect reasonable prior information, and also for the convenience of making draws from posterior. The priors on ψ and δ are chosen to be diffuse normal prior for convenience since they will be dominated by the data in any case. The parameters of the DP prior on ϕ , however, deserve more careful thinking. The baseline distribution G_0 is specified as a conjugate Normal-Gamma distribution to allow for convenience of computing the marginal density of the data which is required in the DP algorithm, as well as that of drawing new values in the "Remix" step. When choosing the hyper parameters, we let them be as diffuse as possible, but at the same time some

support is given for the variance to exclude absurd choices of mean parameters. Specifically, the hyper parameters are specified as follows:

- $\underline{H_\phi}$: The precision matrix is a diagonal matrix. That is, a priori, the parameters in ϕ vector are independent. For each parameter, the prior variance is 16 times of the OLS variance.
- $\underline{\phi}$: This is an all-zero vector.
- $\underline{\underline{H}}_\phi$: This is a diffuse diagonal matrix with 0.01 for each diagonal value.
- a: This is the shape parameter of the Gamma distribution for G_0, fixed at 1.
- b: This is the rate parameter of the Gamma distribution for G_0, fixed at 2.
- $\underline{\psi}$: This is an all-zero vector.
- $\underline{H}\psi$: This is a diffuse diagonal matrix with 0.01 for each diagonal value.
- $\underline{\delta}$: This is an all-zero vector.
- $\underline{H}_\delta$: This is a diffuse diagonal matrix with 0.01 for each diagonal value.
- a_α: This is the shape parameter of the Gamma distribution for the DP precision parameter α, fixed at 1.
- b_α: This is the rate parameter of the Gamma distribution for the DP precision parameter α, fixed at 1.

Given the priors specified above, the Gibbs sampler is as follows:

1. The latent vectors X_i^* $(i = 1, ...N)$ are conditionally independent with normal distribution $X_i^* \overset{iid}{\sim} N\left[\overline{X}_i^*, \overline{H_i}^{-1}\right]$

where

$$\begin{aligned}\overline{H_i} &= 1+\rho_i^2\sigma_i^{-2},\\ \overline{X_i^*} &= Z_i\delta+\overline{H_i}^{-1}\rho_i\sigma_i^{-2}(Y_i-U_i\psi-\tilde{Y_i})\end{aligned}$$

Each variable is truncated such that

$$X_i^* \geqslant 0 \text{ if } X_i=1 \text{ and } X_i^*<0 \text{ if } X_i=0.$$

2. The full conditional distribution of δ is $\delta \sim N[\overline{\delta}, \overline{H}_\delta^{-1}]$ where

$$\begin{aligned}\overline{H}_\delta &= \underline{H}_\delta+\sum_{i=1}^{N} Z_i'(1+\rho_i^2\sigma_i^{-2})Z_i\\ \overline{\delta} &= \overline{H}_\delta^{-1}[\underline{H}_\delta\underline{\delta}+\sum_{i=1}^{N}\{Z_i'(1+\rho_i^2\sigma_i^{-2}))X_i^*\\ &\quad -Z_i'\rho_i\sigma_i^{-2}(Y_i-U_i\psi-\tilde{Y_i})\}].\end{aligned}$$

3. The full conditional distribution of ψ is $\psi \sim N\left[\overline{\psi}, \overline{H}_\psi^{-1}\right]$ where

$$\begin{aligned}\overline{H}_\psi &= \underline{H}_\psi+\sum_{i=1}^{N} U_i'\sigma^{-2}U_i\\ \overline{\psi} &= \overline{H}_\psi^{-1}[\underline{H}_\psi\underline{\psi}+\sum_{i=1}^{N}U_i'\sigma^{-2}(Y_i-\tilde{Y_i})]\end{aligned}$$

4. Based on the Polya Urn representation of the Dirichlet Process, we draw θ_i conditional on θ_{-i} (The conditioning on ψ_i and δ_i is assumed here since we take $\tilde{Y_i}$ as "data"), where θ_{-i} is all the θ except the ith.

   (a) Calculate the marginal densities of data after we observe the ith individual. The marginal density is just the product of normalizing constants of likelihood (Univariate normal

distribution) and prior(Normal-Gamma distribution), divided by the normalizing constant of the posterior, which is also a Normal-Gamma distribution due to the conjugacy. Specifically,

$$
\begin{aligned}
P(\tilde{Y_i}) &= \int p(\tilde{Y_i}|\theta_i)G_0(\theta_i|\lambda)d\theta_i \\
&= \int N(\tilde{Y_i}|\phi, H_\phi^{-1})MVN(\underline{\phi}, \underline{H}_\phi^{-1}|h)Gamma(a,b)d\phi dh \\
&= \int \frac{\sqrt{h}}{\sqrt{2\pi}}exp(-\frac{h}{2}(\tilde{Y_i}-\phi V_i)^2)\frac{1}{\sqrt{2\pi}}\frac{b^a\sqrt{Det(\underline{H})}}{\Gamma a}h^{a-\frac{1}{2}}exp(-bh) \\
&\quad exp(-\frac{h}{2}(\phi-\underline{\phi})'\underline{H}(\phi-\underline{\phi}))d\phi dh \\
&= \frac{\sqrt{h}}{\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}\frac{b^a\sqrt{Det(\underline{H})}}{\Gamma a}\int exp(-\frac{h}{2}(\tilde{Y_i}-\phi V_i)^2)h^{a-\frac{1}{2}}exp(-bh) \\
&\quad exp(-\frac{h}{2}(\phi-\underline{\phi})'\underline{H}(\phi-\underline{\phi}))d\phi dh \\
&= \frac{\sqrt{h}}{\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}\frac{b^a\sqrt{Det(\overline{H})}}{\Gamma a}\int h^{\overline{a}-\frac{1}{2}}exp(-\overline{b}h)exp(-\frac{h}{2}(\phi-\overline{\phi})'\overline{H}(\phi-\overline{\phi}))d\phi dh \\
&= \frac{\frac{\sqrt{h}}{\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}\frac{b^a\sqrt{Det(\overline{H})}}{\Gamma a}}{(\frac{\overline{b}^{\overline{a}}\sqrt{Det(\overline{H})}}{\Gamma\overline{a}})^{-1}}\int(\frac{\overline{b}^{\overline{a}}\sqrt{Det(\overline{H})}}{\Gamma\overline{a}})h^{\overline{a}-\frac{1}{2}}exp(-\overline{b}h) \\
&\quad exp(-\frac{h}{2}(\phi-\overline{\phi})'\overline{H}(\phi-\overline{\phi}))d\phi dh \\
&= \frac{\frac{\sqrt{h}}{\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}\frac{b^a\sqrt{Det(\overline{H})}}{\Gamma a}}{(\frac{\overline{b}^{\overline{a}}\sqrt{Det(\overline{H})}}{\Gamma\overline{a}})^{-1}}
\end{aligned}
$$

where

$$
\begin{aligned}
\overline{H} &= \underline{H} + V_i'V_i \\
\overline{\phi} &= \underline{\phi} + V_i'\tilde{Y_i} \\
\overline{a} &= a + 0.5 \\
\overline{b} &= b + 0.5 * (\underline{\phi}'\underline{H}\underline{\phi} + \tilde{Y_i}^2 + \overline{\phi}'\overline{H\phi})
\end{aligned}
$$

(b) Calculate the probabilities for the multinomial draw for θ_i

$$\begin{aligned} q_0 &= \frac{\alpha}{\alpha + (N-1)} \int p(\tilde{Y_i}|\theta_i) G_0(\theta_i|\lambda) d\theta_i \text{ if } k = i \\ q_k &= \frac{1}{\alpha + (N-1)} p(\tilde{Y_i}|\theta_k) \text{ for } k = (1, \ldots, i-1, i+1, \ldots, N) \end{aligned}$$

Normalize the probabilities:

$$\begin{aligned} q_0^* &= \frac{q_0}{q_0 + \sum_{k=1}^{N-1} q_k} \\ q_k^* &= \frac{q_k}{q_0 + \sum_{k=1}^{N-1} q_k} \end{aligned}$$

(c) With probability of q_k^*, set θ_i as θ_k, one of the existing clusters; with probability of q_0^* draw a new θ from the one observation posterior distribution of θ_i, creating a new cluster.

(d) Perform the "Remix" step. Denote θ_j^* as unique values of θ_i, where $j = 1, \ldots, K$ and K is the number of clusters. For each θ_j^* redraw it from the posterior distributions after observing all the individuals in this cluster. The posterior distribution is from a Normal-Gamma distribution with the following parameters:

$$\begin{aligned} \overline{H} &= \underline{H} + V_j' V_j \\ \overline{\phi} &= \underline{\phi} + V_j' \tilde{Y_j} \\ \overline{a} &= a + 0.5 * N_j \\ \overline{b} &= b + 0.5 * (\underline{\phi}' \underline{H} \underline{\phi} + \tilde{Y_j}' \tilde{Y_j} + \overline{\phi}' \overline{H} \overline{\phi}) \end{aligned}$$

where N_j is the number of observations that fall into the jth cluster. We first draw a new precision parameter:

$$h_j \sim Gamma(\overline{a}, \overline{b})$$

then conditional on the newly drawn h we draw the new ϕ_j:

$$\phi_j \sim N(\overline{\phi}, \frac{1}{h_j \overline{H}})$$

(e) Draw the precision parameter of DP α . Escobar and West (1998) shows that, with a Gamma prior $Gamma(a_\alpha, b_\alpha)$ on α and data augmentation, the posterior of α is a mixture of two Gamma distributions:

i. Conditional on the current value of α, draw the auxiliary variable ζ:

$$\zeta \sim Beta(\alpha + 1, N)$$

where N is the number of observations.

ii. Calculate the weights for the two Gamma distributions:

$$\pi_\zeta = \frac{(a_\alpha + I^* - 1)/(N * (b_\alpha - log(\zeta))}{1 + (a_\alpha + I^* - 1)/(N * (b_\alpha - log(\zeta))}$$

where I^* is the number of components

iii. Draw a new α from the following mixture of two Gamma distributions:

$$\alpha \sim \pi_\zeta Gamma(a_\alpha + I^*, b_\alpha - log(\zeta)) + (1 - \pi_\zeta) Gamma(a_\alpha + I^* - 1, b_\alpha - log(\zeta))$$

(f) Draw θ_{N+1}. For each iteration the MCMC algorithm generates a set of θ_is, and often we are more interested in the mixture distribution of θ :

$$p(\theta_{N+1}|Data) = \int p(\theta_{N+1}|\Theta) p(\Theta|data) d\Theta$$

Since we have I^* unique θ_is for each iteration, we can easily draw θ_{N+1} with the following probability:

i. With probability $\frac{\alpha}{\alpha+N}$ draw a new value from the base distribution G_0.

ii. With probability $\frac{1}{\alpha+N}$ take the θ_i that is associated with the ith individual.

4.3 Application

4.3.1 Data Set: Medicare Current Beneficiary Survey

The data used in this application is closely related to the one used in MT2010. The main difference is that we only use observations for seniors who have purchased Medigap plans. The data come from 2003-05 Medicare Current Beneficiary Survey (MCBS).

MCBS provides information on demographics (age, gender, race, education level, family income, marital status, and children), socioeconomics (e.g. income), and an extensive list of health status variables (including chronic conditions, disability, and activity limitations). The "Cost and Use" file provides information on Medicare utilization such as expenditures. MCBS also contains information on supplemental insurance, source of the plan, the premium paid, and coverage of prescription drug expenses, of which some plan attribute information can serve as instrumental variables for modeling plan choice.

A few notes are in order on how the sample is constructed from MCBS data:

- This sample does not include enrollees with Medicaid or other public plans as well as those who hold more than one health plan or who switch plans during the year.
- Due to the panel structure of the data, we have multi-year observations on some individuals. As this is a cross-sectional study, only one observation per individual is retained in sample – following the convention of using the first time the individual is sampled.
- Since 93% of this sample has positive expenditures, I avoid the two-part hurdle structure and focus on the conditional model of positive expenditures.
- As the result of trimming, the final sample contains 2,309 observations from 3 years, the

breakdown being 1,055 from 2003, 642 from 2004, and 612 from 2005.

4.3.2 Description of Variables in the Model

The logarithm of total drug expenditure, LNAAMTTOT, is the dependent variable and drug coverage (COVRX) is the binary endogenous treatment variable. See Table (1) for the full list of variables used in the model.

The covariates consist of six blocks of variables:

- Self-reported health status: VEGOOD, GOOD, FAIR and POOR. EXCELENT is the baseline dummy that is not included in the model.
- Indicators for heart and other chronic conditions: HEARTCOND and OTHCOND.
- Indicators of present and past smoking habits: SMOKNOW and SMOKEVER.
- Geographic dummies: NOREAST, WEST, SOUTH and MSA. The excluded region indicator is Middle West.
- Socioeconomic status: AGE, WHITE, MALE, MARRY, WIDOW, LVALONE, DEGRCV, number of children living CHNLNM, employment status JOBSTAT, and, the natural logarithm of income LNINCOME.
- Two year dummies: YEAR2004 and YEAR2005. YEAR2003 is the excluded baseline level.

In order to identify the endogenous key parameter, and thus separate the treatment effects from selection effects, one needs to have a good instrumental variable. In this dataset the monthly premium, MOAMT, serves as the IV. As the first check the correlation and scatter plot between COVRX and MOAMT show that there is a quite strong correlation between premiums and health insurance

choices. This positive correlation makes intuitive sense: The sample used in this study only includes Medigap enrollees, and those seniors choose to pay higher premiums to have prescription drug coverage. Therefore MOAMT could potentially be a strong instrument.

The harder argument is to prove that the instrument is a valid one, which is something we will never be able to do with certainty. "One can only subject candidate instruments to intuitive, empirical, and theoretical scrutiny to reduce the risk of using invalid instruments." (Murray, 2006). The main rationale behind this choice is that since the premium is a sunk cost after the purchase of the insurance plan, it should not, in theory, have an impact on health-care utilization. Thus the assumption that the premium only affects drug expenditure through plan attributes is reasonable, as a higher premium does not necessarily translate into better drug coverage. Also, we find that current drug expenditure, after controlling the covariates, has little relationship to the premium paid. The residual partial correlation between premium and drug expenditure is indeed quite small. Hence, MOAMT appears to be a suitable instrument.

As mentioned in the introduction, in order to interpret the IV estimate as a valid treatment effect, we need to look at the relationship between propensity score and the instrumental variable. If the IV is a monotonic transformation of the propensity score, then we will have the confidence to state that the IV estimate indeed identifies the treatment effect we plan to estimate.

The propensity score is estimated using a boosted logistic regression. I plot the instrument variable with the propensity score in Figure (1), and we can see that the monotonicity condition roughly holds for almost the entire region.

4.3.3 Sample Summary Statistics

The summary statistics, broken down by the insurance status COVRX, are provided in Table (1). Figure (2) is the kernel density plot of the dependent variable. It is quite clear that the distribution is not normal and shows substantial skewness in the left tail.

Even though the differences in the means for many variables in Table (1) are not substantial for the two groups, the difference for the dependent variable is very large – the difference of the logarithm of drug expenditure is 0.23, and in the dollar terms it is \$337. A similar difference is noted in the levels of income.

To get a better gauge of the differences between treatment and control groups, two estimates of differences for all the covariates are reported here. One is the traditional t-statistic:

$$\Delta X = \frac{\overline{X}_1 - \overline{X}_0}{\sqrt{S_0^2/N_0 + S_1^2/N_1}}$$

where S_0^2 and S_1^2 are the sample variances of X in the treatment and control group. Another one is the differences in averages scaled by the sum of variances:

$$\Delta X = \frac{\overline{X}_1 - \overline{X}_0}{\sqrt{S_0^2 + S_1^2}}$$

Imbens and Wooldridge (2009) argue that the latter provides a better measure of the differences. The reasoning for their preference for the latter is as follows: A larger sample size does not necessarily make the imbalance between the two subsamples bigger, but a larger sample will lead to higher t-statistics, which artificially amplifies the differences. The normalized difference, however, does not change with sample size systematically, thus providing a more reasonable measure of the difference between the subsamples.

As a rule of thumb, the linear regression method, which assumes the control group and treatment

are similar, could be problematic if the scaled difference in mean is greater than one quarter (Imbens and Wooldridge, 2009). Table (2) shows the results of these two estimates. Measured by the normalized differences in means, only the instrumental variable MOMAT and year dummy YEAR2005 are quite different between these two groups, thus the imbalance is not a big concern in this sample.

4.3.4 Model Selection and Fit

Table (4) lists the marginal likelihood values calculated by Importance Sampling (IS), Reciprocal Importance Sampling (RI) and Bridge Sampling (BS) methods. From the table we can see the BS estimator gives the most stable results. The model where a DP prior is put on all the parameters of the expenditure equation appears to be the best model, thus I report the model fit and results for that model only.

One of the objectives of this model is to improve the model fit of MT2010. Figure (3) shows that this DPM model predicts quite well in both tails. Instead of lumping relatively healthy individuals all into one group, DPM is able to devote many components to those individuals, leading to substantial improvement of the model fit in the tails.

4.3.5 Estimation Results

A common feature of mixture models with latent variables, especially when endogeneity is also modeled, is that the MCMC chain often shows a high degree of autocorrelation for parameters on the endogenous variable and covariance term. As a result, it often takes a very large number of iterations for the chain to converge.

The MCMC algorithm produced well-mixed posterior draws and the chain appears to have converged rather quickly. Figure (4) plots the autocorrelation for the key parameters of interest. They

show little autocorrelation, if at all.

The posterior distribution of the intercept clearly shows bi-modality, and that of the variance term and covariance show strong skewness. The parameter on the endogenous variable COVRX, however, is somewhat unimodal. Figure (5) plots the posterior draws for the key parameters of interest and all of them appear to mix well.

The estimation results from the *Mixed* model are reported in this section. Table (5) shows the posterior means and standard deviations, as well as the p-values, for all the parameters of the selection equation. With increasing age, people in our sample are less likely to have drug coverage. Also, whites, males, the married and the widowed are less likely to have it. Living in a metropolitan statistical area, on the other hand, increases the chance of having drug coverage. HEARTCOND and OTHCOND positively affect the coverage probability, as well as all the health statuses: VEGOOD, GOOD, FAIR, and POOR – this makes sense since EXCELLLENT status is the baseline. Surprisingly, more education does not lead to higher drug coverage, but it is not statistically significant. There is a positive relationship between income and drug coverage, but again it is not statistically significant. Smoking history has virtually no effect on drug coverage, while current smokers are less likely to have coverage. Being employed decreases the probability of coverage, which on the surface seems to go against common sense. However, since this sample only includes those who are enrolled into Medigap plans where prescription drug coverage is optional, those employed very likely already have drug coverage from their employers, and thus would not need additional coverage from Medigap plans.

Table (6) shows the results for the expenditure equation. [1] Since we report the aggregate results

[1]Table (3) lists the estimates from different estimators. The estimate from *Roy-type Model with Normal Errors* is calculated according to Heckman et al. (2001); All the rest are directly from the respective coefficient on the endogenous treatment variable.

without decomposing the sample into different components, the standard deviations in general are quite high, resulting in many parameters on the expenditure equations not being statistically significant. Still we can see that coefficients on the intercept and the endogenous variable, as well as most health related variables, are significant.

We do not attempt to identify different components under DPM, nonetheless we are able to detect some patterns of clustering from the posterior draws. The posterior distribution of the intercept term is clearly bi-modal, and if we plot the draws of the parameter on the intercept against those of the key parameters, we can see two discernible clusters. The one with higher average expenditure, which could be interpreted as the relatively unhealthy group, appears to be a tighter cluster in those plots, while the one with lower average expenditure demonstrates much more heterogeneity.

Figure (6), (7), and (8) are the scatter plots of the posterior draws of coefficients on the intercept and the variance, the intercept and the endogenous variable COVRX, and the intercept and the covariance term, respectively.

4.3.6 Calculation and Interpretation of Treatment Effects

As part of the MCMC algorithm, the treatment effect (TE) and the treatment effect for the treated (TT) are calculated for each individual.

For each iteration after burn-in, they are calculated as follows:

$$TE_i = \exp\left(W_i\gamma + 0.5\left(\rho_i^2 + \sigma_i^2\right)\right)\left[\exp\left(\beta_i\right) - 1\right]. \tag{4.8}$$

$$TT_i = \frac{\Phi\left(Z_i\delta + \rho_i\right)}{\Phi\left(Z_i\delta\right)} \exp\left(W_i\gamma + 0.5\left(\rho_i^2 + \sigma_i^2\right)\right)\left[\exp\left(\beta_i\right) - 1\right]. \tag{4.9}$$

where r stands for the rth draw from the posterior, and i denotes the ith individual. The average treatment effect (ATE) and average treatment effect on the treated (ATT) are then calculated for

each draw:

$$ATE^r = \frac{1}{N}\sum_{i=1}^{N} TE_i^r$$

$$ATT^r = \frac{1}{N_1}\sum_{i=1}^{N_1} TT_i^r$$

where N is the full sample size, and N_1 is the sample size of the treatment group.

The posterior mean of ATE is estimated to be $\$1,132$, and that of ATT is $\$858$. The density plots of TE and TT are shown in Figure (9) and (10). Clearly they show substantial heterogeneity in treatment effects across individuals. Figure (11) shows the jitter plot of TT and HEARTCOND and Figure (12) shows the jitter plot of TT and OTHCOND. For better illustration, I use jitter plots instead of scatter plots to avoid the superimposition of data points, since both HEARTCOND and OTHCOND are binary variables. We can see from these plots that the treatment effects are higher for those who have either the heart condition or other chronic conditions.

5 Modeling Heterogeneity in Treatment Effects: A Roy-type DPM Model

5.1 Model Specification

I first present a variant of the Roy-type model without the DPM part. For brevity I suppress the subscripts for each individual before introducing the DPM into the model. The three-equation Roy-type model is formally defined as follows:

$$Y_1 = X\beta_1 + U_1 \tag{5.1}$$

$$Y_0 = X\beta_0 + U_0 \tag{5.2}$$

$$D^* = Z\delta + U_D \tag{5.3}$$

where Y_1 denotes the outcome with treatment and Y_0 the outcome without treatment. U_0 and U_1 are the error terms for each outcome equation. D^* is the latent utility of receiving treatment, U_D is the unobserved utility, and Z is the instrument.

Denote D as the binary treatment variable. Define $D = 1$ if the individual is in the treatment group, and $D = 0$ if not. The latent utility D^* determines the treatment decision based on the following rule:

$$D(Z) = I(D^* > 0) = I(Z\delta + U_D > 0)$$

where $I(A)$ is the indicator function that takes the value 1 if the event A is true. The realized

outcome Y is defined as:

$$Y = \begin{cases} Y_0 & if\ D = 0 \\ Y_1 & if\ D = 1 \end{cases}$$

The usual way of modeling the endogeneity is to assume a trivariate normal distribution for the three error terms, with the variance of U_D normalized to 1 for identification:

$$\begin{pmatrix} U_D \\ U_1 \\ U_0 \end{pmatrix} \sim \begin{pmatrix} 0, \begin{bmatrix} 1 & \sigma_{1D} & \sigma_{0D} \\ \sigma_{1D} & \sigma_1^2 & \sigma_{10} \\ \sigma_{0D} & \sigma_{10} & \sigma_0^2 \end{bmatrix} \end{pmatrix}$$

However one difficulty with this parametrization is to draw from the posterior distribution of the covariance matrix – we can not draw directly from the conjugate Wishart distribution with U_D normalized to 1. To get around this issue, I follow Koop and Poirier (1997) to rewrite the joint distribution above as the product of a conditional and a marginal distribution:

$$P(U_0, U_1, U_D) = P(U_0, U_1|U_D)P(U_D)$$

Since we only observe either Y_0 or Y_1, but never both of them at the same time, the correlation term between them, σ_{10}, does not enter the likelihood and therefore is unidentified. One way of dealing with unidentified parameters is to not update them in the MCMC algorithms; thus we can fix them at specific values for computational convenience. Specifically I follow Deb, Munkin, and Trivedi (2006) to directly include the unobserved utility U_D into the conditional means of Y_0 and Y_1:

$$Y_1 = X\beta_1 + \rho_1 U_D + \varepsilon_1 \tag{5.4}$$

$$Y_0 = X\beta_0 + \rho_0 U_D + \varepsilon_0 \tag{5.5}$$

where $\text{cov}(U_D, \varepsilon_0|X) = 0$, $\text{cov}(U_D, \varepsilon_1|X) = 0$, and $\varepsilon_0 \sim \text{N}(0, \sigma_0^2)$, $\varepsilon_1 \sim \text{N}(0, \sigma_1^2)$, and $U_D \sim \text{N}(0, 1)$. With this parametrization, we can simply fix σ_{10} at 0 and not update it in the MCMC

algorithm. The endogeneity under this setup is modeled by controlling the unobserved heterogeneity in the two regression equations, and the remaining parts of Y_0 and Y_1, ε_0 and ε_1, are therefore assumed to be independent of the treatment.

Now rewrite (5.4) and (5.5) with subscripts for individuals and outcomes (The subscript D on U_D is dropped for notational simplicity):

$$D_i^* = Z_i\delta + U_i \tag{5.6}$$

$$Y_{ij} = X_{ij}\beta_{ij} + \rho_{ij}U_i + \varepsilon_{ij} \tag{5.7}$$

where

$$\varepsilon_{ij} \sim N(0, \sigma_{ij}^2)$$

and $i = 1, \cdots, N$ for N independent individuals, $j = 0, 1$ for the two outcomes.

Note here the subscript i on β, ρ and σ^2 indicates that they are random effect parameters: coefficients of different individuals can be drawn from different distributions. If panel data are available, this setup naturally leads to a hierarchical model in Bayesian inference. In the case of cross section data, some form of clustering is needed to be able to estimate individual parameters, resulting in a mixture model.

To illustrate the mixture model, denote $\theta_{ij} = (\beta_{ij}, \rho_{ij}, \sigma_{ij}^2)$, and augment the data with indicators s_{ij}:

$$\theta_{ij} = \theta_{kj} \text{ if and only if } s_{ij} = k_j (k_j = 1, \cdots, K_j)$$

where K_j is the number of components, or latent classes, for the jth outcome. We can rewrite (5.7) conditional on the indicators:

$$Y_{ij} \mid s_{ij} = X_{ij}\beta_{s_{ij}} + \rho_{s_{ij}}U_i + \varepsilon_{ij} \tag{5.8}$$

where

$$\varepsilon_{ij} \mid s_{ij} \sim N(0, \sigma^2_{s_{ij}})$$

Omitting the subscript j for the moment since it applies equally to each outcome, let p_k denote the probability that an individual belongs to the kth component, where $k = 1, \cdots, K$. The vector of probabilities is denoted as $P = (p_1, \cdots, p_K)$. If we assume the indicators s_i are drawn from the multinomial distribution given the prior P, and P is from the conjugate Dirichlet distribution:

$$\begin{aligned} s_i \mid P &\sim Mutli(p_1, \cdots, p_K) \\ P &\sim Dirichlet(\alpha/K, \cdots, \alpha/K) \end{aligned} \tag{5.9}$$

then this setup results in the commonly used FMM model if K is finite.

Interesting enough, Neal (2000) shows that when K goes to infinity this setup leads to a Dirichlet Process Mixture model (DPM), with α as the precision parameter of the corresponding Dirichlet Process. This connection demonstrates that DPM can well be an alternative to FMM, especially when we do not have good information on the probable number of components in the data. Therefore, we specify separate DP priors on θ_{i0} and θ_{i1}, resulting in rather flexible DPM models for each outcome.

5.2 MCMC Algorithm

The estimation algorithm progresses in three steps:

1. Conditional on θ_{i0} and θ_{i1}, draw D_i^* and δ.

2. Conditional on D_i^* and δ, draw θ_{i0}.

3. Conditional on D_i^* and δ, draw θ_{i1}.

The step 1 can be implemented with a Gibbs sampler after we augment the data with the latent variable D_i^*; step 2 and 3 are implemented with a marginal DP sampler.

Denote $\Delta_{ij} = (X_i, Z_i, \delta, \theta_{ij})$ The joint density of the observed data and the latent variable for the ith individual, conditional on Δ_{ij}, is as follows:

$$\begin{aligned} f(D_i^*, D_i, Y_{i0}, Y_{i1} | \Delta_{i0}, \Delta_{i1}) &= \frac{\exp\left(-0.5\left(D_i^* - Z_i\delta\right)^2\right)}{\sqrt{2\pi}} \\ &\times \left[D_i I_{\{D_i^* \geqslant 0\}} + (1 - D_i) I_{\{D_i^* < 0\}}\right] \\ &\times \left[(1 - D_i) \frac{\exp(-0.5\sigma_{i0}^{-2}(Y_{i0} - X_i\beta_{i0} - \rho_{i0}(D_i^* - Z_i\delta))^2)}{\sqrt{2\pi\sigma_{i0}^2}}\right. \\ &\left. + D_i \frac{\exp(-0.5\sigma_{i1}^{-2}(Y_{i1} - X_i\beta_{i1} - \rho_{i1}(D_i^* - Z_i\delta))^2)}{\sqrt{2\pi\sigma_{i1}^2}}\right] \end{aligned}$$

The priors are given as follows:

$$\begin{aligned} &\theta_{ij} \mid G \overset{iid}{\sim} G_j \\ &G_j \sim DP(\alpha G_0) \\ &G_0 \equiv MNV(\underline{\beta}, \underline{H}_\beta^{-1}) Gamma(a, b) \\ &\underline{\beta} \sim MVN(\underline{\beta}, \underline{\underline{H}}_\beta^{-1}) \\ &\delta \sim MVN(\underline{\delta}, \underline{H}_\delta^{-1}) \end{aligned} \tag{5.10}$$

Priors are chosen to reflect reasonable prior information, and also for the convenience of making draws from the posterior distribution. The priors on δ are chosen to be diffuse normal priors for convenience since they will be dominated by the data in any case. The parameters of the DP prior on β , however, deserve more careful thinking. The baseline distribution G_0 is specified as a semi-conjugate Normal-Gamma distribution. For choosing the hyper parameters, I let them to be as diffuse as possible, but at the same time some support is given for the variance to exclude absurd choices of mean parameters. Specifically, the hyper parameters are specified as follows:

- $\underline{H_\beta}$: The precision matrix is a diagonal matrix. That is, a priori, the parameters in β vector are independent. For each parameter, the prior variance is 30 times of the variance obtained from Heckman's two-step estimator.

- $\underline{\beta}$: This is an all-zero vector.

- $\underline{\underline{H}}_\beta$: This is a diffuse diagonal matrix with 0.01 for each diagonal value.

- a: This is the shape parameter of the Gamma distribution for G_0, fixed at 1.

- b: This is the rate parameter of the Gamma distribution for G_0, fixed at 2.

- $\underline{\delta}$: This is an all-zero vector.

- $\underline{H}_\delta$: This is a diffuse diagonal matrix with 0.01 for each diagonal value.

- a_α: This is the shape parameter of the Gamma distribution for the DP precision parameter α, fixed at 1.

- b_α: This is the rate parameter of the Gamma distribution for the DP precision parameter α, fixed at 1.

Given the priors specified above, the Gibbs sampler is as follows:

1. The latent vectors D_i^* $(i = 1, ...N)$ are conditionally independent with normal distribution:

   $D_i^* \overset{iid}{\sim} N\left[\overline{D^*}_i, \overline{H_i}^{-1}\right]$ where

   $$\begin{aligned} \overline{H_i} &= 1 + (1 - D_i)\rho_{i0}^2\sigma_{i0}^{-2}, \ + D_i\rho_{i1}^2\sigma_{i1}^{-2}, \\ \overline{D^*}_i &= Z_i\delta + (1 - D_i)\overline{H_{i0}}^{-1}\rho_{i0}\sigma_{i0}^{-2}(Y_{i0} - X_{i0}\beta_{i0}) + D_i\overline{H_{i1}}^{-1}\rho_{i1}\sigma_{i1}^{-2}(Y_{i1} - X_{i1}\beta_{i1}) \end{aligned}$$

   Each variable is truncated such that

   $$D_i^* \geqslant 0 \text{ if } D_i = 1 \text{ and } D_i^* < 0 \text{ if } D_i = 0.$$

2. The full conditional distribution of δ is $\delta \sim N[\overline{\delta}, \overline{H}_{\delta}^{-1}]$ where

$$\begin{aligned}
\overline{H}_{\delta} &= \underline{H}_{\delta} + \sum_{i=1}^{N} Z_i'(1 + (1 - D_i)\rho_{i0}^2\sigma_{i0}^{-2} + D_i\rho_{i1}^2\sigma_{i1}^{-2})Z_i \\
\overline{\delta} &= \overline{H}_{\delta}^{-1}[\underline{H}_{\delta}\underline{\delta} + \sum_{i=1}^{N}\{Z_i'(1 + (1 - D_i)\rho_{i0}^2\sigma_{i0}^{-2} + D_i\rho_{i1}^2\sigma_{i1}^{-2}))D_i^* \\
&\quad -Z_i'((1 - D_i)\rho_{i0}\sigma_{i0}^{-2}(Y_{i0} - X_{i0}\beta_{i0}) + D_i\rho_{i1}\sigma_{i1}^{-2}(Y_{i1} - X_{i1}\beta_{i1})\}].
\end{aligned}$$

3. Since the DP prior is separate for each outcome, I suppress the outcome subscript j in the exposition of the DP part of the model. Based on the Polya Urn representation of the Dirichlet Process, we draw θ_i conditional on θ_{-i} (the conditioning on D_i^* and δ_i is assumed here), where θ_{-i} are all the θs except the ith one. The algorithm #8 provided in Neal (2000) is used to draw θ_i, with the number of auxiliary slot, m, as 3.

   (a) Create m auxiliary clusters. For the ith individual, denote K as the number of distinct clusters among all the θs, and K^- as the number of distinct clusters without counting the ith individual itself. If K is equal to K^-, i.e., the ith individual is not associated with a singleton cluster, then create m new clusters by drawing from the baseline distribution G_0. Otherwise designate the singleton cluster associated with the ith individual as the first auxiliary cluster, and create $m - 1$ new clusters by drawing from the baseline distribution.

   (b) Rearrange the clusters so that the first K^- ones are the existing clusters, and last m are the auxiliary ones.

   (c) Calculate the density of Y_i at each cluster $\phi(Y_i, \theta_k)$, where θ_ks are the unique values for the clusters, $k = 1, \ldots, K^- + m$.

(d) Assign the ith individual to a new cluster, k, based on the following multinomial draw:

$$P(s_i = k|Y_i, s_{-i}, \theta_1, \ldots, \theta_k) = \begin{cases} \frac{N_{-i,k}}{N+1-\alpha}\phi(Y_i, \theta_k) & for\ 1 \leq k \leq K^- \\ \frac{\alpha/m}{N+1-\alpha}\phi(Y_i, \theta_k) & for\ K^- < k \leq K^- + m \end{cases}$$

where s_i is the cluster indicator for the ith individual; s_{-i} denotes all the indicators except the ith individual; $N_{-i,k}$ is the number of individuals associated with the kth cluster without counting the ith individual.

(e) Repeat the four steps above for all individuals, $i = 1, \ldots, N$.

(f) Perform the "Remix" step. Denote θ_k^* as unique values of θ_i, where $k = 1, \ldots, K$ and K is the number of clusters. For each θ_k^* redraw from the posterior distribution after observing all the individuals in the cluster. The posterior distribution is from a semi-conjugate Normal-Gamma distribution with the following parameters:

$$\begin{aligned} \overline{a} &= a + 0.5 * N_k \\ \overline{b} &= b + 0.5 * (Y_k - X_k\beta_k)'(Y_k - X_k\beta_k) \\ \overline{H} &= \underline{H} + h_k X_k' X_k \\ \overline{\beta} &= \overline{H}^{-1}(\underline{H}\underline{\beta} + h_k X_k' Y_k) \end{aligned}$$

where N_k is the number of observations that fall into the kth cluster. A mini Gibbs sampler is used to draw new values for h_k and β_k as follows:

$$h_k \sim Gamma(\overline{a}, \overline{b})$$

$$\beta_k \sim N(\overline{\beta}, \overline{H}^{-1})$$

(g) Draw the precision parameter of DP α . Escobar and West (1998) show that, with a Gamma prior $Gamma(a_\alpha, b_\alpha)$ on α and data augmentation, the posterior of α is a mixture of two Gamma distributions:

i. Conditional on the current value of α, draw the auxiliary variable ζ:

$$\zeta \sim Beta(\alpha + 1, N)$$

where N is the number of observations.

ii. Calculate the weights for the two Gamma distributions:

$$\pi_\zeta = \frac{(a_\alpha + I^* - 1)/(N * (b_\alpha - log(\zeta))}{1 + (a_\alpha + I^* - 1)/(N * (b_\alpha - log(\zeta))}$$

where I^* is the number of components

iii. Draw a new α from the following mixture of two Gamma distributions:

$$\alpha \sim \pi_\zeta Gamma(a_\alpha + I^*, b_\alpha - log(\zeta)) + (1 - \pi_\zeta) Gamma(a_\alpha + I^* - 1, b_\alpha - log(\zeta))$$

(h) Draw θ_{N+1}. For each iteration the MCMC algorithm generates a set of θ_is, and often we are more interested in the mixture distribution of θ :

$$p(\theta_{N+1}|Data) = \int p(\theta_{N+1}|\Theta)p(\Theta|data)d\Theta$$

Since we have I^* unique θ_is for each iteration, we can easily draw θ_{N+1} with the following probability:

i. With probability $\frac{\alpha}{\alpha+N}$ draw a new value from the base distribution G_0.

ii. With probability $\frac{1}{\alpha+N}$ take the θ_i that is associated with the ith individual.

5.3 Application

5.3.1 Data Set: Medicare Expenditure Panel Survey

The data set used in this application is constructed from the 2002-08 Medical Expenditure Panel Survey (MEPS), the latest seven years of data that are available from the MEPS website. We are

interested in estimating the pure incentive effects of having private medical insurance on the outpatient expenditures. Here the outpatient expenditures are calculated as the total medical expenditures minus hospital expenditures. The rationale behind this is that hospital expenditures are often of an involuntary nature, whereas outpatient expenditures are those involved with deliberate decisions, with the status of medical insurance expected to be part of the decision making.

MEPS provides information on demographics (age, gender, race, education level, family income, marital status, and children), socioeconomics (e.g. income), and an extensive list of health status variables (including chronic conditions, disability, and activity limitations). MEPS also contains information on the status of insurance: Does one individual have any type of medical insurance when surveyed? And if yes, what type of insurance plan: public or private.

This sample does not include any individuals enrolled in any public medical insurance plans such as Medicare or Medicaid. I restrict our sample to those who have jobs, but are not self-employed. For the same reason we only include individuals between 25 and 65 years of age. Due to the panel structure of the data, we have multi-year observations for some individuals. As this is a cross-sectional study, only one observation per individual is retained in the sample, following the convention of using the observation when the individual is sampled for the first time. Since more than 80% of this sample has positive expenditures, I avoid the two-part hurdle structure and focus only on the conditional model of positive expenditures. As the result of trimming, the final sample contains $33,081$ individuals, with the majority of them, $28,863$, having some type of private medical insurance.

5.3.2 Description of Variables in the Model

The logarithm of total drug expenditure, LOUTEXP, is the dependent variable and the private medical insurance status (PRVEV) is the binary endogenous treatment variable. See Table (7) for the full list of variables used in the model.

The covariates consist of four blocks of variables. Self-reported health status: VEGOOD, GOOD, and FAIRPOOR (Either FAIR or POOR). EXCELENT is the baseline dummy that is not included in the model. Also included in the model is the indicator for having a limitation of physical activities PHYSLIM. Geographic dummies: NOREAST, WEST, and SOUTH. The excluded region indicator is MIDDLEWEST. MSA indicates whether the individual lives in a metropolitan statistical area. Socioeconomic status: AGEX (age divided by 10), EDUCYR (years of education), FAMSZEYR (family size), MALE, MARRY, BLACK, HISP (Hispanic) and INCOME (income in thousand dollars). Six year dummies: YEAR2003 to YEAR2008. YEAR2002 is the excluded baseline level.

In order to identify the endogenous parameter, and thus separate the treatment effects from the selection effects, one needs to have a good instrumental variable. In this dataset the firm size, FIRMSIZE (firm size divided by 10), serves as the IV. The first check of the correlation and scatter plot between PRVEV and FIRMSIZE shows that there is a quite strong correlation between firm size and health insurance choices. This positive correlation makes intuitive sense: larger firms tend to provide employer-sponsored medical insurance plans as part of the benefits.

The harder argument is that the chosen instrument is a valid one, which is something we will never be able to do with certainty. "One can only subject candidate instruments to intuitive, empirical, and theoretical scrutiny to reduce the risk of using invalid instruments." (Murray, 2006). In our case, it seems reasonable to assume that the firm size should not have a direct impact on one

individual's medical expenditures.

5.3.3 Sample Summary Statistics

The summary statistics, broken down by the insurance status PRVEV, are provided in Table (7). Figure (13) is the kernel density plot of the dependent variable. It is quite clear that the distribution is not normal and shows some skewness in the left tail.

A quick look at the summary statistics reveals that the control group and the treatment group are quite different. The treatment group is much bigger than the control group. It has $28,862$ individuals relative to $4,219$ individuals in the control group. More importantly, the expenditures and key covariates also differ substantially. On average those without insurance spend $\$1,230$ on outpatient care, while those with insurance spend $\$2,492$, more than double the amount of the former. Income level and education also indicate that they belong to different socioeconomic groups: Those with insurance, on average, have two and a half more years of education, and their income is almost double of those without insurance ($\$76,200$ vs. $\$39,890$).

To get a more precise gauge of the differences between the treatment and the control groups, two estimates of differences for all the covariates are reported here. One is the traditional t-statistic:

$$\Delta X = \frac{\overline{D}_1 - \overline{D}_0}{\sqrt{S_0^2/N_0 + S_1^2/N_1}}$$

where S_0^2 and S_1^2 are the sample variances of X in the treatment and control group. The other is the differences in averages scaled by the sum of variances:

$$\Delta X = \frac{\overline{D}_1 - \overline{D}_0}{\sqrt{S_0^2 + S_1^2}}$$

Imbens and Wooldridge (2009) argue that the latter provides a better measure. The reasoning for their preference for the latter is as follows: A larger sample size does not necessarily make the

imbalance between the two subsamples bigger, but a larger sample will lead to higher t-statistics, which artificially amplifies the differences. The normalized difference, however, does not change with sample size systematically, thus providing a more reasonable measure of the difference between the subsamples.

As a rule of thumb, the linear regression method, which assumes the control group and treatment are similar, could be problematic if the scaled difference in means is greater than one quarter (Imbens and Wooldridge, 2009). Table (8) shows the results of these two estimates. The normalized differences in means clearly show those two subsamples are quite different from each other, especially in key covariates such as income, education, firm size, etc., and it is thus reasonable to assume that individuals in those two groups would behave differently in insurance purchase decision and health care resources utilization, which in turn calls for a Roy-type model where essential heterogeneity is explicitly modeled.

5.3.4 Model Selection and Fit

It is well known that calculating the marginal likelihood for mixture models is not a trivial task. The Harmonic mean estimator, the easiest one to calculate since it is a natural byproduct of MCMC algorithms, is known to be unstable for FMM. Fruhwirth-Schnatter (2006) show that, for mixture models, sampling-based techniques are particularly effective in estimating the marginal likelihood. The BS estimator uses both importance sampling draws as well as posterior draws in calculating the marginal likelihood, and Fruhwirth-Schnatter (2004) shows that the variance of the BS estimator is always bounded, thus making it a much more stable one.

Table (10) lists the marginal likelihood values calculated by Importance Sampling (IS), Reciprocal Importance Sampling (RI) and Bridge Sampling (BS) methods. From the table we can see the

BS estimator gives the most stable results. We calculate the marginal likelihoods for various FMM and DPM models. A few conclusions can be drawn from this table. If we specify a DP prior – again a separate one for each outcome – on only the error term (labeled as *error*), or the error term and the covariance term (labeled as *rho*), the resulting likelihoods show that they are not nearly flexible enough to capture the heterogeneity. This is supported by the posterior draws from the DPM model: Most parameters, especially those in the treatment group, are multi-modal and show strong skewness in one or both tails. The likelihoods show that FMM with six mixing components is the best model, and the likelihood improves substantially each time the number of components increases by one until it reaches six. Since the sample sizes of two groups are quite different (80% of the full sample has medical insurance), it is logical that the treatment group has more components. Therefore, in an ideal setting, we would like to test all the combinations of number of components for each outcome, and a combination of five components in the control group and ten components of the treatment group, for example, might give better model fit than the one with six components for both groups. Apparently this approach to model selection is tedious and time-consuming, and this is precisely where the auto-tuning mechanism of DPM can be quite helpful. Indeed, the likelihood for the DPM model dominates that of any FMM so that it does not seem necessary to exhaust every possible combination for FMM.

The model where a DP prior is put on all the parameters of the expenditure equation (labeled as *mixed*) is clearly the best model, and thus we report the model fit and results for that model only. Additionally, figure (14), the density plots of the predicated values, shows that this DPM model predicts well for both the subsample and the full sample.

5.3.5 Estimation Results

The MCMC algorithm produces well-mixed posterior draws and the chain appears to have converged rather quickly. Figure (15) and (16) plot the autocorrelation for the key parameters for both the control and treatment groups. They show little auto correlation, if at all.

Figure (17) plots the posterior draws for the key parameters of interest for the control group, and Figure (18) plots the posterior draws for the key parameters of interest for the treatment group. We can see from those plots that the parameters of the treatment group show much more multi-modality and skewness. All of the parameters in both groups appear to mix well.

The estimation results from the mixed model only are reported here. Table (11) shows the posterior means and standard deviations, as well as the p-values, for all the parameters of the selection equation. All the signs of the year dummies are negative and significant, indicating people are less likely to have medical insurance since the start of the economic recession in 2001 and 2002. With increasing age, people in our sample are more likely to have medical insurance. Also, Blacks, Hispanics, and males are less likely to have medical insurance. Living in a metropolitan statistical area, on the other hand, increases the chance of having medical insurance as expected. Relatively unhealthy people are less likely to have insurance as the signs on GOOD and FAIRPOOR are negative. The strongest indicators for having medical insurance are income, education and marital status: People with higher income and more education, as well as married ones, are significantly more likely to have insurance. And the chosen instrument, FIRMSIZE, is strongly correlated with the probability of having insurance: People working for larger firms are more likely to have medical insurance, which makes sense as larger companies usually provide employer-sponsored insurance as part of their benefits.

Table (12) shows the results for both groups. [2] The coefficients from both groups have the same signs; however, the magnitudes are quite different for most relevant parameters. The intercept is much larger for the treatment group, indicating the existence of sizable treatment effects. Magnitudes of the effects of health status are quite different too: Relatively unhealthy people – those with self-reported health status as VEGOOD, GOOD, FAIR or POOR, as well as those with physical limitations – spend much more on medical care if they have medical insurance. The control group is more homogeneous than the treatment group, as it has on average 8 components vs. 17 components in the treatment group.

While we do not attempt to identify different components under DPM, nonetheless we are able to detect some patterns of clustering from the posterior draws. The posterior distribution of the intercept term is clearly multi-modal, and if we plot the draws of the parameter on the intercept against those of the key parameters, we can see many discernible clusters.

Figure (19) is the scatter plot of the posterior draws of coefficients for the intercept and the variance for the both the control and treatment group. It clearly shows that there are more components in the treatment group with regard to variance and intercept. Figure (20) is the scatter plot of the intercept and the covariance, and it shows some interesting patterns. In the treatment group, the coefficient on the covariance is close to zero for different clusters, indicating the selection effects are minimal for high spenders and low spenders. The control group, however, shows a different pattern: The high spenders show some advantageous selection effects and the low spenders show some adverse selection effects. Finally, figure (21) is a scatter plot of the intercept and MARRY. In general we see married people tend to spend more on health care, however we see a tight cluster of

[2] Table (9) lists the estimates from different estimators. The estimates from *Roy-type Model with Normal Errors* and *Roy-type DPM Model* are calculated according to Heckman et al. (2001); All the rest are directly from the respective coefficient on the endogenous treatment variable.

people who spend more on health care regardless of martial status in the treatment group.

5.3.6 Calculation and Interpretation of Treatment Effects

As part of MCMC algorithm, the treatment effect (TE) and the treatment effect for the treated (TT) are calculated in this section for each individual. For each iteration after burn-in, they are calculated as follows:

$$TE_i^r = \exp\left(Z_i\delta + 0.5\left(\rho_2^2 + \sigma_2^2\right)\right) - \exp\left(Z_i\delta + 0.5\left(\rho_1^2 + \sigma_1^2\right)\right) \tag{5.11}$$

$$TT_i^r = \frac{\Phi\left(Z_i\delta + \rho_2\right)}{\Phi\left(Z_i\delta\right)} \exp\left(Z_i\delta_2 + 0.5\left(\rho_2^2 + \sigma_2^2\right)\right) - \frac{\Phi\left(Z_i\delta + \rho_1\right)}{\Phi\left(Z_i\delta\right)} \exp\left(Z_i\delta_1 + 0.5\left(\rho_1^2 + \sigma_1^2\right)\right) \tag{5.12}$$

where r stands for the rth draw from the posterior, and i denotes the ith individual. The average treatment effect (ATE) and average treatment effect on the treated (ATT) are then calculated for each draw:

$$ATE^r = \frac{1}{N}\sum_{i=1}^{N} TE_i^r$$

$$ATT^r = \frac{1}{N_1}\sum_{i=1}^{N_1} TT_i^r$$

where N is the full sample size, and N_1 is the sample size of the treatment group.

The posterior mean of ATE is estimated to be \$799, and that of ATT is \$764. The density plots of TE and TT are shown in Figure (22) and (23). Clearly they show substantial heterogeneity in treatment effects across individuals.

Figure (24) shows the scatter plot of TE and INCOME, and there is a apparent negative correlation between them. One possible reason for this correlation could be that people with higher income tend to be healthier, thus spending less on health care.

Figure (25) shows the scatter plot of TE and EDUCYR, and it tells a similar story: those with

higher education, especially those with more than twelve years of education, are likely to be healthier and spend less on health care.

Figure (26) shows the scatter plot of TE and AGEX, and we see older people tend to have bigger treatment effects – this is quite logical in that people will utilize more health care as they age, and if they have insurance, they will be more likely to see a doctor if needed. More importantly, the variance on the treatment effects increase with age, which again makes sense in that younger people tend to be uniformly healthier, and as people age, some of them start to have health issues while others maintain good health well into their 50s and 60s.

6 Conclusion and Future Research

6.1 Conclusion

This dissertation studies the endogenous treatment effects in the presence of heterogeneous responses. I show that a Bayesian Nonparametric approach can be quite fruitful in capturing the heterogeneity in treatment effects. When compared to an FMM approach, a DPM model captures the heterogeneity quite well by providing more flexible functional forms and thus providing better model fit. Rather than fixing the number of components in a mixture model, DPM allows the data and prior knowledge to determine the number of components in the data, providing an automatic mechanism for model selection.

I propose two DPM models in this dissertation. Built on top of a canonical linear selection model, the first DPM model specifies a DP prior on some or all the parameters of the structural equation, and marginal likelihoods are calculated to select the best DPM model. This model is used to study the incentive and selection effects of having prescription drug coverage on total drug expenditures among Medicare beneficiaries, and the posterior draws and predicted density both show this model works quite well for that application in producing sensible results and clustering patterns.

Specifically, I find that there are strong incentive effects for seniors having drug coverage – on average those who have drug coverage spend over one thousand dollars more than those who do not. Also I find there are advantageous selection effects, and the selection effect is found to be more heterogeneous than the incentive effect. Under DPM we do not label groups within the sample, but

nonetheless two discernible clusters can be seen in the scatter plots of posterior draws. If we accept the generally acknowledged findings in health economics that the population consists of two groups – the healthy and less healthy groups – with the latter being likely to consume more health care resources, we can show that the relatively unhealthy group is much more homogeneous than the other group. Graphically, the group with higher average expenditure is a much tighter cluster than the one with lower average expenditure. This in fact provides evidence for utilizing a DPM model where we can devote more components to the healthy groups, rather than lumping all of them into one group in the estimation.

While the two-equation DPM model works well for data sets where the treatment group and the control group are not too far apart from each other, a Roy-type model is warranted when the imbalance between those two sub samples are substantial. Thus, in the second DPM model, I utilize a three-equation Roy-type framework to model the observed heterogeneity that arises due to the treatment status, while the unobserved heterogeneity is handled by separate DPM models for the treated and untreated outcomes. I show that for the studied data the DPM model outperforms the FMM. In addition, the DPM models automatically select the number of components within each outcome to give the best model fit, bypassing the cumbersome task of trying out all the possible combinations of number of components for both the control and the treatment groups under FMM. Utilization of the low-level language C++ substantially improves the efficiency of the DP algorithm boosting the computational speed by a factor of somewhere between 60 and 100 relative to R making it well suited for implementing DP algorithms.

This Roy-type DPM model is applied to a data set consisting of $33,081$ independent individuals from the Medical Expenditure Panel Survey (MEPS), and the treatment effects of having private medical insurance on the outpatient expenditures are estimated. I find that there are strong incentive

effects – those with insurance, on average, spend $799 more on outpatient care than those without insurance. The posterior draws of most parameters show multimodality and substantial skewness, especially in the treatment group, where the DPM model is able to generate more components, or clusters, than that of the control group. The selection effect, however, is found to be quite small after the heterogeneity is adequately modeled. This is similar to the finding of Conley et al. (2008) who also report less "endogeneity" bias if more flexible function forms are specified on the error terms.

6.2 Future Research

One idea for future research is to specify a FMM with a DP prior on the error terms. While DPM gives better model fit, it does not give us results for each component, and sometimes we are interested in the estimates for those components. The idea behind this FMM-DP model is to let FMM identify the components in the data, while the DP prior is employed to handle the unknown heterogeneity in the error terms. The model is specified as follows:

$$D_i = W_i\alpha + u_i \tag{6.1}$$

$$M_i = V_i\gamma + \xi_i \tag{6.2}$$

$$Y_{ij} = X_i\beta_j + \rho_j d_i + \delta_j u_i + \varepsilon_{ij} \tag{6.3}$$

with

$$\varepsilon_{ij} \sim G(\alpha G_{0j})$$

The first two equations are the same as in MT2010, and a separate DP prior is specified on the error term of each component as seen in equation (6.3), instead of the normal prior as in MT2010.

I ran the model using the same data set as the one used in the DPM selection model. The coefficients of the selection equation (6.2) show a high level of autocorrelation, and the model seems quite unstable. There could be a couple of reasons for this phenomena. It is possible that the selection equation is indeed misspecified given this data set; or the MNP selection mechanism and DPM are working on the same source of heterogeneity, which makes the parameters on the selection equation unidentifiable, which leads to the instability of the model. Future research in this area should be quite interesting.

References

J.H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

O. Ashenfelter. Estimating the effect of training programs on earnings. *The Review of Economics and Statistics*, 60(1):47–57, 1978.

O. Ashenfelter and D. Card. Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics*, pages 648–660, 1985.

D. Blackwell and J.B. MacQueen. Ferguson distributions via Pólya urn schemes. *The annals of statistics*, 1(2):353–355, 1973.

M. Burda, M. Harding, and J. Hausman. A bayesian mixed logit-probit model for multinomial choice. *Journal of Econometrics*, 147(2):232–246, 2008.

D. Card. Sullivan,(1988),measuring the effect of subsidized training programs on movements in and out of employment. *Econometrica*, 56(3):497–530, 1988.

S. Chib. Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, 90(432), 1995.

S. Chib and B.H. Hamilton. Bayesian analysis of cross-section and clustered data treatment models. *Journal of Econometrics*, 97(1):25–50, 2000.

S. Chib and B.H. Hamilton. Semiparametric Bayes analysis of longitudinal data treatment models. *Journal of Econometrics*, 110(1):67–89, 2002.

S. Chib and I. Jeliazkov. Marginal likelihood from the metropolis-hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001.

T.G. Conley, C.B. Hansen, R.E. McCulloch, and P.E. Rossi. A semi-parametric Bayesian approach to the instrumental variable problem. *Journal of Econometrics*, 144(1):276–305, 2008.

P. Deb, M.K. Munkin, and P.K. Trivedi. Private Insurance, Selection, and Health Care Use. *Journal of Business and Economic Statistics*, 24(4):403–415, 2006.

M.D. Escobar and M. West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the american statistical association*, 90(430), 1995.

M.D. Escobar and M. West. Computing nonparametric hierarchical models. *Practical nonparametric and semiparametric Bayesian statistics*, 133:1–22, 1998.

H. Fang, M.P. Keane, and D. Silverman. Sources of advantageous selection: Evidence from the medigap insurance market. *The Journal of Political Economy*, 116(2):303–350, 2008.

T.S. Ferguson. Bayesian density estimation by mixtures of normal distributions. *Recent advances in statistics*, pages 287–303, 1983.

T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *The annals of statistics*, 1(2): 209–230, 1973.

S. Fruhwirth-Schnatter. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96(453):194–209, 2001.

S. Fruhwirth-Schnatter. Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *Econometrics Journal*, 7(1):143–167, 2004.

S. Fruhwirth-Schnatter. *Finite mixture and Markov switching models*. Springer Verlag, 2006. ISBN 0387329099.

A.E. Gelfand and A.F.M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

J. Geweke and M. Keane. Smoothly mixing regressions. *Journal of Econometrics*, 138(1):252–290, 2007.

J. Geweke, G. Gowrisankaran, and J. Robert. Town (2003).bayesian inference for hospital quality in a selection model.. *Econometrica*, 71(4):1215–1238, 2003.

P.J. Green and S. Richardson. Modelling heterogeneity with and without the Dirichlet process. *Scandinavian journal of statistics*, 28(2):355–375, 2001. ISSN 1467-9469.

J. Heckman, J.L. Tobias, and E. Vytlacil. Four parameters of interest in the evaluation of social programs. *Southern Economic Journal*, 68(2):211–223, 2001.

J.J. Heckman and E. Vytlacil. Structural Equations, Treatment Effects, and Econometric Policy Evaluation1. *Econometrica*, 73(3):669–738, 2005. ISSN 1468-0262.

J.J. Heckman, S. Urzua, and E. Vytlacil. Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, 88(3):389–432, 2006. ISSN 0034-6535.

P.W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81 (396):945–960, 1986.

G.W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

G.W. Imbens and J.D. Angrist. Identification and estimation of local average treatment effects. *Econometrica: Journal of the Econometric Society*, 62(2):467–475, 1994.

G.W. Imbens and J.M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86, 2009.

H. Ishwaran and L.F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

H. Ishwaran and L.F. James. Approximate Dirichlet Process Computing in Finite Normal Mixtures. *Journal of Computational and Graphical Statistics*, 11(3):508–532, 2002.

G. Koop and D.J. Poirier. Learning about the across-regime correlation in switching regression models. *Journal of Econometrics*, 78(1):217–227, 1997.

R.J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620, 1986.

A.Y. Lo. On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12(1):351–357, 1984.

S.N. MacEachern and P. Muller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.

X.L. Meng and W.H. Wong. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 6:831–860, 1996.

M.K. Munkin and P.K. Trivedi. Disentangling incentives effects of insurance coverage from adverse selection in the case of drug expenditure: a finite mixture approach. *Health Economics*, 19(9): 1093–1108, 2010.

M.P. Murray. Avoiding invalid instruments and coping with weak instruments. *The Journal of Economic Perspectives*, 20(4):111–132, 2006. ISSN 0895-3309.

R. Neal. Erroneous results in 'marginal likelihood from the gibbs output'. 1999. URL `http://www.cs.toronto.edu/ radford/publications.html`.

R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

O. Papaspiliopoulos and G.O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 2008.

A.D. Roy. Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3(2):135–146, 1951.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

S.G. Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics-Simulation and Computation*, 36(1):45–54, 2007.

Appendix A: Tables and Figures for the DPM Selection Model

Table 1: Varible Definitions and Summary Statistics: DPM Selection Model

	Full sample	No drug coverage	Drug coverage
AAMTTOT	1687.85(1647.29)	1587.93(1563.75)	1924.75(1809.47)
LNAAMTTOT	6.95(1.15)	6.88(1.17)	7.11(1.11)
COVRX	0.30(0.46)	0.00(0.00)	1.00(0.00)
AGE	78.17(7.39)	78.36(7.27)	77.71(7.65)
MALE	0.40(0.49)	0.42(0.49)	0.36(0.48)
WHITE	0.95(0.22)	0.96(0.20)	0.93(0.25)
HEARTCOND	0.80(0.40)	0.79(0.41)	0.82(0.38)
OTHCOND	0.62(0.49)	0.61(0.49)	0.64(0.48)
MARRY	0.54(0.50)	0.54(0.50)	0.52(0.50)
WIDOW	0.39(0.49)	0.39(0.49)	0.40(0.49)
LVALONE	0.37(0.48)	0.36(0.48)	0.38(0.48)
CHNLNM	2.90(1.86)	2.88(1.86)	2.96(1.89)
MSA	0.63(0.48)	0.59(0.49)	0.71(0.46)
NOREAST	0.12(0.33)	0.11(0.31)	0.16(0.36)
WEST	0.12(0.33)	0.11(0.31)	0.16(0.37)
SOUTH	0.44(0.50)	0.46(0.50)	0.38(0.49)
VEGOOD	0.30(0.46)	0.30(0.46)	0.32(0.47)
GOOD	0.36(0.48)	0.37(0.48)	0.34(0.47)
FAIR	0.15(0.36)	0.14(0.35)	0.16(0.37)
POOR	0.04(0.20)	0.04(0.19)	0.05(0.22)
SMOKNOW	0.08(0.26)	0.08(0.27)	0.06(0.25)
SMOKEVER	0.56(0.50)	0.56(0.50)	0.54(0.50)
DEGRCV	4.69(1.95)	4.65(1.94)	4.79(1.96)
JOBSTAT	0.11(0.31)	0.11(0.31)	0.10(0.30)
LNINCOME	10.00(0.77)	9.98(0.76)	10.06(0.80)
YEAR2004	0.28(0.45)	0.29(0.46)	0.24(0.43)
YEAR2005	0.27(0.44)	0.21(0.41)	0.40(0.49)
PENET	7.38(10.82)	6.41(10.19)	9.68(11.89)
MOAMT	148.34(89.23)	136.85(64.95)	175.58(125.69)

Table 2: Statistics for Sample Differences: DPM Selection Model

	Imbens and Rubin 2007	T statistics
AGE	-0.06	-1.88
MALE	-0.09	-2.69
WHITE	-0.07	-2.23
HEARTCOND	0.06	1.77
OTHCOND	0.03	1.07
MARRY	-0.03	-1.07
WIDOW	0.01	0.37
LVALONE	0.02	0.54
CHNLNM	0.03	0.97
MSA	0.17	5.47
NOREAST	0.11	3.30
WEST	0.12	3.66
SOUTH	-0.11	-3.39
VEGOOD	0.04	1.14
GOOD	-0.05	-1.50
FAIR	0.03	0.95
POOR	0.05	1.41
SMOKNOW	-0.04	-1.42
SMOKEVER	-0.02	-0.75
DEGRCV	0.05	1.58
JOBSTAT	-0.01	-0.46
LNINCOME	0.07	2.16
YEAR2004	-0.09	-2.77
YEAR2005	0.29	8.73
PENET	0.21	6.29
MOAMT	0.27	7.65

Table 3: Comparison of Estimates from Different Estimators: MCBS Dataset

	Estimate
OLS	0.1163
OLS with Propensity Score	0.0020
Weighted GLM	0.0787
IV	1.1077
Heckman Two-Step	1.0263
Roy-type Model with Normal Errors	1.3170
DPM Selection Model	0.4568

Table 4: Marginal Likelihood Results: DPM Selection Model

	IS	RI	BS
Error	-5525.39	-6199.62	-5935.81
Rho	-5509.76	-6154.67	-5868.21
Both	-5712.16	-6213.09	-5965.51
Mixed	-5519.06	-6053.08	-5785.89

Table 5: Selection Equation Results: DPM Selection Model

	Estimate	Std. Error	Pr(>|t|)
CONST(First Stage)	-0.280	0.562	0.618
AGE(First Stage)	-0.014	0.004	0.001
MALE(First Stage)	-0.157	0.069	0.023
WHITE(First Stage)	-0.367	0.131	0.005
HEARTCOND(First Stage)	0.081	0.074	0.277
OTHCOND(First Stage)	0.059	0.061	0.332
MARRY(First Stage)	-0.058	0.126	0.648
WIDOW(First Stage)	-0.029	0.119	0.805
LVALONE(First Stage)	-0.005	0.087	0.953
CHNLNM(First Stage)	0.010	0.016	0.519
MSA(First Stage)	0.232	0.063	0.000
NOREAST(First Stage)	0.184	0.097	0.058
WEST(First Stage)	0.274	0.094	0.003
SOUTH(First Stage)	-0.131	0.069	0.056
VEGOOD(First Stage)	0.096	0.092	0.298
GOOD(First Stage)	0.028	0.094	0.764
FAIR(First Stage)	0.150	0.110	0.172
POOR(First Stage)	0.261	0.159	0.100
SMOKNOW(First Stage)	-0.115	0.116	0.322
SMOKEVER(First Stage)	0.011	0.065	0.861
DEGRCV(First Stage)	-0.017	0.016	0.312
JOBSTAT(First Stage)	-0.042	0.098	0.672
LNINCOME(First Stage)	0.042	0.043	0.332
YEAR2004(First Stage)	0.071	0.069	0.309
YEAR2005(First Stage)	0.530	0.069	0.000
MOAMT(First Stage)	0.003	0.000	0.000

Table 6: Expenditure Equation Results: DPM Selection Model

	Estimate	Std. Error	Pr(>|t|)
Number of Components	4.994	1.717	0.004
CONST	4.809	1.766	0.007
AGE	-0.002	0.007	0.754
MALE	-0.059	0.126	0.638
WHITE	0.194	0.210	0.355
HEARTCOND	0.539	0.406	0.184
OTHCOND	0.324	0.127	0.011
MARRY	-0.061	0.204	0.764
WIDOW	-0.026	0.178	0.886
LVALONE	0.029	0.140	0.837
CHNLNM	0.008	0.024	0.742
MSA	-0.035	0.123	0.778
NOREAST	0.029	0.162	0.857
WEST	0.008	0.138	0.956
SOUTH	0.165	0.110	0.136
VEGOOD	0.135	0.129	0.297
GOOD	0.389	0.159	0.014
FAIR	0.653	0.309	0.035
POOR	0.594	0.281	0.035
SMOKNOW	-0.179	0.213	0.403
SMOKEVER	0.047	0.128	0.710
DEGRCV	0.006	0.025	0.826
JOBSTAT	-0.243	0.151	0.109
LNINCOME	0.092	0.069	0.180
YEAR2004	0.093	0.103	0.366
YEAR2005	0.276	0.126	0.029
COVRX	0.457	0.184	0.013
Covariance	-0.236	0.150	0.116
Sigma	0.701	0.206	0.001

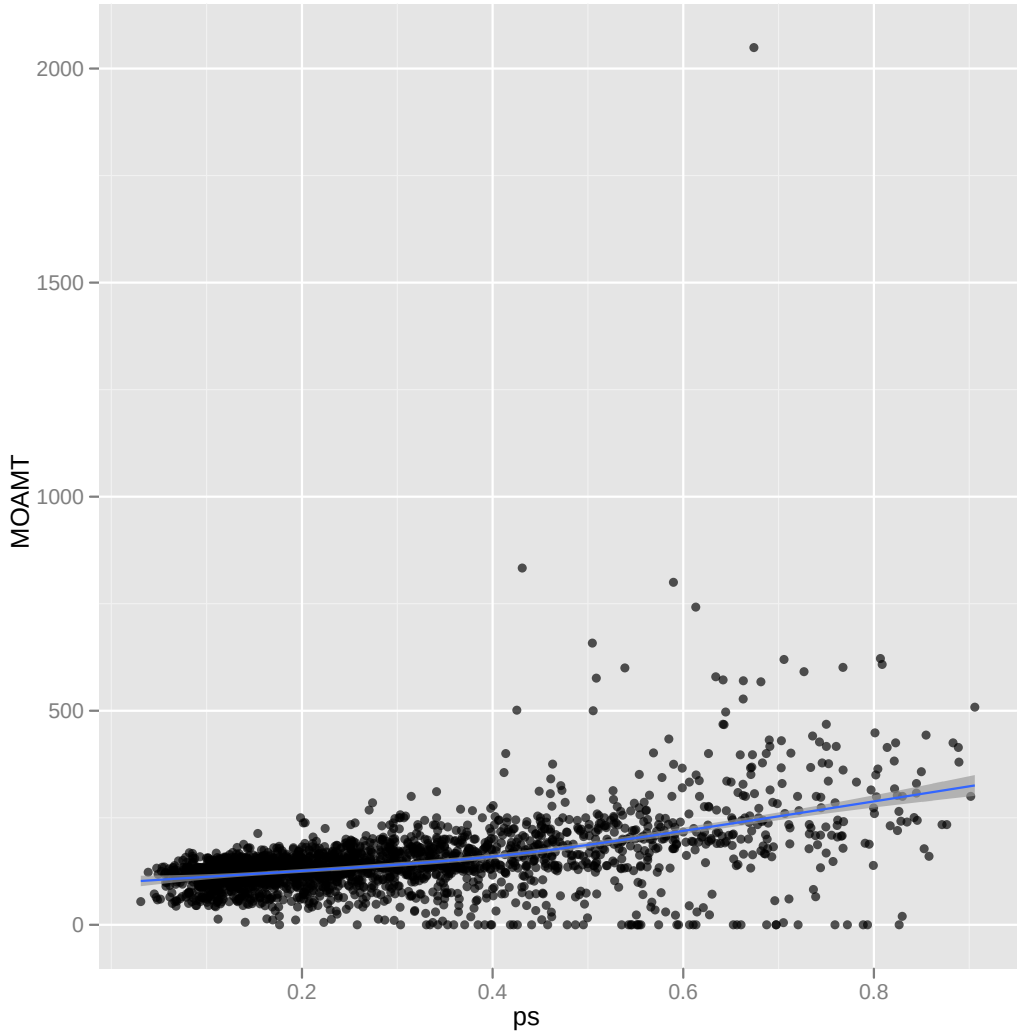

Figure 1: Scatter Plot of Propensisty Score vs IV for DPM Selection Model

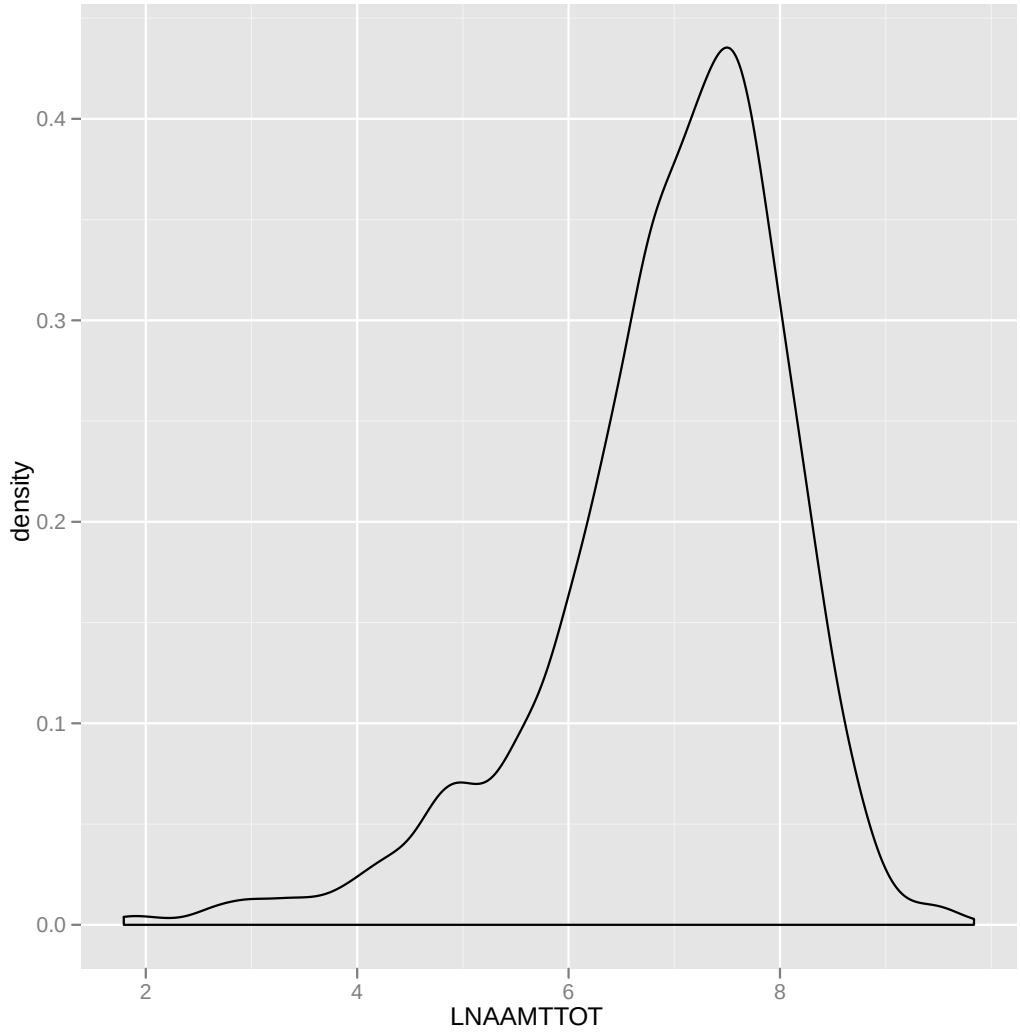

Figure 2: Kernel Density Plot of Data for DPM Selection Model

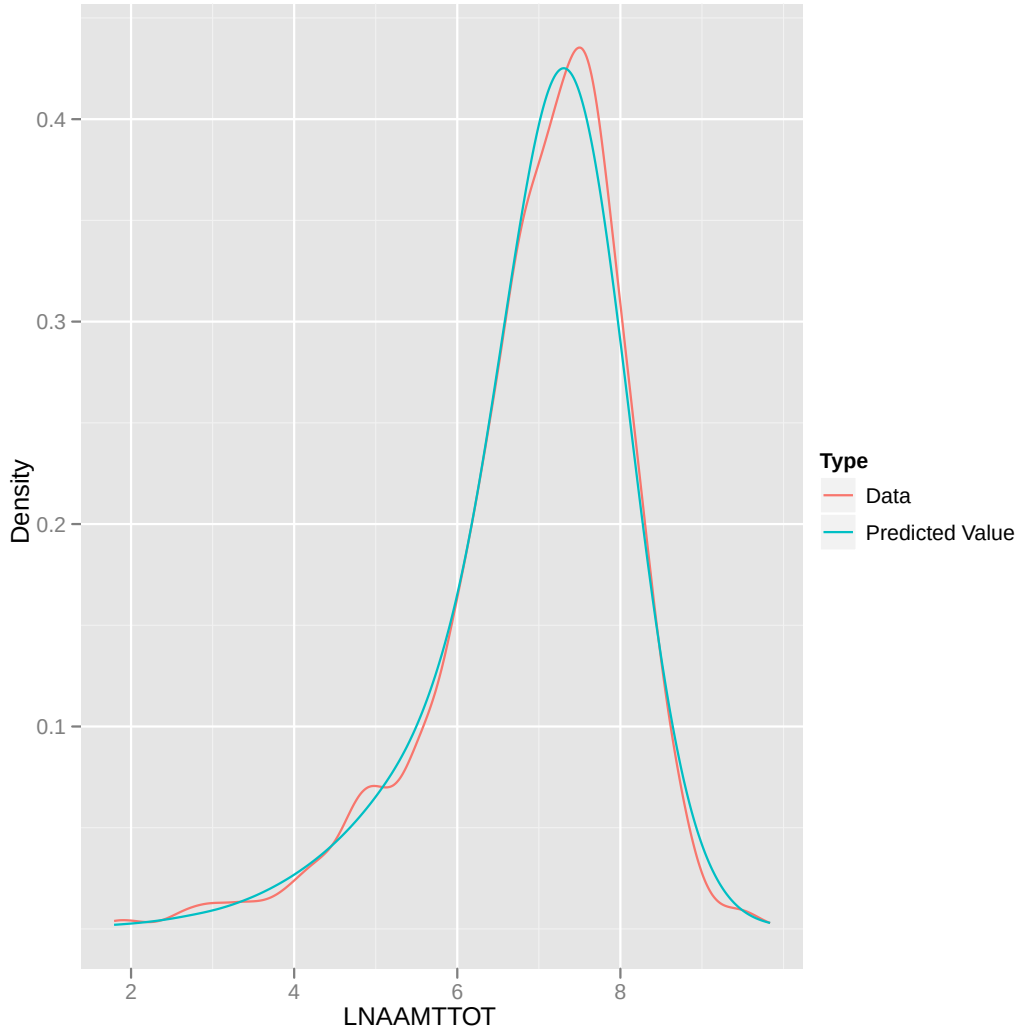

Figure 3: Density Plots of Data and Predicted Values for DPM Selection Model

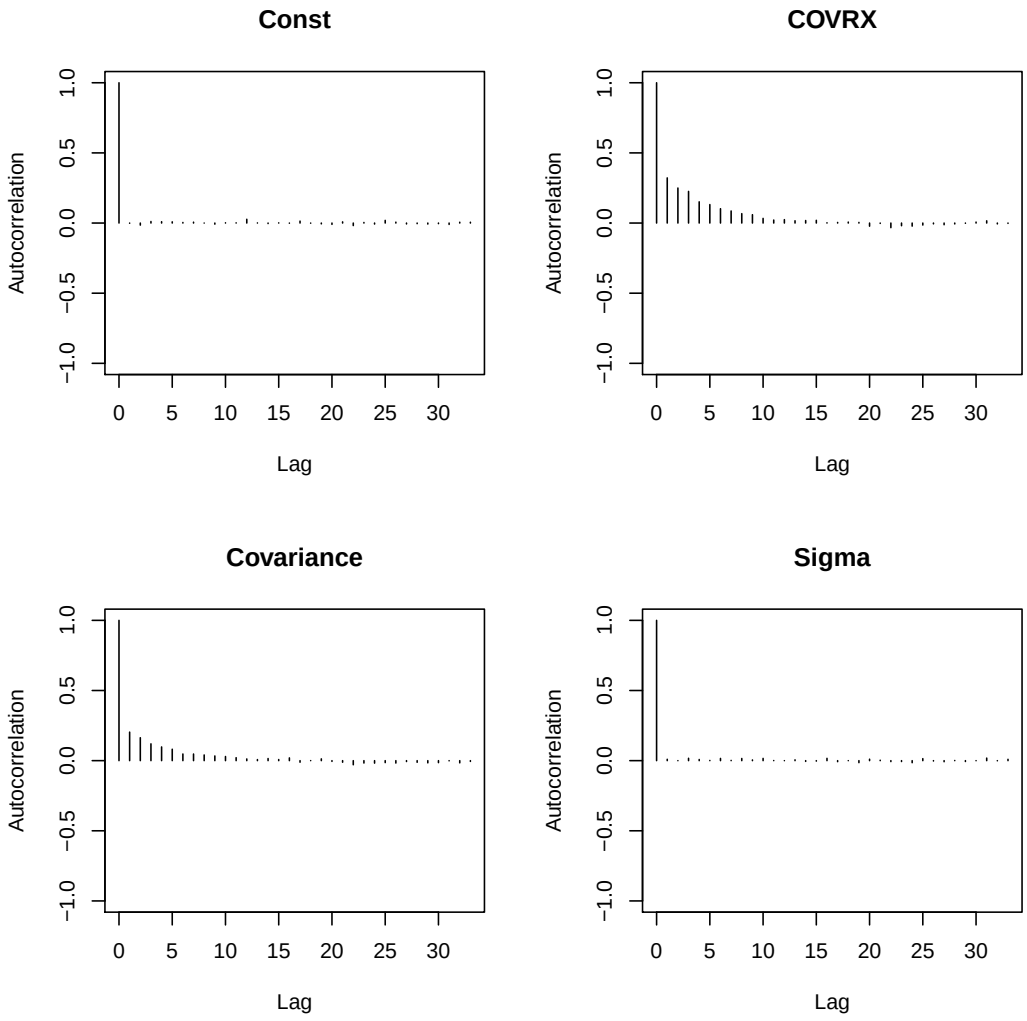

Figure 4: Auto Correlation Plots of Key Parameters for DPM Selection Model

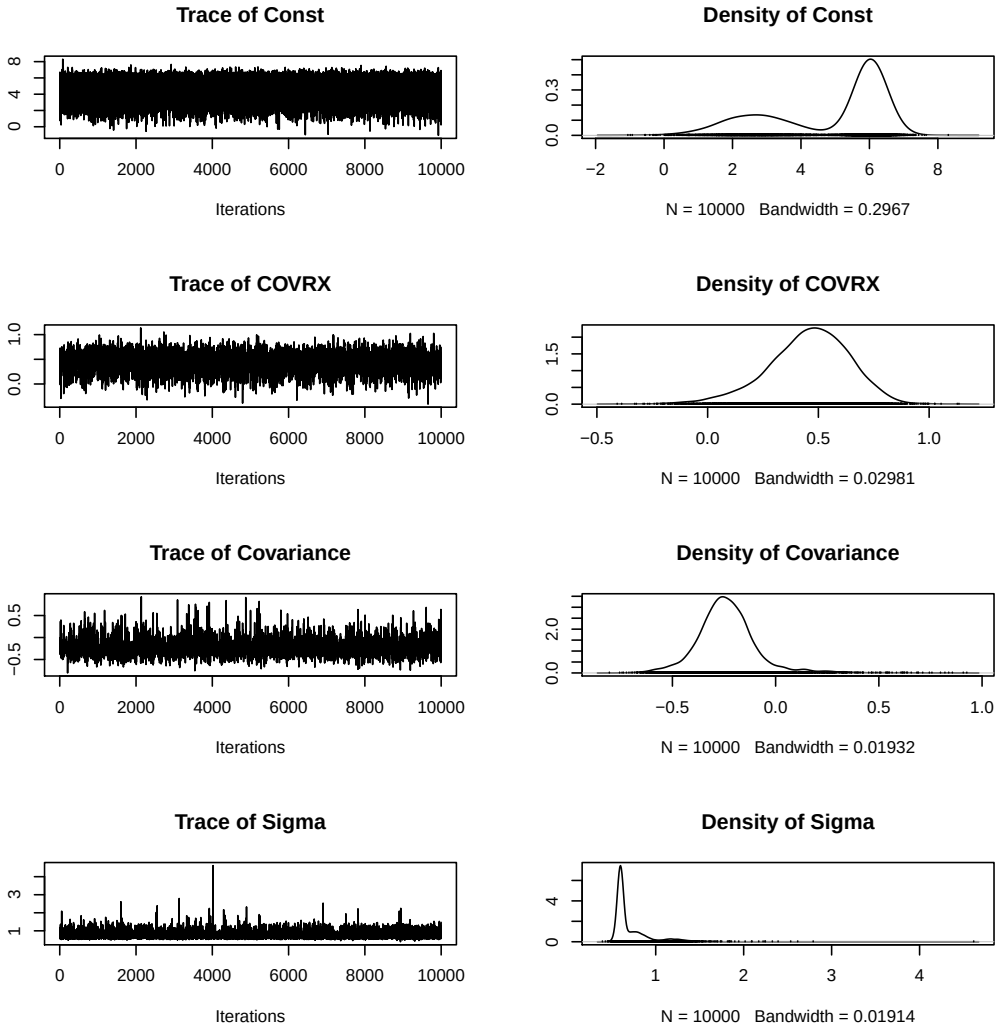

Figure 5: Trace and Density Plots of Key Parameters for DPM Selection Model

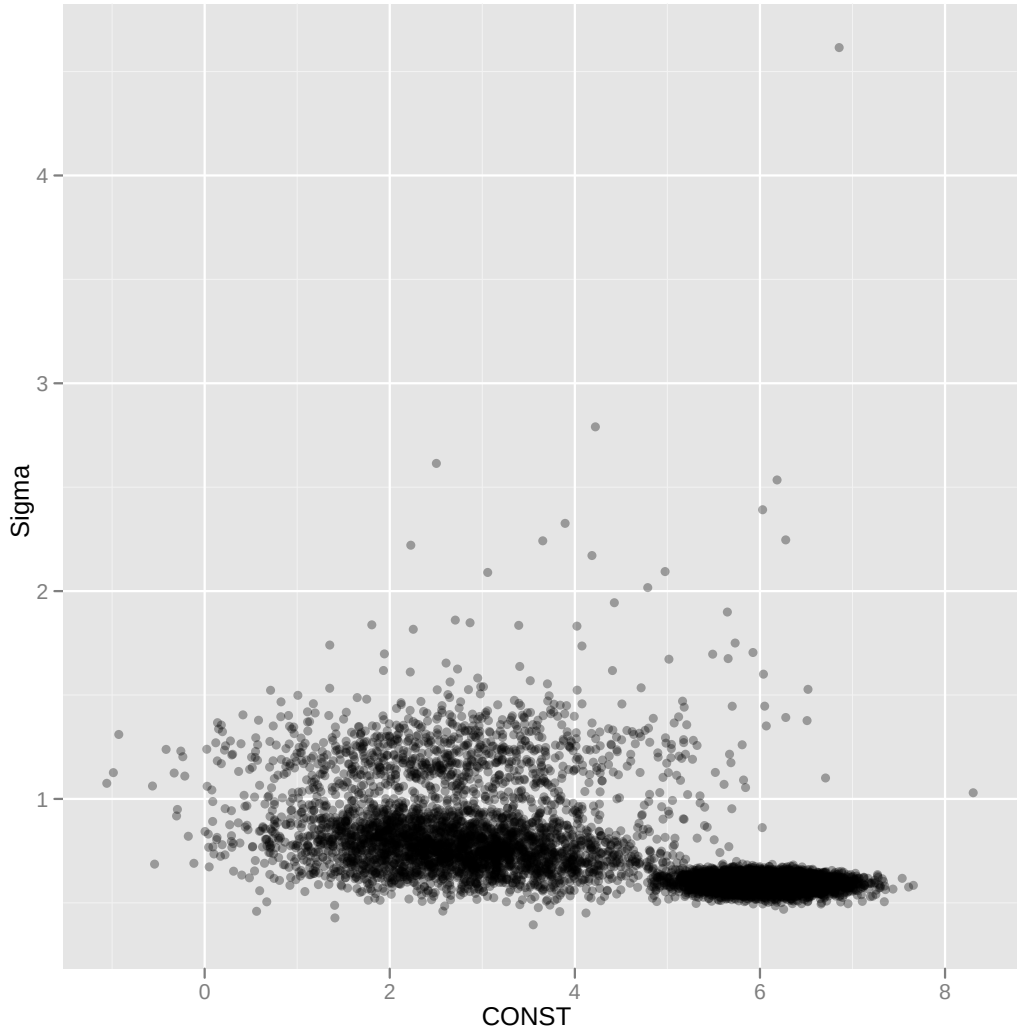

Figure 6: Posterior Scatter Plot of Intercept vs Variance for DPM Selection Model

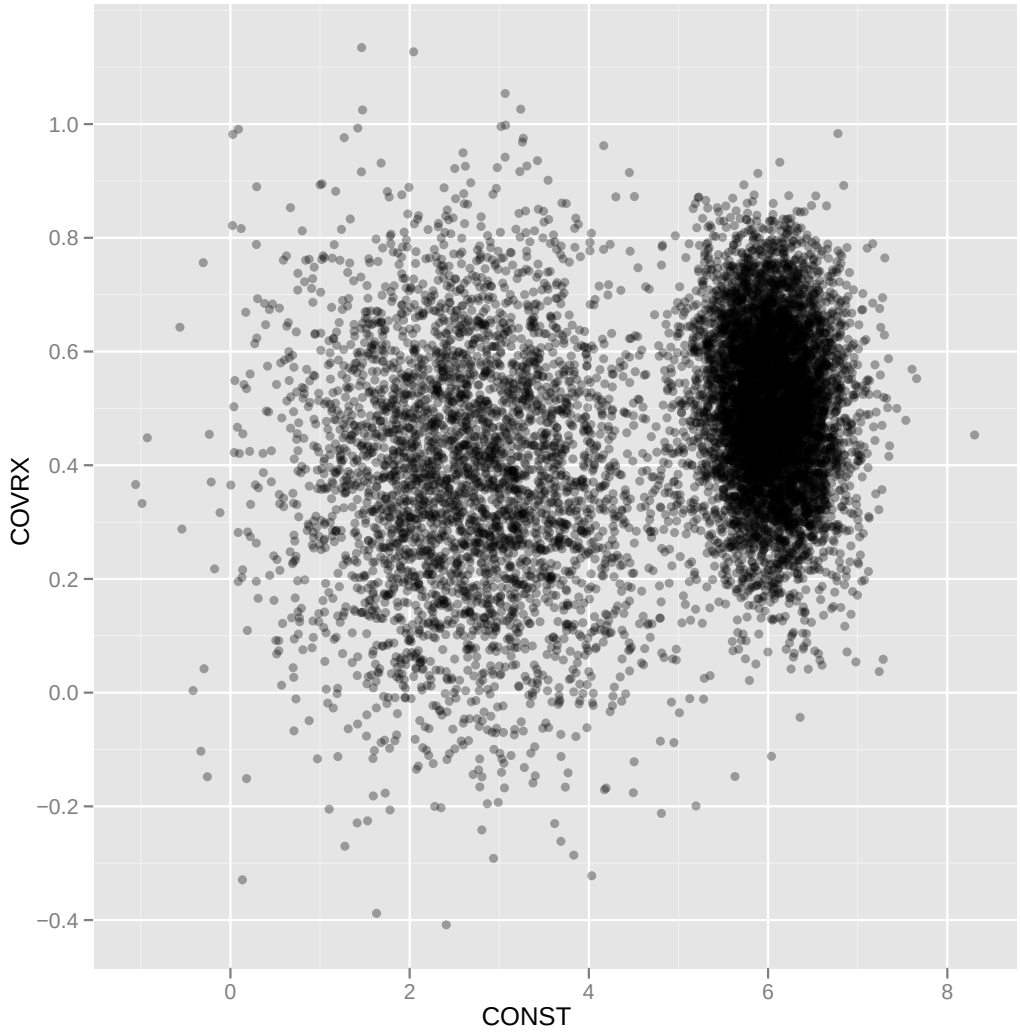

Figure 7: Posterior Scatter Plot of Intercept vs COVRX for DPM Selection Model

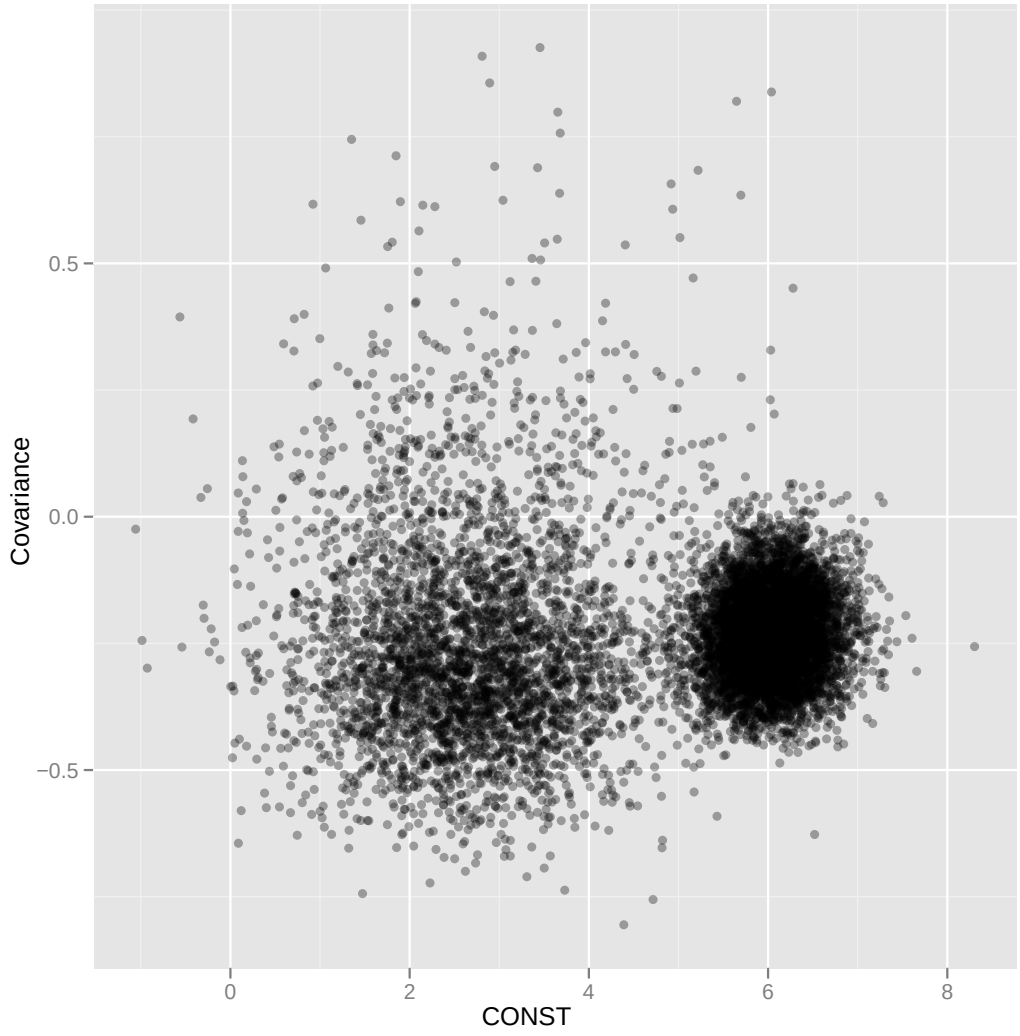

Figure 8: Posterior Scatter Plot of Intercept vs Selection Bias for DPM Selection Model

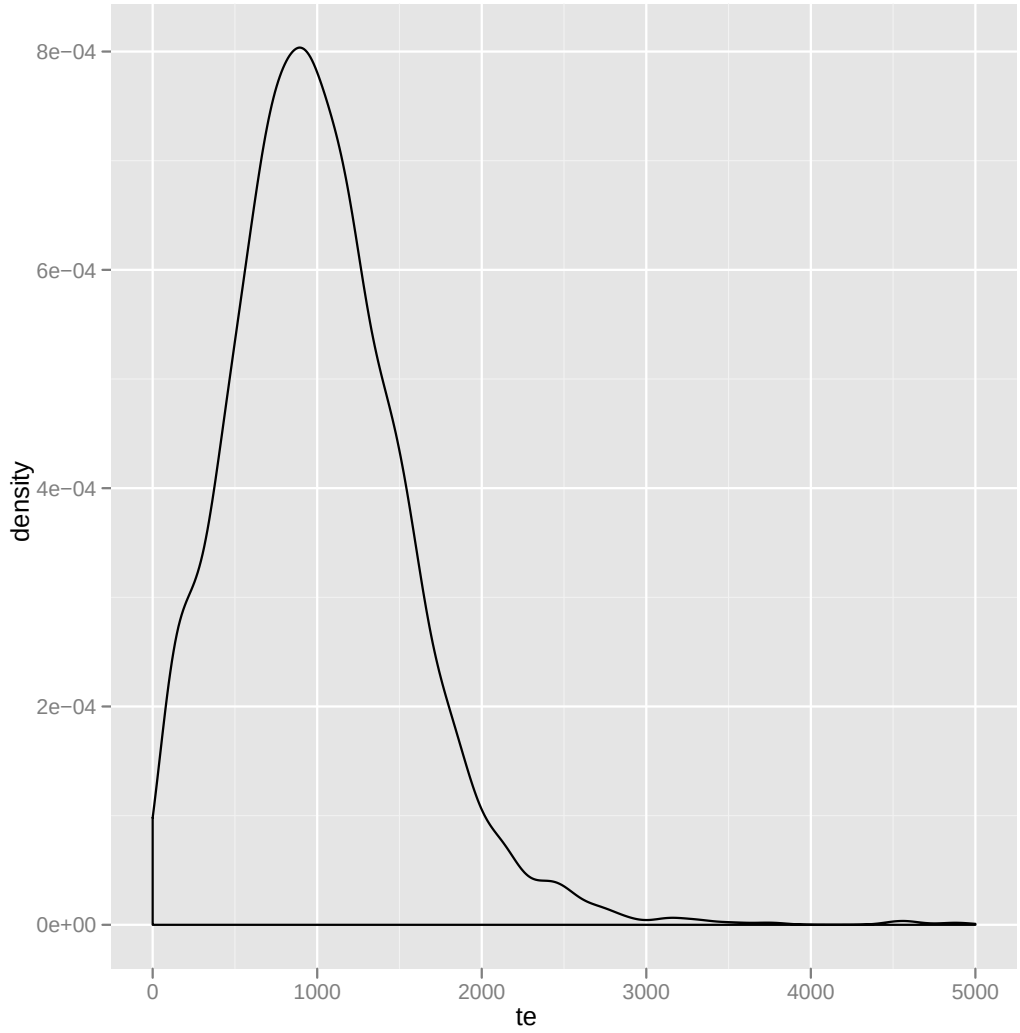

Figure 9: Density Plot of Treatment Effects for DPM Selection Model

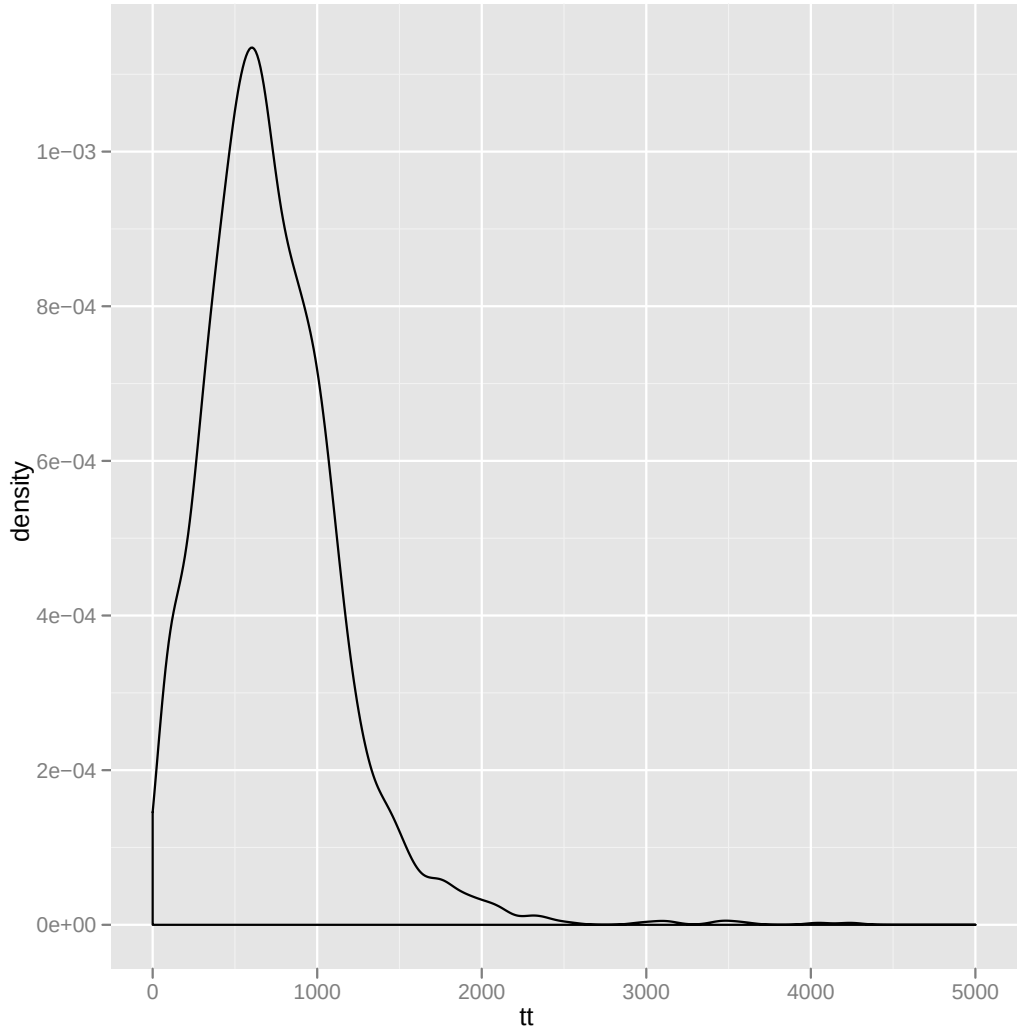

Figure 10: Density Plot of Treatment Effects for the Treated for DPM Selection Model

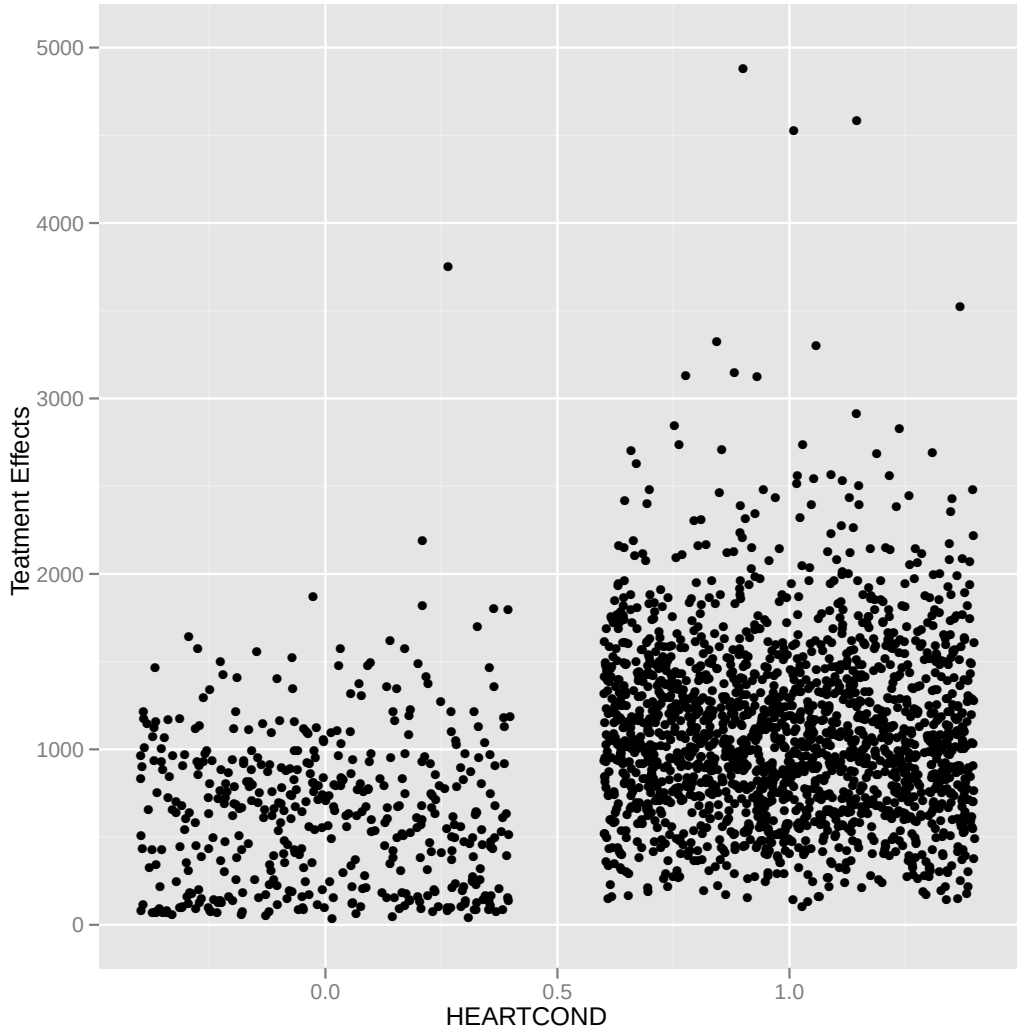

Figure 11: Posterior Jitter Plot of TE vs HEARTCOND for DPM Selection Model

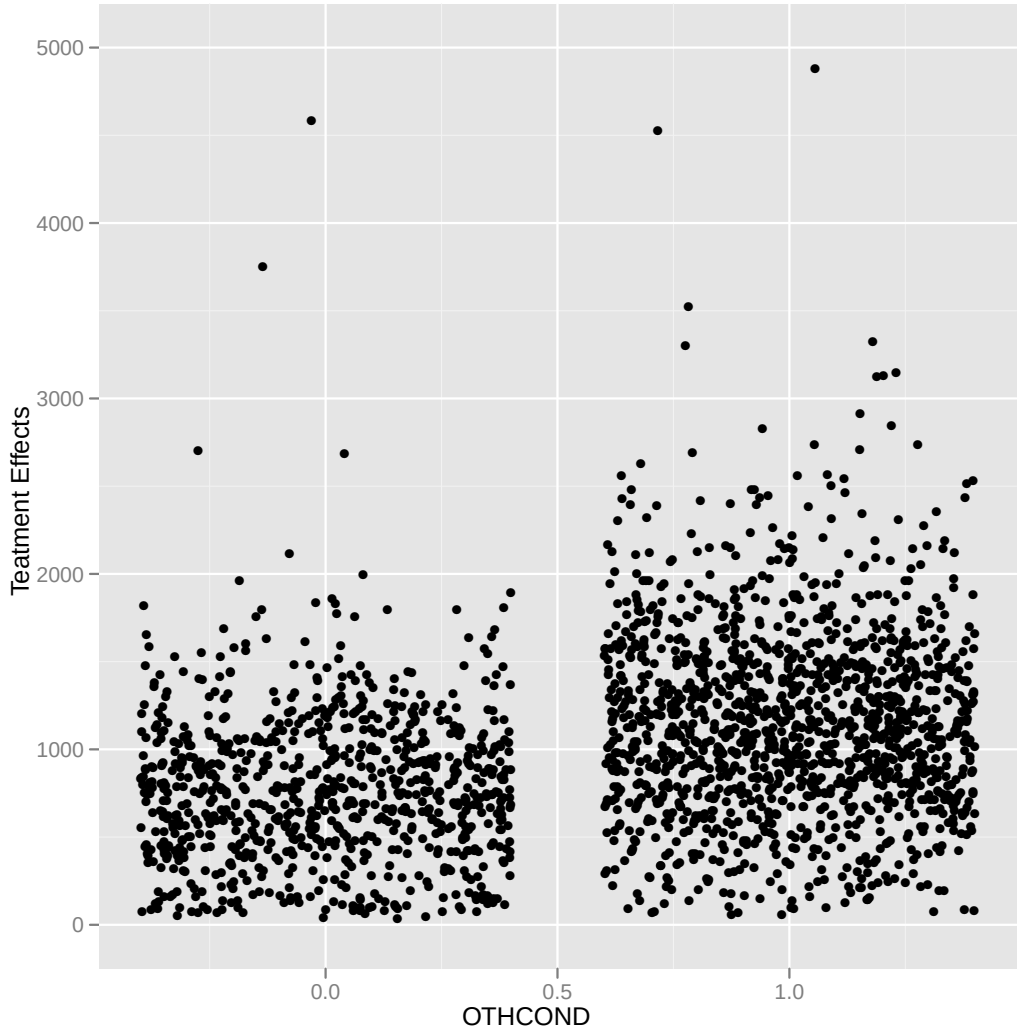

Figure 12: Posterior Jitter Plot of TE vs OTHER COND for DPM Selection Model

Appendix B: Tables and Figures for the Roy-type DPM Model

Table 7: Varible Definitions and Summary Statistics: Roy-type DPM Model

	Full sample	No Insurance	With Insurance
Sample Size	33081	4219	28862
OUTEXP	2331.15(4252.68)	1229.96(2773.06)	2492.12(4404.78)
LOUTEXP	6.85(1.46)	6.04(1.52)	6.97(1.41)
PRVEV	0.87(0.33)	0.00(0.00)	1.00(0.00)
YEAR03	0.12(0.33)	0.12(0.32)	0.12(0.33)
YEAR04	0.12(0.32)	0.13(0.33)	0.12(0.32)
YEAR05	0.11(0.32)	0.13(0.33)	0.11(0.32)
YEAR06	0.13(0.33)	0.14(0.34)	0.12(0.33)
YEAR07	0.10(0.30)	0.10(0.29)	0.10(0.30)
YEAR08	0.14(0.35)	0.16(0.37)	0.14(0.34)
VEGOOD	0.34(0.47)	0.27(0.45)	0.35(0.48)
GOOD	0.28(0.45)	0.33(0.47)	0.27(0.44)
FAIRPOOR	0.11(0.31)	0.19(0.39)	0.09(0.29)
PHYSLIM	0.07(0.25)	0.08(0.27)	0.06(0.24)
FAMSZEYR	3.02(1.49)	3.26(1.73)	2.99(1.44)
AGEX	4.29(1.04)	4.05(1.06)	4.33(1.04)
EDUCYR	13.44(2.79)	11.46(3.25)	13.73(2.59)
FEMALE	0.55(0.50)	0.54(0.50)	0.55(0.50)
BLACK	0.14(0.35)	0.17(0.37)	0.14(0.35)
HISP	0.17(0.38)	0.39(0.49)	0.14(0.35)
MARRY	0.67(0.47)	0.52(0.50)	0.69(0.46)
NOREAST	0.16(0.37)	0.10(0.30)	0.17(0.37)
MIDWEST	0.23(0.42)	0.15(0.35)	0.24(0.43)
SOUTH	0.37(0.48)	0.48(0.50)	0.36(0.48)
MSA	0.83(0.38)	0.79(0.41)	0.83(0.37)
INCOME	71.57(50.02)	39.89(33.82)	76.20(50.33)
FIRMSIZE	16.88(18.94)	8.30(14.15)	18.13(19.22)

Table 8: Statistics for Sample Differences: Roy-type DPM Model

	Imbens and Rubin 2007	T statistics
YEAR03	0.01	0.93
YEAR04	-0.02	-1.39
YEAR05	-0.03	-2.28
YEAR06	-0.03	-2.21
YEAR07	0.01	0.54
YEAR08	-0.05	-4.19
VEGOOD	0.12	10.61
GOOD	-0.09	-7.86
FAIRPOOR	-0.20	-15.75
PHYSLIM	-0.05	-3.84
FAMSZEYR	-0.12	-9.89
AGEX	0.19	15.79
EDUCYR	0.54	43.24
FEMALE	0.00	0.28
BLACK	-0.06	-4.90
HISP	-0.42	-32.06
MARRY	0.25	20.99
NOREAST	0.14	13.23
MIDWEST	0.17	15.76
SOUTH	-0.18	-14.80
MSA	0.07	6.11
INCOME	0.60	60.61
FIRMSIZE	0.41	40.07

Table 9: Comparison of Estimates from Different Estimators: MEPS Dataset

	Estimate
OLS	0.7005
OLS with Propensity Score	0.6296
Weighted GLM	0.6120
IV	1.5641
Heckman Two-Step	1.4248
Roy-type Model with Normal Errors	1.6796
Roy-type DPM Model	0.5975

Table 10: Marginal Likelihood Results: Roy-type DPM Model

	IS	RI	BS
error	-98331.91	-99861.89	-98685.61
rho	-96197.51	-99189.61	-97026.69
mixed	-82225.64	-88136.32	-82327.33
2 components	-98333.57	-99440.19	-98781.91
3 components	-93602.16	-96285.69	-95497.31
4 components	-88948.96	-92202.41	-89781.95
5 components	-85574.08	-92579.34	-87769.80
6 components	-84189.61	-89184.13	-84981.33
7 components	-84447.17	-90142.50	-85352.30

Table 11: Selection Equation Results: Roy-type DPM Model

	Estimate	Std. Error	Pr(>|t|)
CONST(First Stage)	-0.514	0.080	0.000
YEAR03(First Stage)	-0.086	0.035	0.021
YEAR04(First Stage)	-0.127	0.033	0.001
YEAR05(First Stage)	-0.096	0.055	0.094
YEAR06(First Stage)	-0.135	0.051	0.013
YEAR07(First Stage)	-0.153	0.038	0.000
YEAR08(First Stage)	-0.202	0.033	0.000
VEGOOD(First Stage)	0.031	0.026	0.246
GOOD(First Stage)	-0.064	0.027	0.027
FAIRPOOR(First Stage)	-0.203	0.035	0.000
PHYSLIM(First Stage)	-0.017	0.039	0.661
FAMSZEYR(First Stage)	-0.080	0.007	0.000
AGEX(First Stage)	0.067	0.010	0.000
EDUCYR(First Stage)	0.070	0.004	0.000
FEMALE(First Stage)	0.068	0.020	0.002
BLACK(First Stage)	-0.074	0.028	0.014
HISP(First Stage)	-0.362	0.026	0.000
MARRY(First Stage)	0.340	0.023	0.000
NOREAST(First Stage)	0.151	0.034	0.000
MIDWEST(First Stage)	0.152	0.031	0.000
SOUTH(First Stage)	-0.087	0.027	0.003
MSA(First Stage)	0.068	0.025	0.012
INCOME(First Stage)	0.008	0.000	0.000
FIRMSIZE(First Stage)	0.014	0.001	0.000

Table 12: Outcome Equation Results: Roy-type DPM Model

	Control	Treated
Sigma	1.119(0.2137)	0.8009(0.3054)
Covariance	-0.005(0.0872)	-0.0110(0.0445)
Intercept	4.221(0.6964)	4.8143(1.2675)
YEAR03	-0.099(0.0858)	0.1973(0.0802)
YEAR04	0.002(0.1188)	0.2716(0.0687)
YEAR05	0.085(0.0764)	0.1096(0.0794)
YEAR06	0.090(0.0643)	0.2571(0.0877)
YEAR07	0.123(0.0922)	0.3473(0.0945)
YEAR08	0.175(0.1265)	0.3001(0.0861)
VEGOOD	0.152(0.0501)	0.2615(0.0750)
GOOD	0.342(0.0560)	0.4878(0.1018)
FAIRPOOR	0.645(0.1574)	0.9410(0.1999)
PHYSLIM	0.449(0.0865)	0.6068(0.1188)
FAMSZEYR	-0.041(0.0200)	-0.0913(0.0370)
AGEX	0.231(0.0283)	0.2423(0.0720)
EDUCYR	0.045(0.0077)	0.0365(0.0139)
FEMALE	0.233(0.0630)	0.3689(0.1050)
BLACK	-0.223(0.0584)	-0.3037(0.1072)
HISP	-0.278(0.0828)	-0.2900(0.0978)
MARRY	0.125(0.0860)	0.0903(0.0841)
NOREAST	-0.049(0.0801)	0.0642(0.0733)
MIDWEST	-0.081(0.0608)	0.1813(0.0652)
SOUTH	-0.044(0.0540)	0.0483(0.0466)
MSA	-0.022(0.0492)	0.0165(0.0502)
INCOME	0.004(0.0006)	0.0009(0.0007)
DP Precision	0.721(0.3669)	1.6481(0.7260)
Number of Components	7.971(2.6378)	17.4576(5.9086)

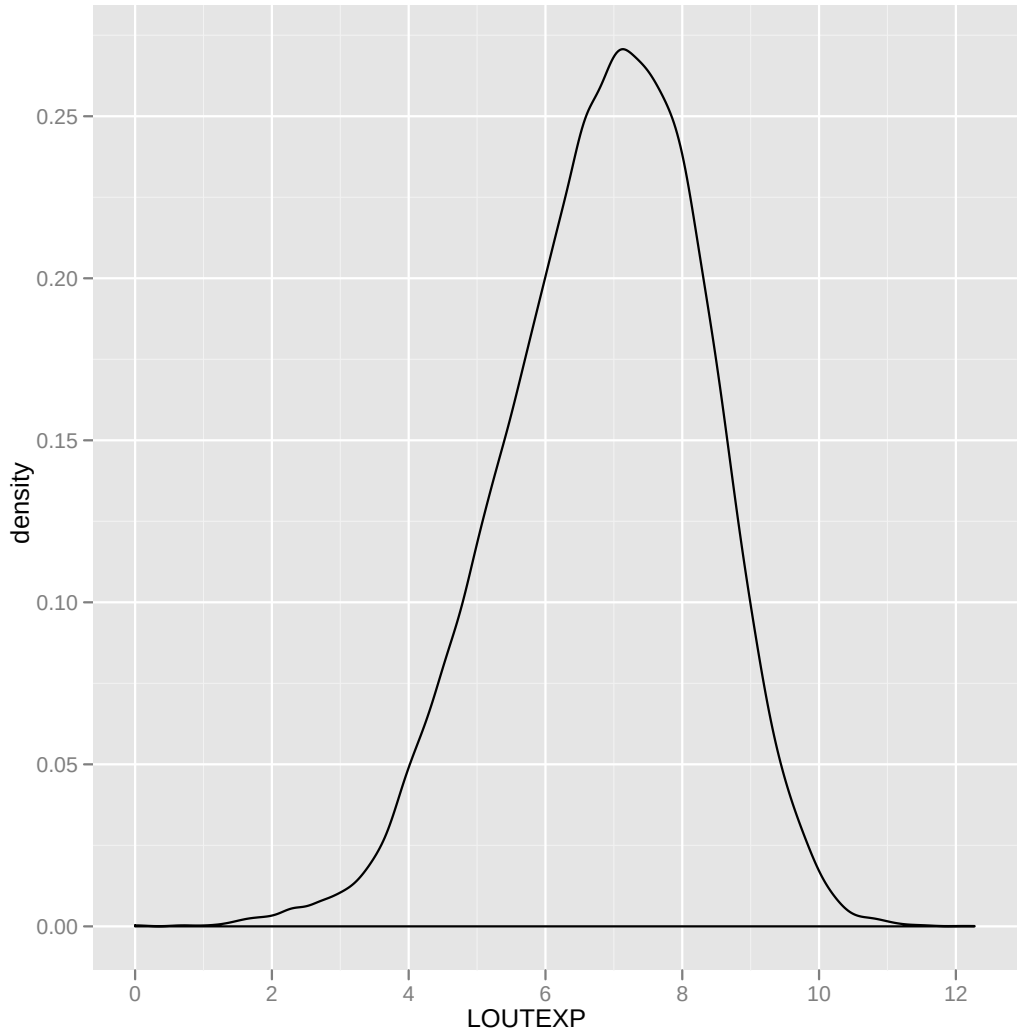

Figure 13: Kernel Density Plot of Data for Roy-type DPM Model

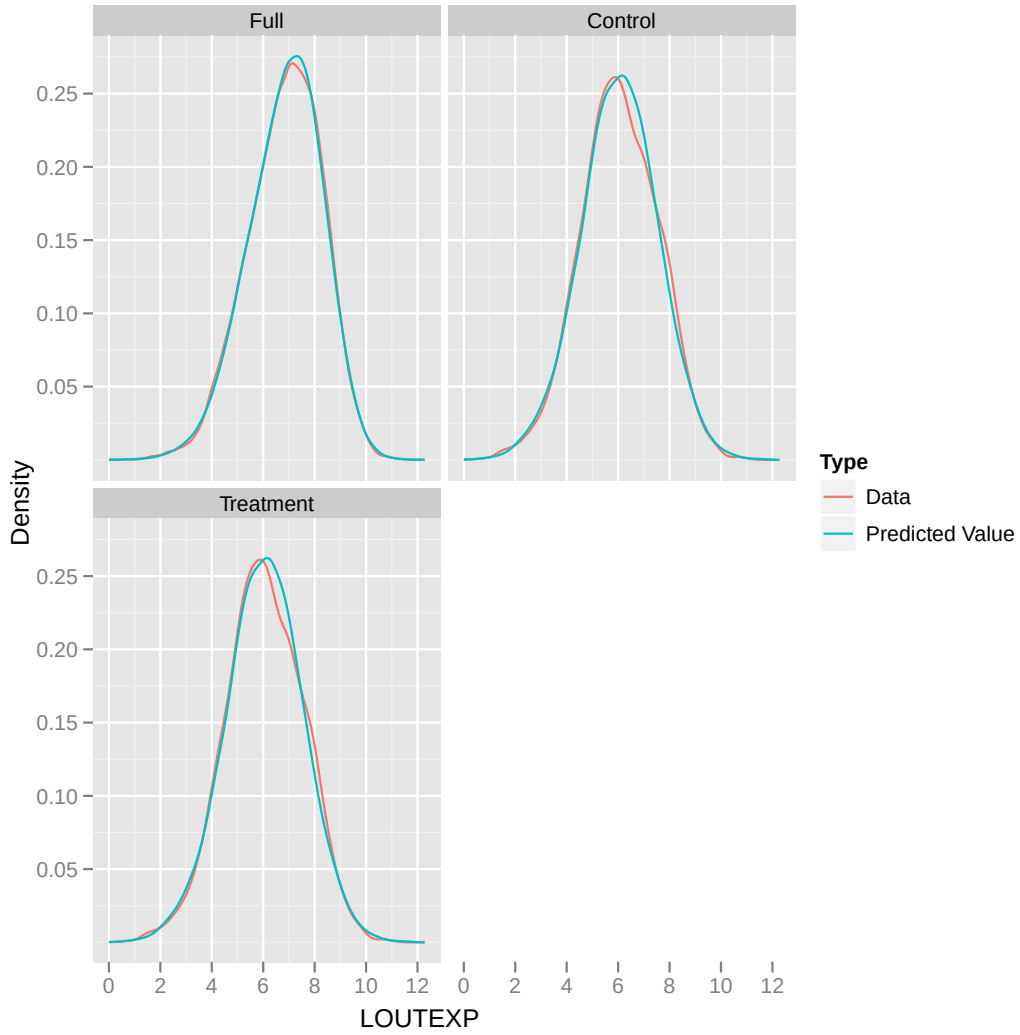

Figure 14: Density Plots of Data and Predicted Values for Roy-type DPM Model

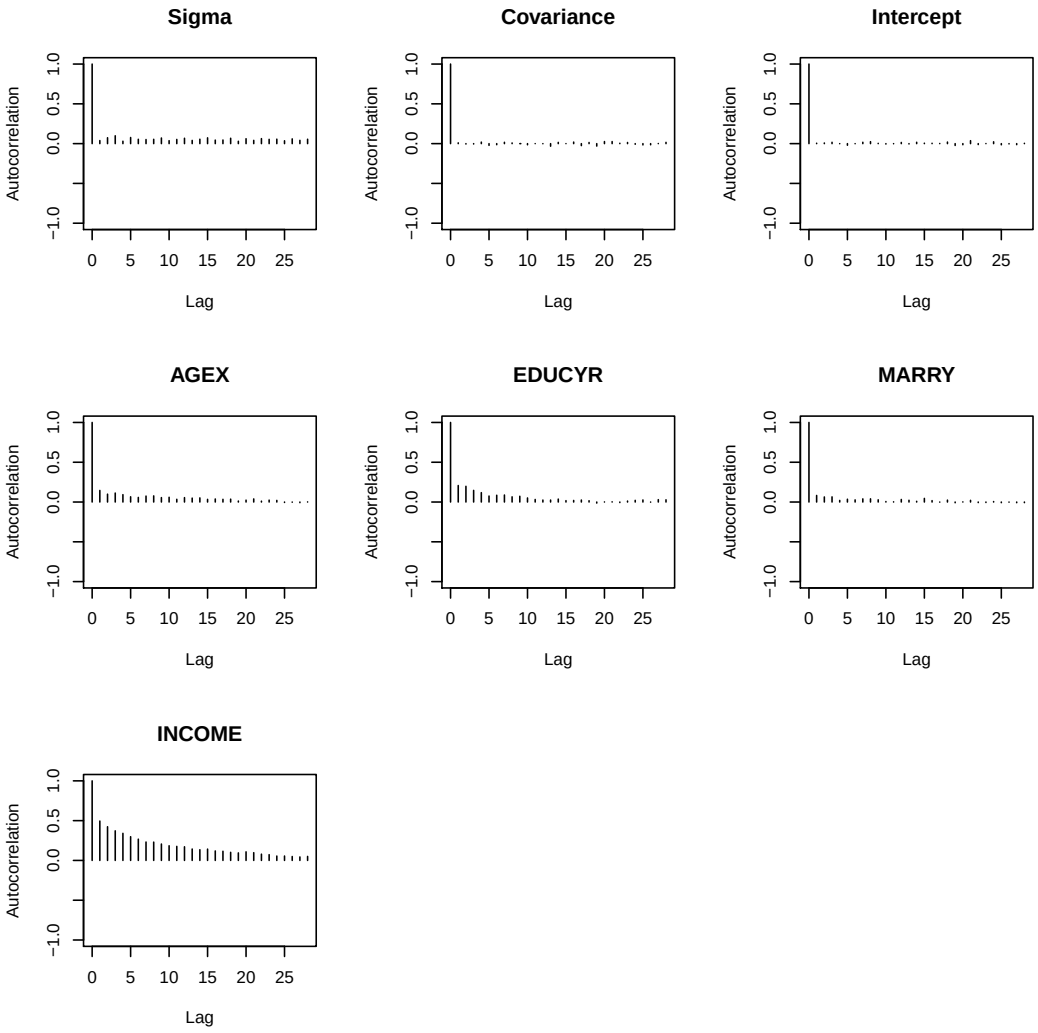

Figure 15: Auto Correlation Plots of Key Parameters for Roy-type DPM Model: Control Group

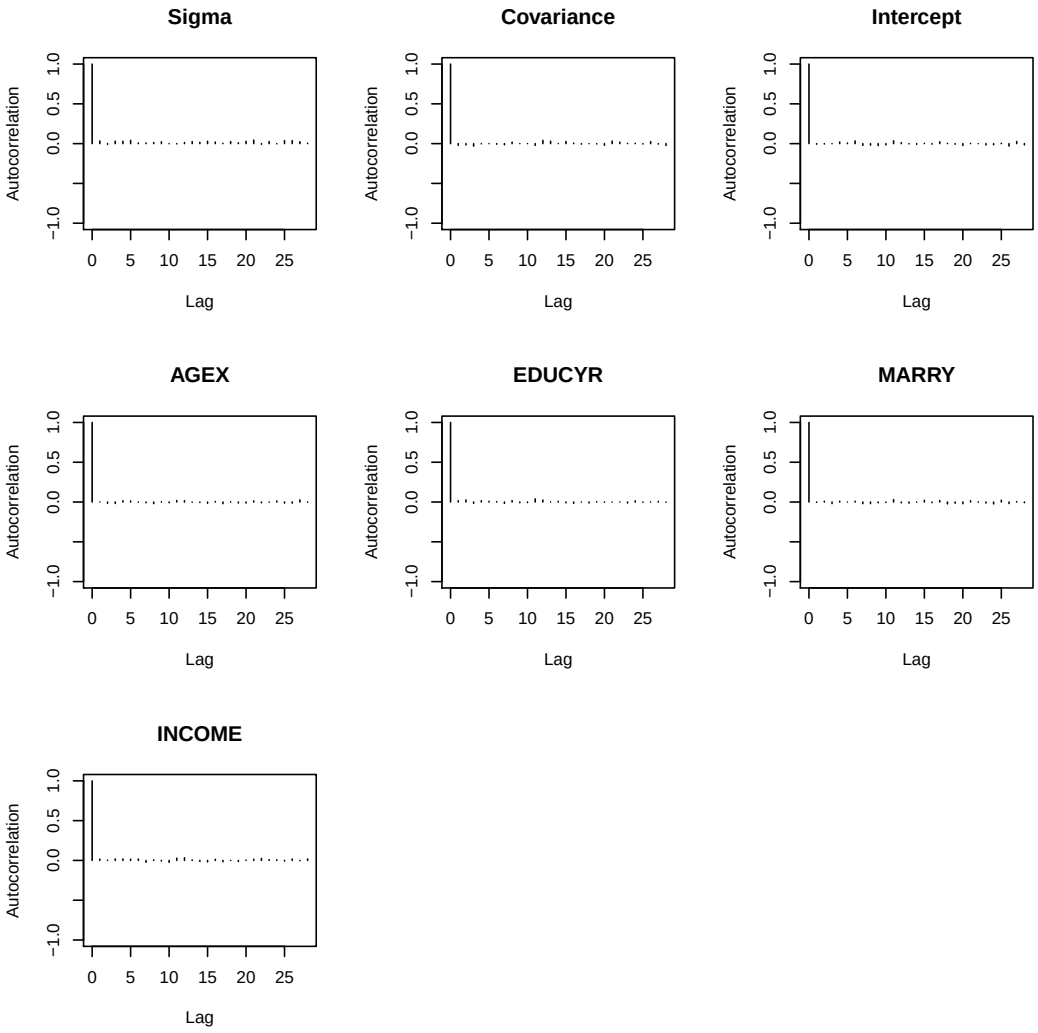

Figure 16: Auto Correlation Plots of Key Parameters for Roy-type DPM Model: Treatment Group

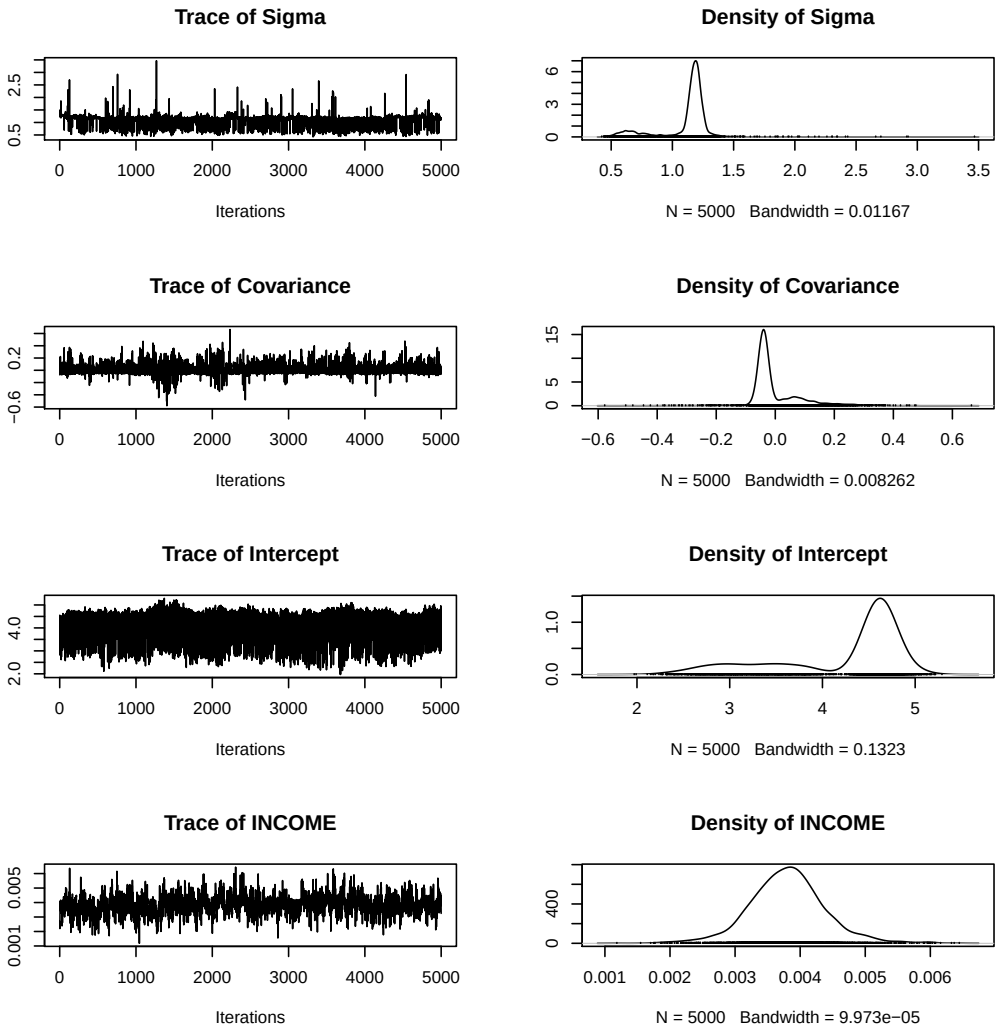

Figure 17: Trace and Density Plots of Key Parameters for Roy-type DPM Model: Control Group

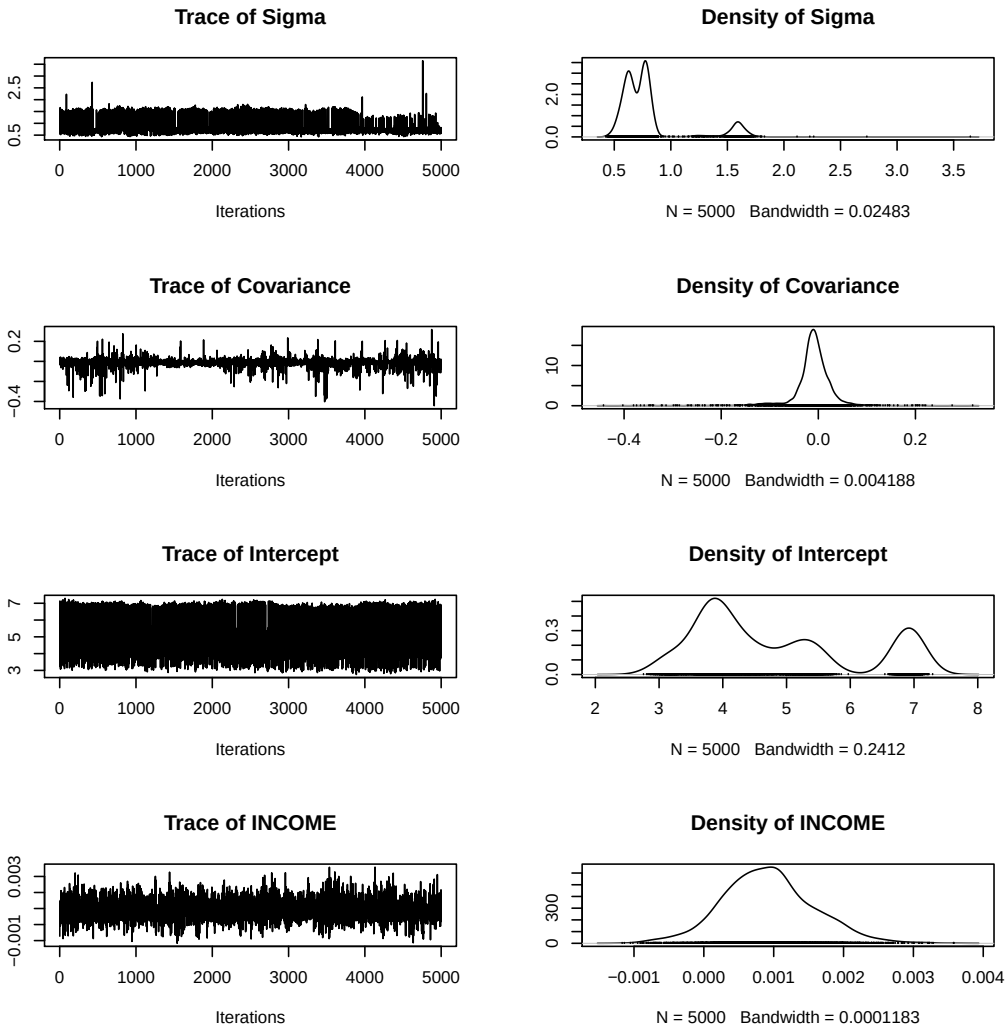

Figure 18: Trace and Density Plots of Key Parameters for Roy-type DPM Model: Treatment Group

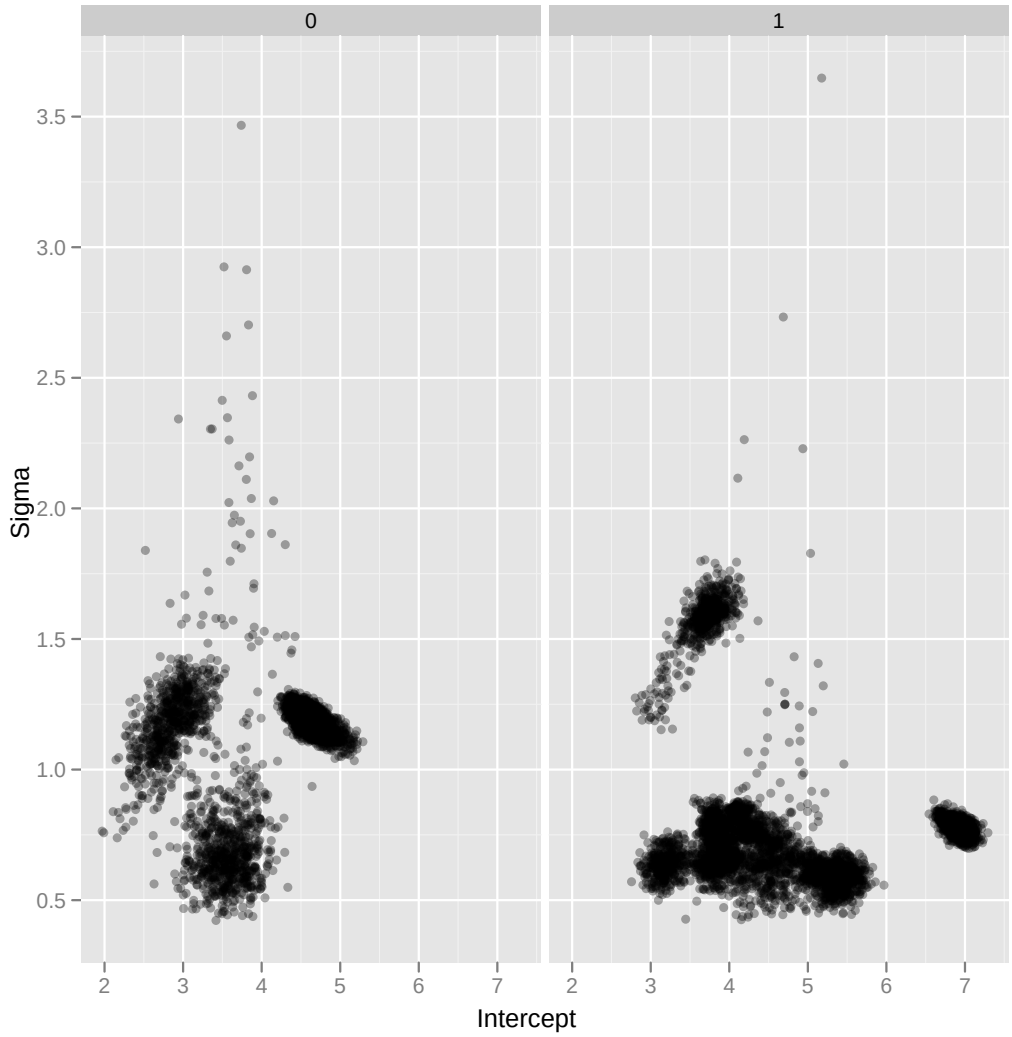

Figure 19: Posterior Scatter Plot of Intercept vs Variance for Roy-type DPM Model

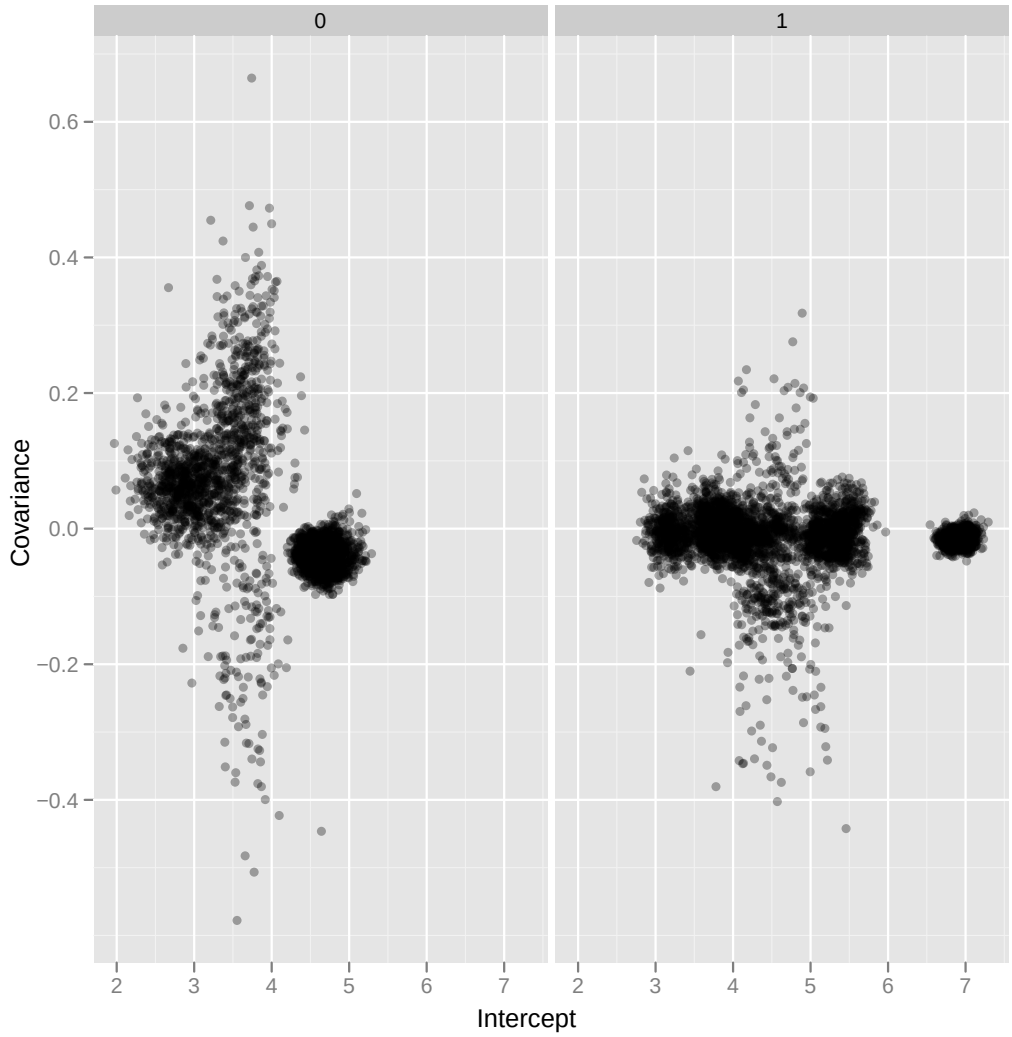

Figure 20: Posterior Scatter Plot of Intercept vs Selection Bias for Roy-type DPM Model

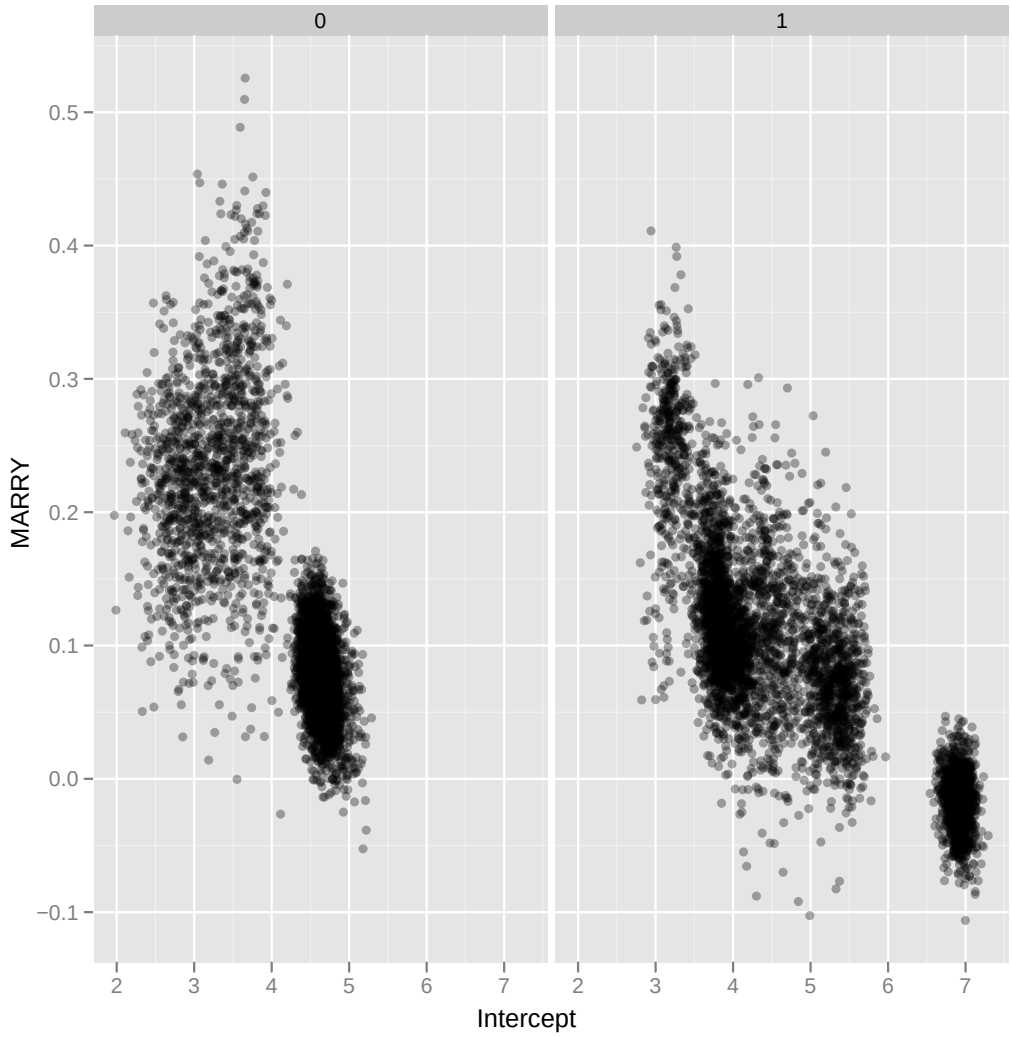

Figure 21: Posterior Scatter Plot of Intercept vs MARRY for Roy-type DPM Model

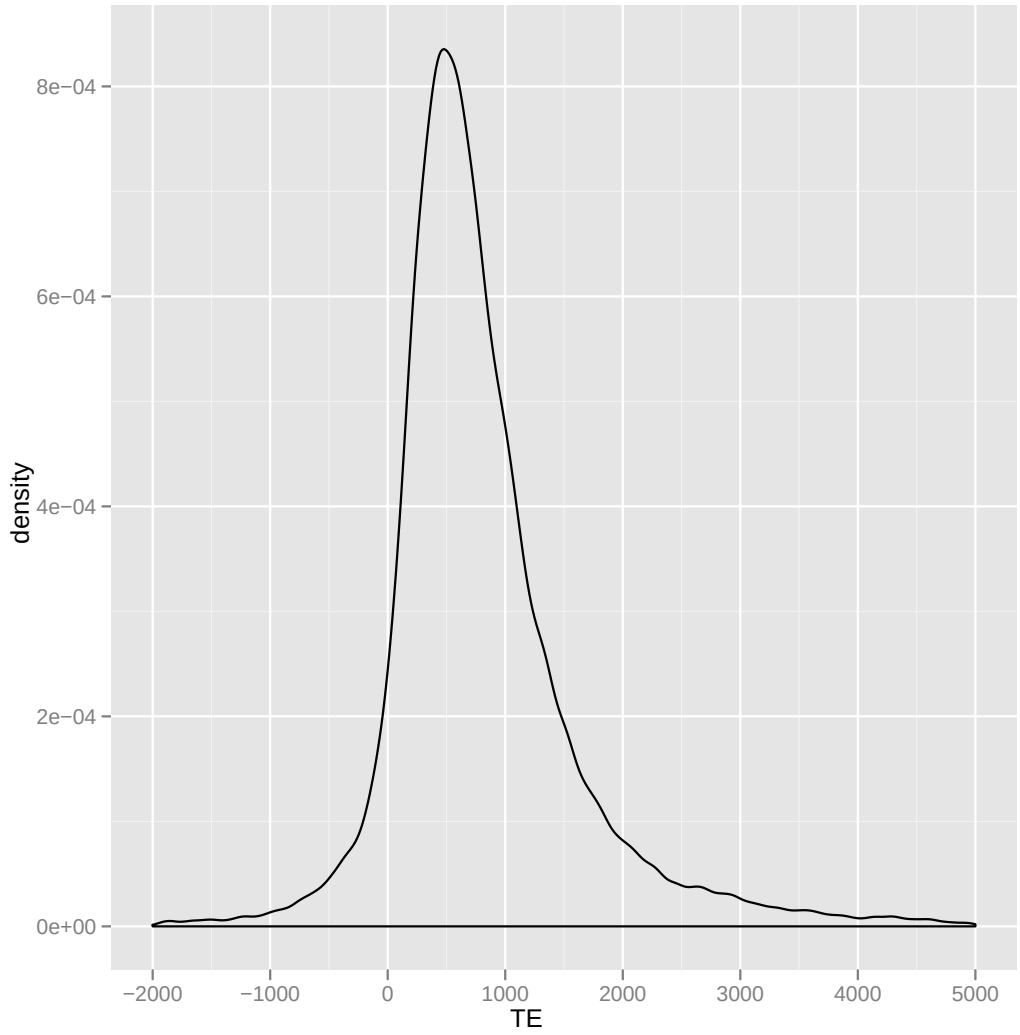

Figure 22: Density Plot of Treatment Effects for Roy-type DPM Model

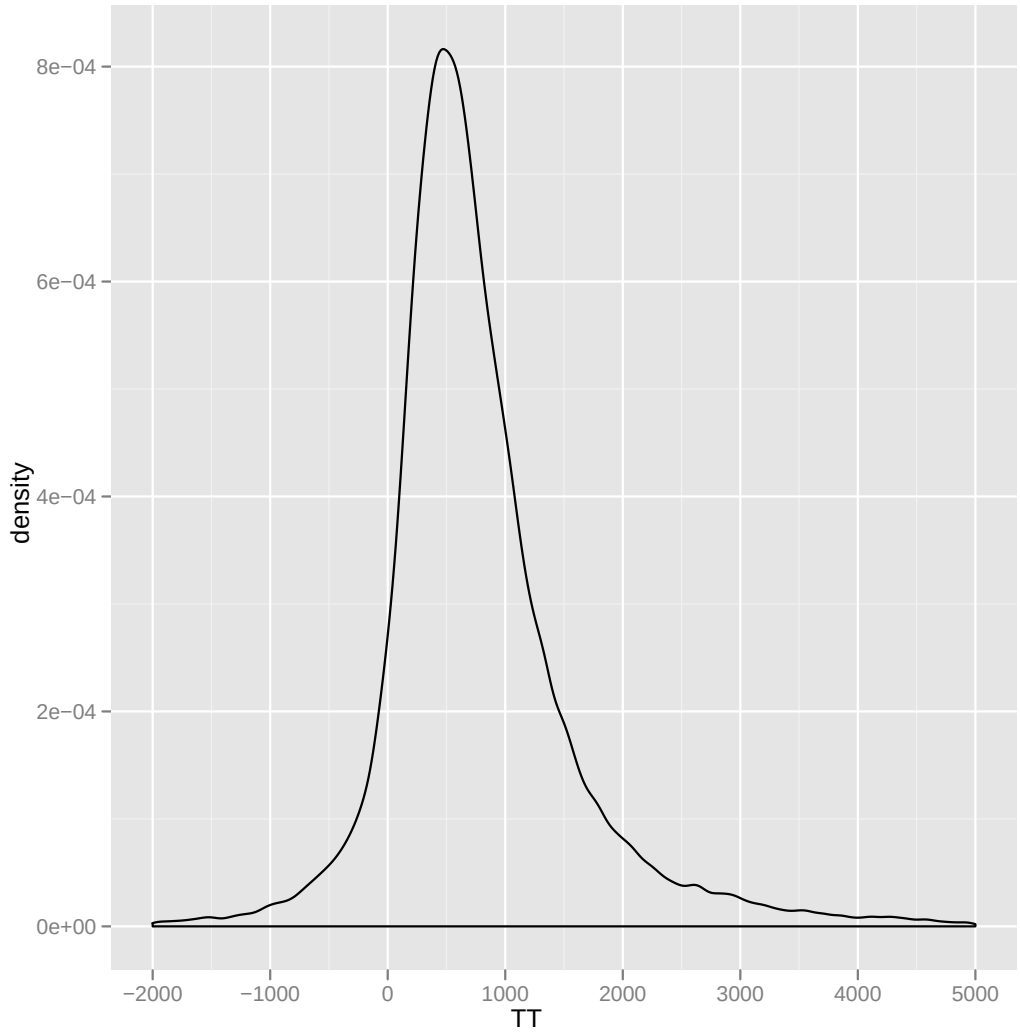

Figure 23: Density Plot of Treatment Effects for the Treated for Roy-type DPM Model

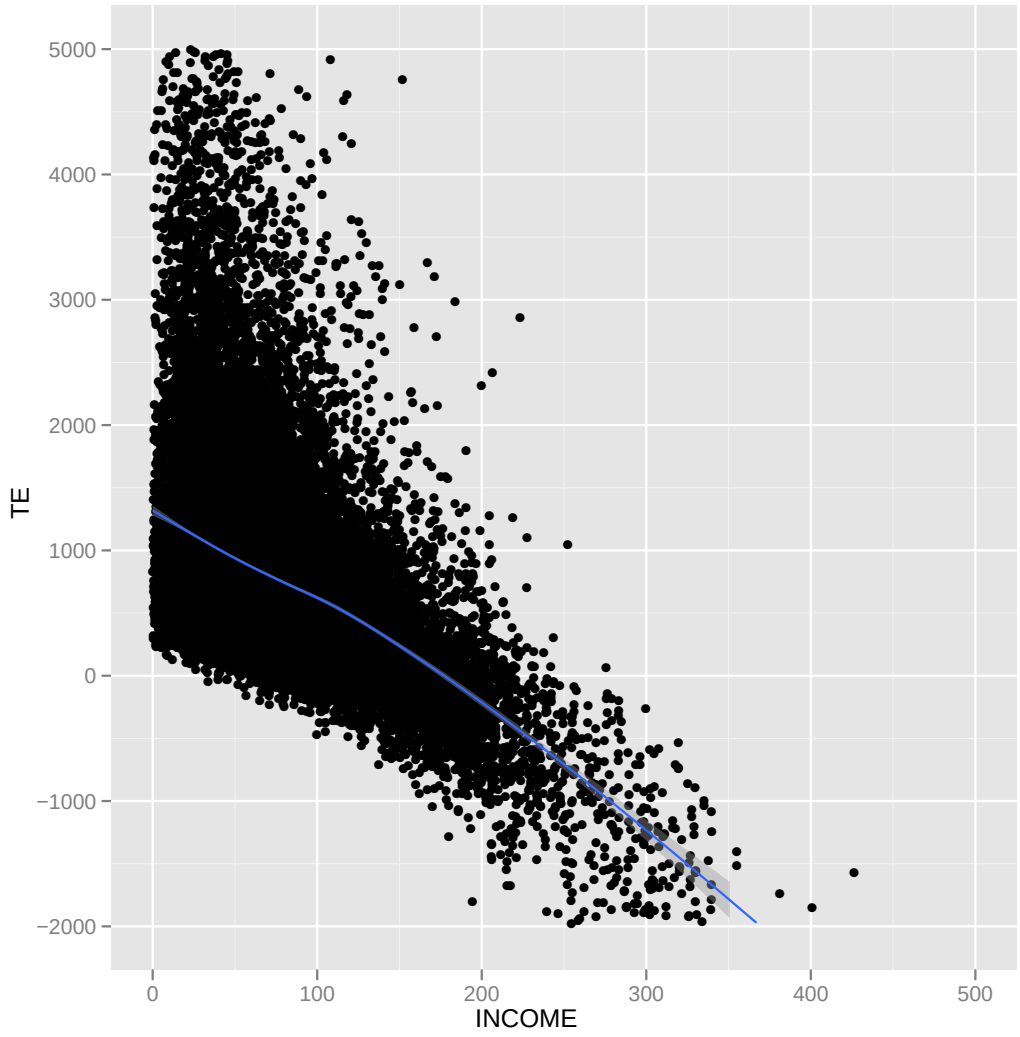

Figure 24: Posterior Scatter Plot of TE vs INCOME for Roy-type DPM Model

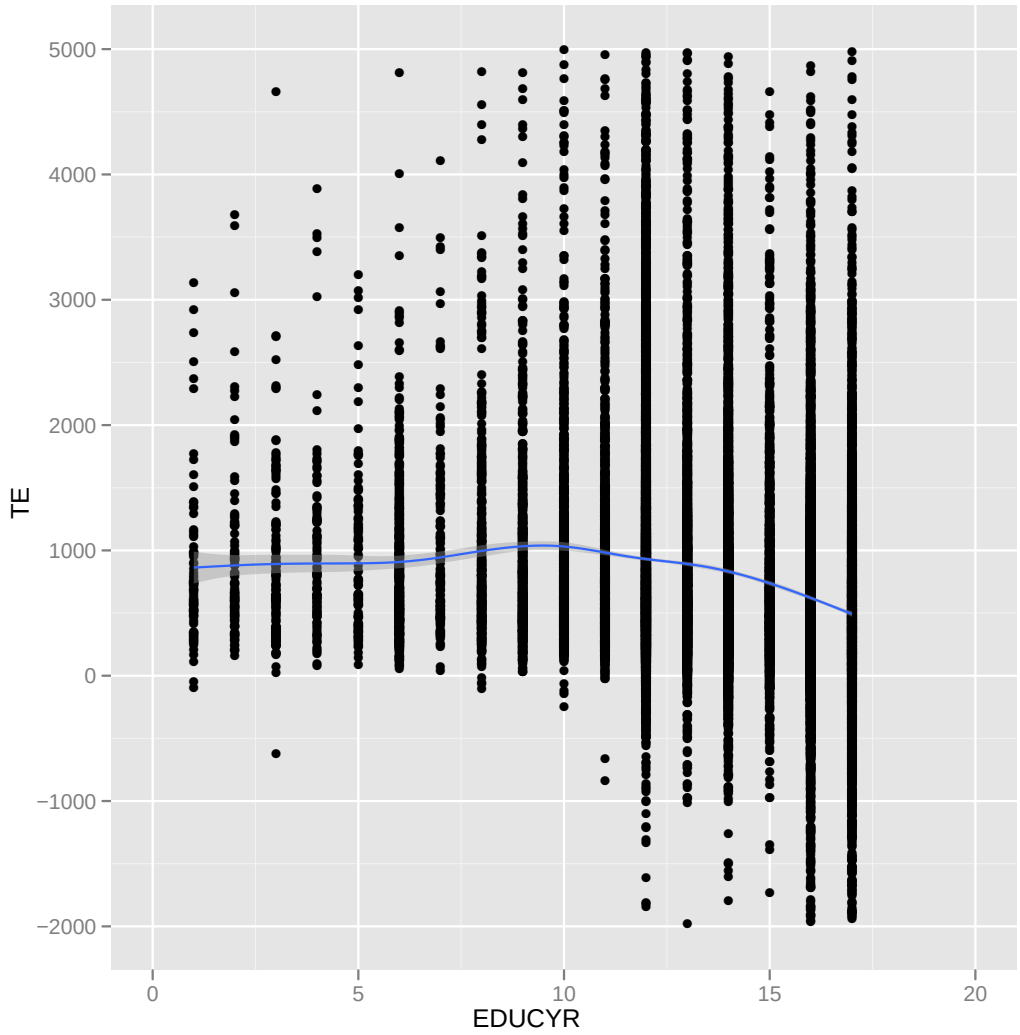

Figure 25: Posterior Scatter Plot of TE vs EDUCYR for Roy-type DPM Model

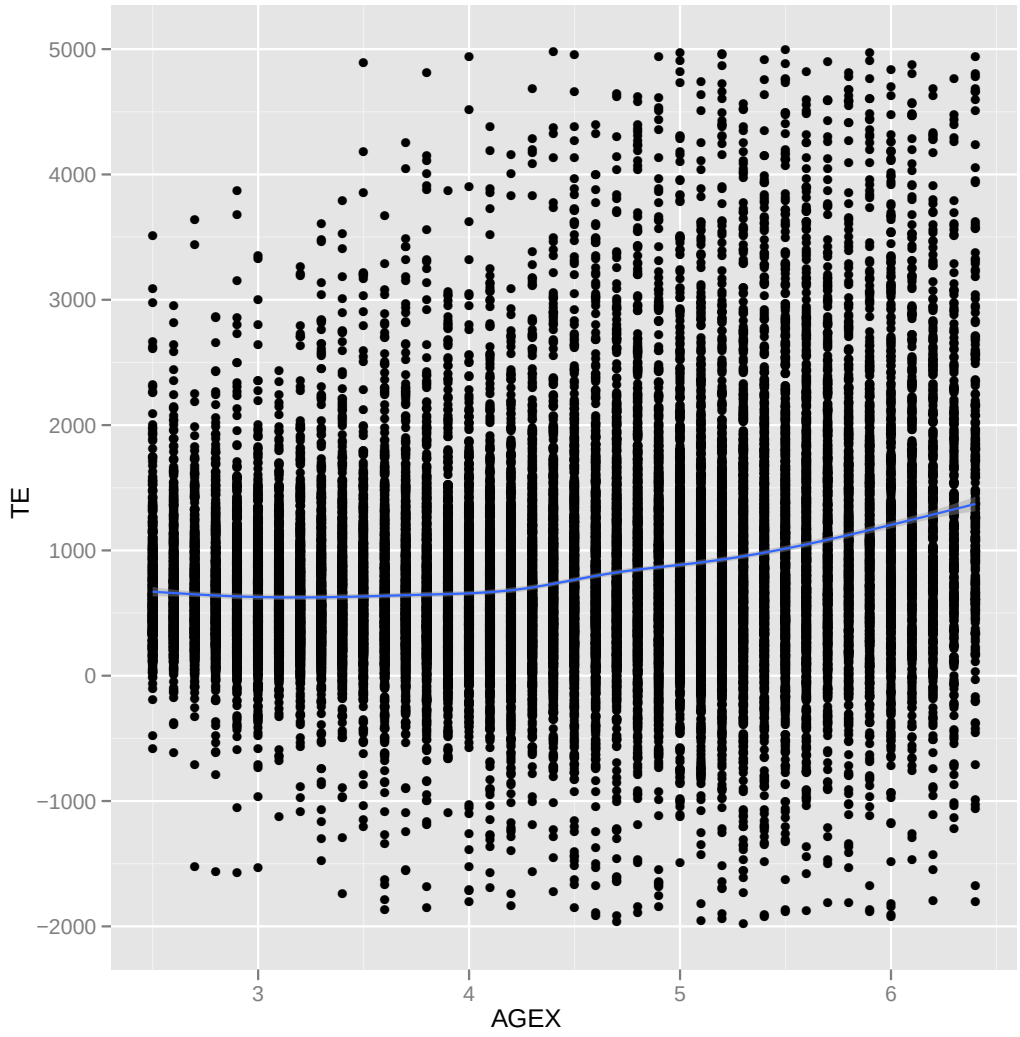

Figure 26: Posterior Scatter Plot of TE vs AGEX for Roy-type DPM Model