Count Roy Model with Finite Mixtures

Murat K. Munkin
Department of Economics
4202 E Fowler Avenue, CMC 208M
University of South Florida
Tampa, FL, 33620, U.S.A.
Email: mmunkin@usf.edu
Phone: +1 (813) 974 6517

July 8, 2021

Abstract

This paper develops the Count Roy Model with Finite Mixtures to model heterogeneity in count variables, that display different overdispersed patterns in two states, defined by an endogenous binary treatment variable. The random permutation sampler is adopted to estimate the numbers of mixture components in the model based on the marginal likelihood. The smoothly mixing regression approach is utilized to model the probabilities of the components. A continuous instrumental variable is allowed to enter the treatment equation nonparametrically.

The model is applied to estimating treatment effects of Medigap private insurance on the demand for prescription drugs for the US elderly unemployed Medicare population. A strong evidence is found that there are two components both in the treated and untreated states. The estimated treatment effects produce evidence of adverse selection.

Key words: Finite Mixtures; Count Roy Model; Random Permutation Sampler; Smoothly Mixing Regressions; Nonparametric Instrument.

JEL Codes: C11, C14

1. Introduction

This paper uses a semiparametric Finite Mixture approach to extend the Roy model with endogenous selection for count outcomes, developed by Deb, Munkin and Trivedi (2006), to account for unobserved heterogeneity in a more flexible semiparametric way. The model specification is motivated by observed consumption patterns for prescription drugs, which display overdispersion and multimodal frequencies with long tails, different in the treated and untreated states. In such cases the Roy model for count outcomes, later referred to as the Count Roy (CR) model, can produce unrealistic marginal and treatment effects. The objective is to identify components within the data for both states, where the assumption of homogeneity in marginal and treatment effects is potentially more realistic.

Various extensions of the Poisson model have been developed in economics to analyze overdispersed count data. Most commonly used models, Negative-Binomial and Poisson-Lognormal, can be viewed as continuous mixtures, in which the random variables, representing unobserved heterogeneity, are distributed as Gamma and Lognormal respectively. The Roy modeling framework controls for observed heterogeneity captured by the treatment variable. The proposed Finite Mixture Roy (FMR) model further controls for unobserved heterogeneity using finite mixtures, assuming that the components follow Poisson-Lognormal distributions both in the treated and untreated states. The developed FMR model has several additional features, influenced by the empirical application, which analyzes the effect of supplemental Medigap private insurance on the demand for prescription drugs for the elderly US Medicare population. Medigap status defines the endogenous treatment variable and the numbers of components in the mixtures can be different for those with and without Medigap insurance coverage.

To better understand what factors influence belonging to a latent component the corresponding

probabilities are modeled as functions of observed covariates. I assume that these probabilities are generated by a multinomial probit model with a constrained covariance matrix as in the smoothly mixing regressions (SMR) approach by Geweke and Keane (2007). Other specifications are possible, including multinomial logit in Villani, Kohn and Nott (2012) and Frühwirth-Schnatter et al. (2012, 2016), however, the normality assumption, adopted in this paper, is a convenient framework for modeling unobserved heterogeneity and potential endogeneity of the binary treatment variable.

Identification of the treatment effects relies on a constructed continuous instrumental variable, defined as the ratio of the annual supplemental security income received by individuals to their annual personal income. The application section argues that the instrument is likely to affect the treatment variable nonlinearly, and therefore, the FMR model allows it to enter the treatment equation nonparametrically. The results are then compared with those of the linear specification. The estimated marginal likelihoods of the competing models give strong support to the nonparametric specification.

Estimation of finite mixtures both in the frequentist and Bayesian frameworks have some subtle points (Celeux et al., 2019b). Gormley and Frühwirth-Schnatter (2019) discuss identifiability of mixture of experts (ME) models and distinguish among three types of non-identifiability: invariance to relabelling the components, potential overfitting and generic non-identifiability, occurring only for certain distributions. The label switching problem was studied by Celeux, Hurn and Robert (2000), Frühwirth-Schnatter (2001) and Jasra, Holmes and Stephens (2005). Given the invariance of the likelihood to label switching of k components, the sampler fails to visit some of $k!$ regions in the support of the posterior distribution. In case of overfitting the corresponding Markov chains usually display poor mixing properties, something that also happens in the application section when more components are specified than the data support. Therefore, it is important to identify the

correct numbers of components.

Gormley and Frühwirth-Schnatter (2019) indicate that generic identifiability is an open research question for general mixture-of-experts models and must be verified on a case-by-case basis. Once the correct numbers of components in the model are identified in order to conduct component-specific inference this paper applies the method of Geweke (2007), in which separation of the components is done by imposing valid inequality constraints on the draw. The approach works well when there is a large number of observations and a relatively small number of components in the mixtures, which is the case in the application. The estimation results show that both in the treated and untreated states the components of the finite mixture model are well identified. The corresponding posterior distributions have clear signs of convergence for all parameters, component specific means and weights.

This paper takes the approach of Chib (1995) combined with the random permutation sampler to calculate the marginal likelihoods in selecting the numbers of components, although other approaches exist. Frühwirth-Schnatter, Celeux and Robert (2019), Celeux et al. (2019a) provide a comprehensive review of such methods for finite mixture models. Frühwirth-Schnatter (2004) considers three sampling-based techniques, which include importance sampling, reciprocal importance sampling and bridge sampling, and shows that reliability of the first two estimators for finite mixtures depends on the tail behavior of the proposal density. Lee and Robert (2016) point out that all three methods are subjects to biases unless the importance density exhibits the same kind of multimodality as the posterior. That includes the method of Chib (1995) which can be derived as a bridge sampling estimator, as shown by Mira and Nicholls (2004). Frühwirth-Schnatter (2004) shows that the method of Chib (1995) for finite mixture models is biased without perfect symmetry in label switching, however, the bias can be corrected if the estimator is combined with the random

permutation sampler. Frühwirth-Schnatter (2001) designs the random permutation sampler that randomly switches labels among all $k!$ regions so that all labeling subspaces are visited in a balanced fashion. This idea was also utilized by Berkhof et al. (2003), Fruhwirth-Schnatter (2004, 2006), Lee and Robert (2016) and Fruhwirth-Schnatter (2019). Since the marginal likelihood function of a k component mixture is permutation invariant the output from the random permutation sampler can be used to estimate the Bayes factors using the approach of Chib (1995).

There are alternative approaches to identifying the correct numbers of components that treat them as unknown parameters. These include overfitting mixtures by Rousseau and Mengersen (2011), reversible jump MCMC (Markov chain Monte Carlo) by Richardson and Green (1997) and Bayesian nonparametric (BNP) methods with potentially infinite but countable numbers of components. The most widely used BNP prior is the Dirichlet process (DP) prior with applications given in Burda, Harding and Hausman (2008), Müller and Mitra (2013) and Hu, Munkin and Trivedi (2015). Additional discussion of the methods and references can be found in Celeux et al. (2019a).

Finite mixtures for count variables in the frequentist framework are studied in Wedel et al. (1993), Deb and Trivedi (1997), Bago d'Uva (2006), Hyppolite and Trivedi (2012), Deb and Trivedi (2013) and Karlis, Papatla and Roy (2016). Bayesian treatments of Poisson mixture models include Chib (1996), Viallefont, Richardson and Green (2002), Hurn, Justel and Robert (2003), Villani et al. (2012), Burda, Harding and Hausman (2012).

The rest of the paper is organized as follows. Section 2 specifies the FMR model and connects it to related studies, defines parameter priors and develops an MCMC algorithm. Section 3 outlines estimation of marginal likelihoods to identify the numbers of components in the mixtures. Section 4 presents an application and Section 5 concludes. The details of the MCMC algorithm and steps

in calculating the marginal likelihoods are given in the computational appendix.

2. The FMR Model

This section defines the Finite Mixture Roy model, discusses and motivates the choice of parameter priors and outlines the MCMC inference.

2.1. The model

Assume that dependent count variable Y_i is observed for N independent individuals ($i = 1, ..., N$). First, define the treatment equation and let d_i denote a potentially endogenous binary treatment variable. It is assumed that the observed value of d_i is generated by latent variable D_i, which measures the difference in utility derived by individual i in the treated and untreated states respectively such that

$$d_i = I_{[0,+\infty)}\left(D_i\right), \tag{2.1}$$

where $I_{[0,+\infty)}$ is the indicator function for the set $[0, +\infty)$. A valid instrumental variable is allowed to affect D_i in a flexible nonparametric way. The treatment equation is specified as

$$D_i = f(s_i) + \mathbf{W}_i\boldsymbol{\alpha} + u_i, \tag{2.2}$$

where $\mathbf{W}_i$ is a vector of exogenous regressors, $\boldsymbol{\alpha}$ is a conformable vector of parameters, which does not include an intercept, function $f(.)$ is unknown, s_i is a continuous instrumental variable and the distribution of the error term is $u_i \overset{iid}{\sim} N\left(0, 1\right)$.

Next, define the count dependent variable Y_i assuming that there are two potential outcomes Y_i^1 and Y_i^2 and the observability condition is

$$Y_i = \begin{cases} Y_i^1 & \text{if } d_i = 1 \\ Y_i^2 & \text{if } d_i = 0 \end{cases}.$$

Assume also that Y_i^1 and Y_i^2 are distributed as finite mixtures of Poisson-lognormal densities with means $\exp(\mu_{ij}^1)$ and $\exp(\mu_{ij}^2)$ respectively, where subscript j indicates that observation i belongs to component j. In general the numbers of components are different for the treated and untreated states for which additional notations are introduced below. Variables μ_{ij}^1, μ_{ij}^2 are assumed to be linear in the set of exogenous regressors $\mathbf{X}_i$ and u_i, the error of the treatment equation, such that

$$\begin{aligned} \mu_{ij}^1 &= \mathbf{X}_i\boldsymbol{\beta}_{1j} + \delta_{1j}u_i + \varepsilon_{ij}^1, \\ \mu_{ij}^2 &= \mathbf{X}_i\boldsymbol{\beta}_{2j} + \delta_{2j}u_i + \varepsilon_{ij}^2, \end{aligned}$$

where $\varepsilon_{ij}^1 \sim N\left(0, \sigma_{1j}^2\right)$ and $\varepsilon_{ij}^2 \sim N\left(0, \sigma_{2j}^2\right)$ represent unobserved heterogeneity. Random variable u_i is introduced in the conditional means to control for the unobservable factors affecting both the treatment choice and utilization. Once these factors are controlled for, the error terms ε_{ij}^1 and ε_{ij}^2 can be assumed to satisfy the conditions $cov(u_i, \varepsilon_{ij}^1|\mathbf{X}) = 0$, $cov(u_i, \varepsilon_{ij}^2|\mathbf{X}) = 0$. The correlation between ε_{ij}^1 and ε_{ij}^2 is not identifiable because just one outcome is observed for each individual. I restrict it to zero, $cov\left(\varepsilon_{ij}^1, \varepsilon_{ij}^2|u_i\right) = 0$. Since u_i, ε_{ij}^1 and ε_{ij}^2 are normally distributed the resulting distributions of the components, unconditional of latent variables u_i, ε_{ij}^1 and ε_{ij}^2, are Poisson-Lognormal distributions, which do not have a closed form solution.

It is important to mention that this model could be introduced differently: through a joint normal distribution of u_i, ε_{ij}^1, ε_{ij}^2, denoted as $N(\mathbf{0}, \boldsymbol{\Sigma})$. Specifically, Deb et al. (2006) start by defining the joint distribution of the errors. Then positive definiteness of covariance matrix $\boldsymbol{\Sigma}$ imposes restrictions on the correlation parameters $corr\left(\varepsilon_{ij}^1, \varepsilon_{ij}^2\right)$. Poirier and Tobias (2003) and Li, Poirier and Tobias (2004) show that learning about this non-identified correlation can take place through the identified correlation between u_i and ε_{ij}^1, ε_{ij}^2 respectively, and carefully chosen priors. However, I present the model differently first specifying the marginal distribution of u_i and then the conditional distribution of ε_{ij}^1 and ε_{ij}^2 given u_i. Then the assumed correlation between ε_{ij}^1 and

ε_{ij}^2 is, in fact, the conditional correlation $corr\left(\varepsilon_{ij}^1, \varepsilon_{ij}^2 | u_i\right)$. Poirier and Tobias (2003) indicate that no learning takes place about the non-identified correlation parameter after taking into account the linear effect of u_i. Subsequently this correlation in the FMR model is restricted to zero, which might not be consistent with the true correlation value, and, therefore, the results must be interpreted only conditionally on this assumption. Alternatively, Chib (2007) defines a potential outcomes model specifying full bivariate joint distributions $\left(\varepsilon_{ij}^1, u_i\right)$ and $\left(\varepsilon_{ij}^2, u_i\right)$, thus avoiding specification of the joint distribution of $\left(\varepsilon_{ij}^1, \varepsilon_{ij}^2, u_i\right)$. However, a limitation of this approach is that it allows to estimate only mean treatment effects without distributional treatment effects, which might have a particular policy relevance. Heckman, Lopes, and Piatek (2014) and Jacobi, Wagner and Frühwirth-Schnatter (2016) specify potential outcomes models with a latent factor structure, which allows to estimate the distributional treatment effects.

The general approach by Diebolt and Robert (1994) in estimating finite mixture models augments the posterior with latent variable z_{ij}, defined as

$$z_{ij} = \begin{cases} 1 & \text{if observation } i \text{ belongs to component } j \\ 0 & \text{otherwise} \end{cases}.$$

It draws z_{ij} at each iteration, assigning each observation to a component, and makes it a part of the parameters space. I introduce latent variables z_{ij}^t to differentiate between the treated $(t = 1)$ and untreated $(t = 2)$ states and take the smoothly mixing regressions approach by Geweke and Keane (2007) to allow the corresponding probabilities $\Pr\left(z_{ij}^t = 1\right)$ to depend on covariates. This is done by specifying the multinomial probit model. Let $k_t - 1$ latent variables R_{ij}^t $(j = 2, ..., k_t)$ be defined as

$$R_{ij}^t = \mathbf{V}_i \boldsymbol{\gamma}_{tj} + \xi_{ij}^t, \tag{2.3}$$

where $\mathbf{V}_i$ is a set of covariates (it can be different from $\mathbf{X}_i$), $\boldsymbol{\gamma}_{tj}$ is a conformable vector of para-

meters, $\boldsymbol{\xi}_i^t \overset{iid}{\sim} N(0, \mathbf{I}_{k-1})$ and R_{i1}^t is restricted to zero. Then the components are identified as

$$z_{ij}^t = 1 \text{ if and only if } R_{ij}^t \geqslant R_{il}^t \text{ (for } \forall l,\ l = 1, ..., k_t). \tag{2.4}$$

In general k_1 and k_2 are different, however, the set of covariates $\mathbf{V}_i$ is assumed to be the same for the treated and untreated states.

2.2. Choice of Priors

I use Bayesian semiparametric techniques presented in Koop and Poirier (2004) and Koop and Tobias (2006) and first sort the data by $\mathbf{s}$ where s_1 denotes the smallest and s_N the largest values. In the application $\mathbf{s} \in [0, 1]$ so that $s_1 = 0$ and $s_N = 1$. Denote k_ν the total number of distinct values of the instrument $\mathbf{s}$. The nonparametric approach treats all k_ν values of $f(s_j)$ $(j = 1, ..., k_\nu)$ as parameters. Potentially $\mathbf{s}$ can take the same number of distinct values as the number of observations, in which case k_ν would be close to N. However, larger values of k_ν would result in larger computational costs potentially producing very little practical benefits in refining the shape of the nonparametric function. If the probability of treatment does not change much for small increments in $\mathbf{s}$, then it can be rounded up to such levels. That should reduce the number of distinct values $f(s_j)$ $(j = 1, ..., k_\nu)$, decreasing the computational burden and producing comparable results. The main assumption made regarding function $f(.)$ is that it is differentiable and its slope does not change too fast over small changes in s_j (Shiller, 1984). The constructed continuous instrumental variable must satisfy this "smoothness" condition.

Stacking (2.2) over i obtains

$$\mathbf{D} = \mathbf{P}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\alpha} + \mathbf{u},$$

where

$$\boldsymbol{\nu} = \begin{bmatrix} f(s_1) \\ f(s_2) \\ ... \\ f(s_{k_\nu}) \end{bmatrix},$$

and $\mathbf{P}$ is an $N \times k_\nu$ matrix constructed to select the appropriate element of ν for each observation i.

Define an $k_\nu \times k_\nu$ matrix $\mathbf{R}$ such that $\boldsymbol{\psi} = \mathbf{R}\boldsymbol{\nu}$ is a vector of slope changes of function $f(.)$,

$$\psi_j = \frac{\nu_j - \nu_{j-1}}{s_j - s_{j-1}} - \frac{\nu_{j-1} - \nu_{j-2}}{s_{j-1} - s_{j-2}}, \quad j = 3, ..., k_\nu,$$

and the first two elements are simply $\psi_1 = f(s_1)$ and $\psi_2 = f(s_2)$. One can think of parameters ψ_j $(j = 3, ..., k_\nu)$ as numerical approximations to the second order derivatives of function $f(s_j)$, calculated at $k_\nu - 2$ points corresponding to $j = 3, ..., k_\nu$. Then

$$\mathbf{D} = \mathbf{P}\mathbf{R}^{-1}\boldsymbol{\psi} + \mathbf{W}\boldsymbol{\alpha} + \mathbf{u}.$$

Specifying priors on the numerical second derivatives places priors on the degree of smoothness of $f(.)$. Parameters ψ_1 and ψ_2 set the initial two points which determine the level of the regression curve and parameters ψ_j $(j = 3, ..., k_\nu)$ set its degree of smoothness. Relatively flatter but still proper priors are set on (ψ_1, ψ_2) as $N(\mathbf{0}_2, \mathbf{I}_2)$. Assume an informative prior

$$\psi_j \sim N(0, \eta), \quad j = 3, ..., k_\nu.$$

where parameter η determines the tightness of the prior for ψ_j. Further, assume that

$$\eta \sim IG(a, b).$$

This should allow the model to learn about the degree of smoothness of the regression curve $f(.)$ instead of simply choosing a fixed value for η. After experimenting with different values of the hyperparameters in the application I select $a = 3$ and $b = 10^3$ and find them to produce smooth posteriors. Proper prior distributions are chosen for the rest of the parameters, since improper

priors can lead to improper posterior (Diebolt and Robert, 1994),

$$\begin{aligned}
\pi(\boldsymbol{\alpha}) \tilde{} N\left(\underline{\boldsymbol{\alpha}}, \underline{\mathbf{H}}_{\alpha}^{-1}\right), \underline{\boldsymbol{\alpha}} &= \mathbf{0},\ \underline{\mathbf{H}}_{\alpha} = 10I \\
\pi(\boldsymbol{\beta}_{tj}) \tilde{} N\left(\underline{\boldsymbol{\beta}}, \underline{\mathbf{H}}_{\beta}^{-1}\right),\ \underline{\boldsymbol{\beta}} &= \mathbf{0},\ \underline{\mathbf{H}}_{\beta} = 10I \\
\pi(\delta_{tj}) \tilde{} N\left(\underline{\delta}, \underline{H}_{\delta}^{-1}\right),\ \underline{\delta} &= 0,\ \underline{H}_{\delta} = 10 \\
\pi(\boldsymbol{\gamma}_{tj}) \tilde{} N\left(\underline{\boldsymbol{\gamma}}, \underline{\mathbf{H}}_{\gamma}^{-1}\right), \underline{\boldsymbol{\gamma}} &= \mathbf{0}\ ,\ \underline{\mathbf{H}}_{\gamma} = 10I \\
\sigma_{tj}^{-2} \tilde{} G\left(n/2, g/2\right),\ n &= 6 \text{ and } g = 1,
\end{aligned}$$

where $j = 1, ..., k_t$, $t = 1, 2$.

2.3. MCMC Inference

I use indicators 1 and 2 as superscripts and subscripts to specify variables and coefficients related to the treated and untreated states respectively and denote $\mathbf{z}_i^1 = (z_{i1}^1, z_{i2}^1, ...z_{ik_1}^1)'$, $\mathbf{z}_i^2 = (z_{i1}^2, z_{i2}^2, ...z_{ik_2}^2)'$, $\mathbf{R}_i^1 = \left(R_{i1}^1, ..., R_{ik_1}^1\right)$, $\mathbf{R}_i^2 = \left(R_{i1}^2, ..., R_{ik_2}^2\right)$, $\boldsymbol{\gamma}^1 = \left(\boldsymbol{\gamma}_{12}', ..., \boldsymbol{\gamma}_{1k_1}'\right)'$, $\boldsymbol{\gamma}^2 = \left(\boldsymbol{\gamma}_{22}', ..., \boldsymbol{\gamma}_{2k_2}'\right)'$, $\boldsymbol{\phi}_{1j}' = (\boldsymbol{\beta}_{1j}', \delta_{1j})$, $\boldsymbol{\phi}_1' = \left(\boldsymbol{\phi}_{11}', ..., \boldsymbol{\phi}_{1k_1}'\right)$, $\boldsymbol{\phi}_{2j}' = (\boldsymbol{\beta}_{2j}', \delta_{2j})$, $\boldsymbol{\phi}_2' = \left(\boldsymbol{\phi}_{21}', ..., \boldsymbol{\phi}_{2k_2}'\right)$. Denote also $\boldsymbol{\mu}_i^1 = \left(\mu_{i1}^1, ..., \mu_{ik_1}^1\right)$, $\boldsymbol{\mu}_i^2 = \left(\mu_{i1}^2, ..., \mu_{ik_2}^2\right)$, $\boldsymbol{\sigma} = \left(\sigma_{11}^2, ..., \sigma_{1k_1}^2, \sigma_{21}^2, ..., \sigma_{2k_2}^2\right)$ and $\boldsymbol{\Delta}_i = (\mathbf{X}_i, \mathbf{W}_i, \mathbf{V}_i, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \boldsymbol{\psi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}^{\mathbf{1}}, \boldsymbol{\gamma}^{\mathbf{2}}, \boldsymbol{\sigma})$. The MCMC algorithm is augmented with latent variables $\boldsymbol{\mu}_i^1$, $\boldsymbol{\mu}_i^2$, $\mathbf{R}_i^1$, $\mathbf{R}_i^2$, $\mathbf{z}_i^1$ and $\mathbf{z}_i^2$. The joint density of observed and latent data for individual i is

$$\begin{aligned}
f(D_i, d_i, \boldsymbol{\mu}_i^1, \boldsymbol{\mu}_i^2, \mathbf{R}_i^1, \mathbf{R}_i^2, \mathbf{z}_i^1, \mathbf{z}_i^2, Y_i^1, Y_i^2, Y_i | \boldsymbol{\Delta}_i) &= p(D_i | \boldsymbol{\Delta}_i) p(d_i | D_i, \boldsymbol{\Delta}_i) \\
&\quad \times p(\boldsymbol{\mu}_i^1, \boldsymbol{\mu}_i^2 | d_i, D_i, \boldsymbol{\Delta}_i) \\
&\quad \times p\left(\mathbf{R}_i^1, \mathbf{R}_i^2 | \boldsymbol{\mu}_i^1, \boldsymbol{\mu}_i^2, d_i, D_i, \boldsymbol{\Delta}_i\right) \\
&\quad \times p\left(\mathbf{z}_i^1, \mathbf{z}_i^2, Y_i^1, Y_i^2, Y_i | \mathbf{R}_i^1, \mathbf{R}_i^2, \boldsymbol{\mu}_i^1, \boldsymbol{\mu}_i^2, d_i, D_i, \boldsymbol{\Delta}_i\right)
\end{aligned}$$

or

$$\frac{1}{\sqrt{2\pi}}\exp\left[-0.5\left(D_i-\mathbf{P}_i\mathbf{R}^{-1}\boldsymbol{\psi}-\mathbf{W}_i\boldsymbol{\alpha}\right)^2\right]\left[d_iI_{[0,\infty)}\left(D_i\right)+\left(1-d_i\right)I_{(-\infty,0]}\left(D_i\right)\right]$$
$$\times\prod_{j=1}^{k_1}\left(\frac{\exp\left[-0.5\sigma_{1j}^{-2}\left(\mu_{ij}^1-\mathbf{X}_i\boldsymbol{\beta}_{1j}-\delta_{1j}u_i\right)^2\right]}{\sqrt{2\pi\sigma_{1j}^2}}\right)^{d_i}\prod_{j=1}^{k_2}\left(\frac{\exp\left[-0.5\sigma_{2j}^{-2}\left(\mu_{ij}^2-\mathbf{X}_i\boldsymbol{\beta}_{2j}-\delta_{2j}u_i\right)^2\right]}{\sqrt{2\pi\sigma_{2j}^2}}\right)^{1-d_i}$$
$$\times\left(\prod_{j=2}^{k_1}\frac{\exp\left[-0.5\left(R_{ij}^1-\mathbf{V}_i\boldsymbol{\gamma}_{1j}\right)^2\right]}{(2\pi)^{(k_1-1)/2}}\right)^{d_i}\left(\prod_{j=2}^{k_2}\frac{\exp\left[-0.5\left(R_{ij}^2-\mathbf{V}_i\boldsymbol{\gamma}_{2j}\right)^2\right]}{(2\pi)^{(k_2-1)/2}}\right)^{(1-d_i)}$$
$$\times\left[\sum_{j=1}^{k_1}I_{\left\{z_{ij}^1=1\right\}}I_{\left\{Y_{ij}^1=Y_i\right\}}\left(\prod_{l=1}^{k_1}I_{(-\infty,R_{ij}^1]}\left(R_{il}^1\right)\right)\right]^{d_i}\left[\sum_{j=1}^{k_2}I_{\left\{z_{ij}^2=1\right\}}I_{\left\{Y_{ij}^2=Y_i\right\}}\left(\prod_{l=1}^{k_2}I_{(-\infty,R_{ij}^2]}\left(R_{il}^2\right)\right)\right]^{(1-d_i)}$$
$$\times\left(d_i\sum_{j=1}^{k_1}\frac{\exp\left(-\exp\left(\mu_{ij}^1\right)\right)\exp\left(\mu_{ij}^1Y_i\right)}{Y_i!}I_{\left\{z_{ij}^1=1\right\}}+\left(1-d_i\right)\sum_{j=1}^{k_2}\frac{\exp\left(-\exp\left(\mu_{ij}^2\right)\right)\exp\left(\mu_{ij}^2Y_i\right)}{Y_i!}I_{\left\{z_{ij}^2=1\right\}}\right)$$

The joint density for all observations is

$$f(\mathbf{D},\mathbf{d},\boldsymbol{\mu}^1,\boldsymbol{\mu}^2,\mathbf{R}^1,\mathbf{R}^2,\mathbf{z}^1,\mathbf{z}^2,\mathbf{Y}^1,\mathbf{Y}^2,\mathbf{Y}|\boldsymbol{\Delta})=\prod_{i=1}^{N}f(D_i,d_i,\boldsymbol{\mu}_i^1,\boldsymbol{\mu}_i^2,\mathbf{R}_i^1,\mathbf{R}_i^2,\mathbf{z}_i^1,\mathbf{z}_i^2,Y_i^1,Y_i^2,Y_i|\boldsymbol{\Delta}_i).$$

The full conditional kernel is the product of the joint density and the prior distributions. The details of the MCMC algorithm are presented in Computational Appendix A1.

3. Model Selection

The objective of this section is to calculate marginal likelihood values to select among several model specifications. Denote $M\left(k_1,k_2\right)$ a model in which there are k_1 components in the treated state and k_2 components in the untreated state. The method of Chib (1995) to estimate marginal likelihoods is adopted in combination with the random permutation sampler. The random permutation sampler randomly permutes the components and their corresponding parameters at the end of each iteration. Block all parameters as $\boldsymbol{\sigma}$, $\boldsymbol{\phi}=[\boldsymbol{\phi}_1,\boldsymbol{\phi}_2]$, $\boldsymbol{\gamma}=\left[\boldsymbol{\gamma}^1,\boldsymbol{\gamma}^2\right]$, η, $[\boldsymbol{\psi},\boldsymbol{\alpha}]$ and denote $\boldsymbol{\theta}=\left(\boldsymbol{\sigma},\boldsymbol{\phi}_1,\boldsymbol{\phi}_2,\boldsymbol{\gamma}^1,\boldsymbol{\gamma}^2,\eta,\boldsymbol{\psi},\boldsymbol{\alpha}\right)$. The marginal likelihood can be written as

$$m(y)=\frac{l(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}|y)}.$$

The logarithm of the marginal likelihood is estimated at the posterior mean of the parameters $\boldsymbol{\theta}^*$ as

$$\log m(y) = \log l(y|\boldsymbol{\theta}^*) + \log \pi(\boldsymbol{\theta}^*) - \log \pi(\boldsymbol{\theta}^*|y),$$

where $l(y|\boldsymbol{\theta}^*)$, $\pi(\boldsymbol{\theta}^*)$ and $\pi(\boldsymbol{\theta}^*|y)$ are the likelihood, prior density and posterior ordinate evaluated at the posterior mean. The posterior ordinate can be written by the law of total probability as

$$\pi(\boldsymbol{\theta}^*|y) = \pi(\boldsymbol{\sigma}^*|y)\pi(\boldsymbol{\phi}^*|y,\boldsymbol{\sigma}^*)\pi(\boldsymbol{\gamma}^*|y,\boldsymbol{\sigma}^*,\boldsymbol{\phi}^*)\pi(\eta^*|y,\boldsymbol{\sigma}^*,\boldsymbol{\phi}^*,\boldsymbol{\gamma}^*)\pi(\boldsymbol{\psi}^*,\boldsymbol{\alpha}^*|y,\boldsymbol{\sigma}^*,\boldsymbol{\phi}^*,\boldsymbol{\gamma}^*,\eta^*).$$

Then the logarithm of the posterior ordinate could be estimated as

$$\begin{aligned}\log \widehat{\pi}(\boldsymbol{\theta}^*|y) &= \log \widehat{\pi}\left(\boldsymbol{\sigma}^*|y\right) + \log \widehat{\pi}(\boldsymbol{\phi}^*|y,\boldsymbol{\sigma}^*) + \log \widehat{\pi}(\boldsymbol{\gamma}^*|y,\boldsymbol{\sigma}^*,\boldsymbol{\phi}^*) + \log \widehat{\pi}(\eta^*|y,\boldsymbol{\sigma}^*,\boldsymbol{\phi}^*,\boldsymbol{\gamma}^*) \\ &\quad + \log \widehat{\pi}(\boldsymbol{\psi}^*,\boldsymbol{\alpha}^*|y,\boldsymbol{\sigma}^*,\boldsymbol{\phi}^*,\boldsymbol{\gamma}^*,\eta^*).\end{aligned}$$

The following priors are used when identifying the number of components

$$\begin{aligned}\pi(\boldsymbol{\alpha}) \tilde{} N\left(\underline{\boldsymbol{\alpha}}, \underline{\mathbf{H}}_{\alpha}^{-1}\right),\ \underline{\boldsymbol{\alpha}} &= \mathbf{0},\ \underline{\mathbf{H}}_{\alpha} = 0.1I \\ \pi(\boldsymbol{\beta}_{tj}) \tilde{} N\left(\underline{\boldsymbol{\beta}}, \underline{\mathbf{H}}_{\beta}^{-1}\right),\ \underline{\boldsymbol{\beta}} &= \mathbf{0},\ \underline{\mathbf{H}}_{\beta} = 0.1I \\ \pi(\delta_{tj}) \tilde{} N\left(\underline{\delta}, \underline{H}_{\delta}^{-1}\right),\ \underline{\delta} &= 0,\ \underline{H}_{\delta} = 0.1 \\ \pi(\boldsymbol{\gamma}_{tj}) \tilde{} N\left(\underline{\boldsymbol{\gamma}}, \underline{\mathbf{H}}_{\gamma}^{-1}\right), \underline{\boldsymbol{\gamma}} &= \mathbf{0}\ ,\ \underline{\mathbf{H}}_{\gamma} = 0.1I \\ \sigma_{tj}^{-2} \tilde{} G\left(n/2, g/2\right),\ n &= 6 \text{ and } g = 1,\end{aligned}$$

where $j = 1, ..., k_1$ if $t = 1$ (treated parameters) and $j = 1, ..., k_2$ if $t = 2$ (untreated). The priors are invariant to relabelling of the components. The details of the method in estimating the likelihood and posterior ordinate are given in Computational Appendix A2.

4. Application

This section analyzes the effect of supplemental Medigap private insurance on the demand for prescription drugs by the elderly US Medicare population. Deb and Trivedi (1997) analyzed the

demand for six measures of medical care for the elderly using a frequentist finite mixture approach in which private insurance status was assumed to be exogenous. This is usually done for computational simplicity and because a suitable instrument is difficult to find given limitations in the available data. The six measures of medical care did not include prescription drugs. However, they found strong evidence of two components different in utilization means interpreted as relatively healthy and unhealthy groups although covariates were not included in the component probabilities, which limits understanding of the factors placing individuals in the components.

4.1. Data and Instrument

The data set is derived from eight annual surveys of the Medical Expenditure Panel Survey (MEPS) conducted between 1996–2003. MEPS is a nationally representative survey of health care use, expenditure, sources of payment and insurance coverage for the US civilian non-institutionalized population, and it is publicly available at the Agency for Healthcare Research and Quality (AHRQ). The MEPS data are designed as a two-year overlapping panel, i.e., each calendar year starting from 1997 includes two samples of observations, with one sample in its second year while the other sample is in its first year of responses. Thus, each individual is observed for two years at most. The constructed data set for this application includes only observations on the first round of the survey respondents in each year. The sample is restricted to only those individuals, aged 65 years and older, whose insurance status did not change during the survey period and who are covered by Medicare, which is about 99 percent in the MEPS. An individual is eligible for Medicare if he or his spouse has paid Medicare taxes for at least 40 quarters.

Medicare is the primary public health insurance for most of the elderly in the US. In the studied period the Medicare program consists of two plans. Part A covers inpatient hospital, skilled nursing and some home health care with no premium to pay. However, for the first 60 days of hospital stay

there is a deductible ($840 in 2003). Extra days of hospital care have copayment amounts which rise with the length of stay ($210 per day for days 61-90, and $420 per day for hospital stays beyond the 90th day in 2003). Any stay of more than 150 days is not covered. Medicare Part B covers physician services, outpatient hospital services, certain home health services and durable medical equipment. Part B coverage has a monthly premium ($58.70 in 2003). As a result elderly could be responsible for substantial out of pocket expenses and some would choose to purchase supplemental insurance.

Medicaid, a public insurance program for the low income, can also serve as a supplemental plan, that pays for non-covered by Medicare benefits, but it would likely provide different incentives to individuals than those of private supplemental plans. In addition, the Medicaid population is likely to be different from the privately insured in both health status and risk-aversion. To reduce heterogeneity in treatment effects publicly insured individuals under Medicaid (2092 observations) are excluded from the sample.

Heterogeneity in the data can be reduced further by controlling on observable characteristics related to the type of Medigap coverage. Private insurance plans vary greatly from each other in coverage and benefits. The MEPS distinguishes among private plans received through employment, individually purchased non-group plans, group plans with unknown sources and plans related to self-employment. The instrumental variable for the treatment equation is constructed based on the supplemental security income (SSI) received by individuals, which is likely not to affect the probability of treatment much for those who receive supplemental insurance as an employment benefit. Therefore, the sample is restricted to only unemployed. That excludes 2052 employed individuals, whose Medigap coverage is 67.5 percent and average number of prescription drugs is only 17.48, indicating a healthier group.

In the studied sample Medicare did not provide coverage for prescription drugs. Medicare Part D drug coverage was enacted in 2003 and went into effect only in 2006. However, its anticipation could have affected Medigap choices and utilization, possibly as early as 2004, and therefore, only 1996–2003 surveys of the MEPS are included in the sample. Prior to 2006, drugs administered during a hospital admission and at a doctor's office were paid under Medicare Part A and Part B respectively. Medicare did not cover outpatient prescription drugs, which includes those purchased at retail, mail order, home infusion and long-term care pharmacies. Even though Medicare paid for a substantial amount of prescription drugs prior to 2006, availability of Medigap created incentives to utilize medical services above optimal levels. It is possible that supplemental insurance status is influenced by latent factors which also affect utilization. Omission of relevant unobservable variables, such as risk preferences and health status, could potentially generate endogeneity of the treatment variable. In the constructed data set Medigap patients (5650 observations or 57.5 percent) and those without private coverage (4168 observations) have on average 24.05 and 23.57 prescription drugs respectively. Even though the difference is only about 0.5 drugs or 2 percent, the average drug expenditure for Medigap patients was 6 percent higher ($1237 versus $1169). This is consistent with the fact that even through the coverage of Medicare Part A and Part B was extensive, it did not cover a small number of most expensive drugs.

The instrumental variable, SSIRATIO, is defined as the ratio of the annual supplemental security income received by individuals to their annual personal income. SSI consists of Federal benefits and States supplemental payments, which vary at the state level with some states not providing them at all. The amounts for such additional payments are set by the states, which also determine Medigap premiums. The assumption is that premium affordability influences Medigap coverage. The premiums and SSI payments are not independent, because they are related through the states'

structures. Therefore, SSI will drive Medigap premiums and will affect the probability of treatment, but it is not clear in what functional form, since these two are related through a third variable. Thus, the instrument enters the treatment equation nonparametrically but as discussed in Section 2 it must satisfy the additional "smoothness" conditions. Dividing SSI with the annual personal income restricts the values of the instrument SSIRATIO to $[0, 1]$ interval.

To argue why the constructed variable is a valid instrument consider the following. The average prescription drug expenditure in the sample is $1208 and the average income is $17,912. A large portion of this expenditure is covered either through Medigap or Medicare Part A and Part B. Individuals can be still responsible for either copays or outpatient prescription drugs not covered by Medicare. Once the level of income is controlled for, which enters the outcome equations, I argue that the structure of income, specifically its share paid by SSI, taking values in $[0, 1]$, should not affect the demand for prescription drugs, other than through Medigap coverage. The same instrument was also previously used by Munkin and Trivedi (2008) in the context of physician visits for individuals aged between 55 and 75 years. I **refine** it further by restricting the sample to only unemployed. This eliminates Medigap coverage through employment, when the effect of SSIRATIO on the probability of coverage is only marginal. Thus, the individuals in the sample have only Medigap choices with premiums set by the states.

Definitions and summary statistics of the variables used in the application are given in Table 1. The number of prescription drugs (DRUGS) is the dependent variable and Medigap private insurance (PRIVATE) is the endogenous treatment. The vector of covariates $\mathbf{X}$ in the utilization equations consist of the self-perceived health status variables VEGOOD, GOOD, FAIR and POOR (excellent health status is the excluded category), indicator of chronic diseases CHRONIC, indicator of physical limitation PHYSLIM, geographical location variables NOREAST, MIDWEST,

SOUTH and MSA, variables that proxy for socioeconomic status, BLACK, FAMSIZE, FEMALE, MARRIED, EDUC, INCOME, AGE, and year dummies YEAR97, YEAR98, YEAR99, YEAR00, YEAR01, YEAR02, YEAR03 (year dummy for 1996 is excluded). Vector **W** of the insurance equation includes all variables in **X**, excluding the intercept, and instrumental variable SSIRATIO. The final sample size is 9818.

Figure 1 shows histograms of DRUGS for three samples: all observations, Medigap and no Medigap only individuals. The distributions have a long tail, with about 12.5 percent of all observations exceeding 50, 2 percent exceeding 100 and the maximum of 290 drugs. Since it is difficult to visualize differences in frequencies for larger outcomes, for exposition purposes only the last frequency in Figure 1 is defined as cumulative 50 prescription drugs or more. Figure 2 presents differences in the frequencies across treatment groups placing them against each other. The untreated sample (no Medigap) has a larger proportion of zeroes, slightly smaller mean (23.57 versus 24.05) and larger variance (standard deviation of 26.21 versus 25.42), however, overall the distributions appear to be very similar.

4.2. Results

First, the numbers of components in the FMR model, potentially different in the treated and untreated states, are identified. Table 2 presents the marginal likelihood values, calculated for five model specifications $M(k_1, k_2)$ where $k_1, k_2 = 1, 2, 3$. Because model $M(2, 2)$ produces the best fit there is no need to calculate the marginal likelihood values for specifications $M(1, 1)$ and $M(3, 3)$. The corresponding posterior means of the parameters are calculated based on Markov chains run for 20,000 replications after discarding first 1000 replications of the burn-in phase. The posterior ordinates are estimated based on 10,000 draws following 1000 burn-in phase draws. The chains show very good mixing properties with the autocorrelation function of the parameters dying off after at

most 2-3 lags. In addition to the marginal likelihood, Table 2 presents estimated log-likelihood values. The results provide very strong evidence in favor of two components both in the treated and untreated states.

Next, the FMR model with two components in the treated and untreated states is estimated, imposing inequality constraints, based on the calculated means of the components, assigning each draw to either lower or larger mean. These constraints separate the components very well. In the treated state the conditional means of the estimated components are 10.3 and 30.6, and the estimated probabilities are 0.31 and 0.69 respectively. In the untreated state the conditional means are 13.6 and 34.8, and the component probabilities are 0.47 and 0.53. The Markov chain displays considerable serial correlations so it is run for 100,000 replications after discarding first 10,000 replications. Posterior means and standard deviations of the parameters are presented in Tables 3 and 4. The SMR modeling approach estimates how the included explanatory variables affect the component probabilities. Parameters γ_{12} and γ_{22} in Table 3 must be interpreted as increasing the probabilities of belonging to the higher mean components. Since it is likely that these probabilities both in the treated and untreated states are influenced by health status, additionally constructed variables, measuring the extent of chronic conditions, are included in the place of CHRONIC, defined as an indicator of at least one chronic condition. Specifically, variables CHRONIC1, CHRONIC2, CHRONIC3 and CHRONIC4+ are constructed as indicators of 1, 2, 3 and 4 or more chronic conditions respectively. The excluded category is no chronic conditions.

From the results in Table 3 it can be seen that all chronic condition indicators have strong impacts: the larger the numbers of chronic conditions the greater positive impacts they have on the probability of belonging to the higher utilization groups. A notable exception is CHRONIC1 which has a strong negative impact. Thus, the lower utilization components are likely to have individuals

with either no chronic conditions or just one, while the higher utilization group has two chronic conditions or more. Variables AGE, FAMSZE, EDUCYR, BLACK and INCOME have a strong impact in the untreated state only. Surprisingly variables PHYLIM, MARRY, MSA and HISP do not strongly affect the probabilities of belonging to components. Among health status variables only FAIR has a positive strong effect for both treated and untreated states. Variables VEGOOD and GOOD have a strong positive effect in the treated state only. Overall, it appears reasonable to interpret the components as relatively healthy and unhealthy groups.

Variables MARRY, FAIR, POOR and CHRONIC have a strong impact on utilization for both components in the treated state. It is interesting to notice that the impacts are much stronger in magnitude for the lower utilization group (component 1). CHRONIC has a strong positive impact on utilization for all components but PHYLIM has a strong effect only on the higher utilization components both in the treated and untreated states. Variables EDUCYR and MSA do not strongly affect the components. BLACK and HISP have strong negative impacts only in the untreated state. Variable AGE and MARRY affect the treated and untreated states differently. Health status variables produce mixed results. VEGOOD and GOOD produce strong effects only for the lower utilization component in the untreated state, but POOR has a positive strong effect on each component. INCOME is more important for the untreated group as expected, and surprisingly, it has a negative impact on the higher utilization group in the treated state.

As an alternative, another specification of the FMR model is estimated in which the instrumental variable SSIRATIO enters the treatment equation linearly

$$D_i = \alpha_0 + s_i\rho + \mathbf{W}_i\boldsymbol{\alpha} + u_i.$$

A formal test comparing the nonparametric and linear specifications of the FMR model is conducted based on the marginal likelihood values. Other testing approaches are possible, as discussed

in Tobias and Chan (2019). Table 2 presents the estimated logarithm of the marginal likelihood for the linear specification, -46747.25 (0.85). It is substantially lower than that of the nonparametric model, -46731.41 (0.91). Thus, the nonparametric model produces a better fit. Figure 3 plots the estimated function $f(s_i)$ against the linear regression $\alpha_0 + s_i\rho$. The figure also presents the 95% posterior probability intervals indicated by the dashed lines. Variable SSIRATIO is rounded up to 0.002 which gives $k_\nu = 495$ distinct values. The linear regression predicts a negative relationship between SSIRATIO and the probability of treatment over the entire support of the instrumental variable. However, according to the nonparametric model, the probability of treatment is monotonically increasing from 0 to 0.39, reaching the maximum at 0.39, after which it starts declining first slowly reaching 0.6, after which it drops sharply in a linear way, decreasing at a much faster rate than the slope of the linear regression line.

The estimated posterior means and standard deviations of the covariance parameters, δ_{11}, δ_{12}, δ_{21} and δ_{22}, are given in Table 4 as -1.557 (0.11), -0.306 (0.232), 1.813 (0.089) and 0.658 (0.144) respectively. They are well separated from zero by more than two standard deviations for all parameters except δ_{12}. Therefore, I formally test the null hypothesis that jointly restricts all the covariance parameters to zero, $H_0 : \delta_{11} = 0$, $\delta_{12} = 0$, $\delta_{21} = 0$, $\delta_{22} = 0$, against the alternative that leaves them unconstrained. The Bayes factor can be calculated using the Savage-Dickey density ratio approach (Verdinelli and Wasserman, 1995) as

$$B_0 = \frac{p(\delta_{11}^*, \delta_{12}^*, \delta_{21}^*, \delta_{22}^* | y)}{p(\delta_{11}^*, \delta_{12}^*, \delta_{21}^*, \delta_{22}^*)}, \tag{4.1}$$

where $p(\delta_{11}^*, \delta_{12}^*, \delta_{21}^*, \delta_{22}^* | y)$ is the posterior density and $p(\delta_{11}^*, \delta_{12}^*, \delta_{21}^*, \delta_{22}^*)$ is the prior density of parameters δ_{11}, δ_{12}, δ_{21} and δ_{22} evaluated at the point $\delta_{11}^* = 0$, $\delta_{12}^* = 0$, $\delta_{21}^* = 0$, $\delta_{22}^* = 0$. The choice of the priors is the same as used for model selection in Section 3. The null hypothesis of no endogeneity is overwhelmingly rejected.

4.3. Average Treatment Effects

Next the average treatment effect (ATE) and the average treatment effect for the treated (ATET) parameters are calculated for the nonparametric specification of the FMR model with two components both in the treated and untreated states. Definition of dependent variable Y_i establishes the link between the observed and counterfactual outcomes as

$$Y_i = d_i \sum_{j=1}^{2} I_{\left\{z_{ij}^1=1\right\}} Y_{ij}^1 + (1-d_i) \sum_{j=1}^{2} I_{\left\{z_{ij}^2=1\right\}} Y_{ij}^2.$$

ATE is the expected outcome gain from receipt of treatment for a randomly chosen individual and ATET is the expected outcome gain for those who actually receive the treatment. The expected means

$$E\left[Y_{ij}^t | \boldsymbol{\Delta}_i, D_i, d_i, \boldsymbol{\mu}_i^1, \boldsymbol{\mu}_i^2, \mathbf{R}_i^1, \mathbf{R}_i^2, \mathbf{z}_i^1, \mathbf{z}_i^2\right],$$

$$E\left[Y_{ij}^t | d=1, \boldsymbol{\Delta}_i, D_i, d_i, \boldsymbol{\mu}_i^1, \boldsymbol{\mu}_i^2, \mathbf{R}_i^1, \mathbf{R}_i^2, \mathbf{z}_i^1, \mathbf{z}_i^2\right],$$

$$(t=1,2;\ j=1,2),$$

conditionally on the latent variables, parameters of the model and $\boldsymbol{\Delta}_i$ (defined in Section 2), have closed forms. Thus, the ATE and ATET parameters, $E\left[Y^1 - Y^2 | \mathbf{X}\right]$ and $E\left[Y^1 - Y^2 | \mathbf{X}, \mathbf{W}, d=1\right]$, can be calculated averaging over the estimated components, observations and posterior distributions of the parameters and latent variables. The ATE and ATET parameters in this case can be interpreted as posterior means of the corresponding posterior distributions of the treatment effects. The estimated ATE is 0.479 (0.203) and ATET is 2.677 (1.958). The size of ATET relative to ATE determines whether adverse or favorable selection is present. Since the estimated posterior standard deviation of ATET is large (1.958), the conclusion on the selection effect is ambiguous.

Alternatively, instead of estimating the entire distribution, the treatment effects can be evaluated at the posterior means of the parameters, in which case ATE is 0.450 (0.160) and ATET is 1.075

(0.216). Thus, Medigap coverage provides incentives to increase annual utilization of prescription drugs by about 2 percent for a randomly selected individual, but it is an increase of 4.5 percent for those who actually select to purchase Medigap. The estimated ATET is consistent with adverse selection since the average treatment effect for those who actually select the treatment is about 0.625 drugs larger than the ATE for a randomly chosen individual.

Even though the nonparametric specification of the FMR model dominates in terms of the marginal likelihood, average treatment effects are estimated for three alternative specifications for comparison. All of them produce similar ATE values, however, they differ substantially in ATET. For the linear specification of the FMR model the estimated ATET is 4.871 (1.740). The CR model (Deb et al. 2006) produces even more extreme effects. The nonparametric and linear specifications of the CR model estimate ATET as 12.636 (0.890) and 12.808 (0.886) respectively, which are consistent with incentives effects of more than 50 percent of actual utilization much larger than the observational difference in utilization between treated and untreated groups of only 0.5 drug. Thus, these results underscore the importance of the flexible modelling approach of the FMR model.

5. Conclusion

This paper develops the Count Roy Model with Finite Mixtures to model heterogeneity in count variables. The Roy approach captures heterogeneity generated by an observed binary treatment variable. Finite mixtures further control for unobserved heterogeneity. The assumption is that finite mixtures identify components, in which the marginal and treatment effects are homogeneous. To estimate the numbers of components in the model, the marginal likelihood values are calculated for different model specifications with the help of the random permutation sampler. The model has additional features dictated by the specifics of the application. To better understand the composition of the components, their probabilities are modeled as functions of covariates using the

smoothly mixing regression approach. A continuous instrumental variable is allowed to enter the treatment equation nonparametrically.

The model is applied to identifying the incentives effects of Medigap supplemental private insurance on the demand for prescription drugs for the US elderly unemployed Medicare population. A strong evidence is found that there are two components both in the treated and untreated states interpreted as healthy and unhealthy groups. The results show that Medigap insurance provides incentives to increase prescription drug utilization by 2 percent. The estimated treatment effects are consistent with adverse selection.

References

Bago d'Uva, T. (2006). Latent class models for utilization of health care. *Health Economics*, 15, 329-343.

Berkhof, J., Van Mechelen, I., & Gelman, A. (2003). A Bayesian approach to the selection and testing of mixture models. *Statistica Sinica,* 13, 423-442.

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

Burda, M., Harding, M., & Hausman, J. (2012). A Poisson mixture model of discrete choice. *Journal of Econometrics 166*, 184–203.

Celeux, G., Hurn, M., & Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of American Statistical Association,* 95, 957-970.

Celeux, G., Fruhwirth-Schnatter, S., & Robert, C. P. (2019a). Model selection for mixture models - perspectives and strategies. In Fruhwirth-Schnatter, S., Celeux, G., & Robert, C. P. (Eds.), *Handbook of Mixture Analysis*, Chapter 7, 117-154. Boca Raton, FL: CRC Press.

Celeux, G., Kamary, K., Malsiner-Walli, G., Marin, J.-M., & Robert, C. P. (2019b). Computational solutions for bayesian inference in mixture models. In Fruhwirth-Schnatter, S., Celeux, G., & Robert, C. P. (Eds.), *Handbook of Mixture Analysis*, Chapter 5, 73-96. Boca Raton, FL: CRC Press.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association,* 90, 1313-1321.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics,* 75, 79-97.

Chib, S. (2007). Analysis of treatment response data without the joint distribution of potential outcomes. *Journal of Econometrics,* 140, 401–412.

Deb, P., & Trivedi, P. K. (1997). Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics,* 12, 313-336.

Deb, P., & Trivedi, P. K. (2013). Finite mixture for panels with fixed effects. *Journal of Econometric Methods*, 2(1), 35-54.

Deb, P., Munkin, M. K., & Trivedi, P. K. (2006). Private insurance, selection, and the health care use: a Bayesian analysis of a Roy-type model. *Journal of Business and Economic Statistics,* 24, 403-415.

Diebolt, J., & Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society Series B,* 56, 363-375.

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

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

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models.* Springer Series in Statistics.

Frühwirth-Schnatter, S., Pamminger, C., Weber, A., & Winter-Ebmer, R. (2012). Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts Markov chain clustering. *Journal of Applied Econometrics,* 27, 1116– 1137.

Frühwirth-Schnatter, S., Pamminger, C., Weber, A., & Winter-Ebmer, R. (2016). Mothers' long-run career patterns after first birth. *Journal of the Royal Statistical Society Series A*, 179, 707–725.

Fruhwirth-Schnatter, S. (2019). Keeping the balance - Bridge sampling for marginal likelihood estimation in finite mixture, mixture of experts and Markov mixture models. *Brazilian Journal of Probability and Statistics,* 33, 706-733.

Fruhwirth-Schnatter, S., Celeux, G., & Robert, C. P. (2019). *Handbook of Mixture Analysis.* Boca Raton, FL: CRC Press.

Geweke, J. (2007). Interpretation and inference in mixture models: simple MCMC works. *Computational Statistics & Data Analysis,* 51, 3529-3550.

Geweke, J., & Keane, M. (2007). Smoothly mixing regressions. *Journal of Econometrics,* 138, 252-291.

Gormley, I. C., & Fruhwirth-Schnatter, S. (2019). Mixture of experts models. In Fruhwirth-Schnatter, S., Celeux, G., & Robert, C. P. (Eds.), *Handbook of Mixture Analysis*, Chapter 12, 271-307. Boca Raton, FL: CRC Press.

Heckman, J. J., Lopes, H., & Piatek, R. (2014). Treatment effects: a Bayesian perspective. *Econometric Reviews,* 33, 36–67.

Hu, X., Munkin, M. K., & Trivedi, P. K. (2015). Estimating incentive and selection effects in the Medigap insurance market: an application with Dirichlet process mixture model. *Journal of Applied Econometrics,* 30, 1115–1143.

Hyppolite, J., & Trivedi, P. K. (2012). Alternative approaches for econometric analysis of panel count data using dynamic latent class models (with application to doctor visits data). *Health Economics*, 21, 101-128.

Hurn, M., Justel, A., & Robert, C. P. (2003). Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics,* 12, 55-79.

Jacobi, L., Wagner, H., & Frühwirth-Schnatter, S. (2016). Bayesian treatment effects models with variable selection for panel outcomes with an application to earnings effects of maternity leave. *Journal of Econometrics*, 193, 234-250.

Jasra, A., Holmes, C. C., & Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science,* 20, 50–67.

Karlis, D., Papatla, P., & Roy, S. (2016). Finite mixtures of censored Poisson regression models. *Statistica Neerlandica*, 70, 100–122.

Koop, G., & Poirier, D. J. (2004). Bayesian variants of some classical semiparametric regression techniques. *Journal of Econometrics*, 123, 259-282.

Koop, G., & Tobias, J. L. (2006). Semiparametric Bayesian inference in smooth coefficient models. *Journal of Econometrics*, 134, 283-315.

Lee, J. E. & Robert, C. P. (2016). Importance sampling schemes for evidence approximation in mixture models. *Bayesian Analysis,* 11, 573-597.

Li, M., Poirier, D. J., & Tobias, J. L. (2004). Do dropouts suffer from dropping out? Estimation and prediction of outcome gains in generalized selection models. *Journal of Applied Econometrics*, 19, 203-225.

Müller, P. & Mitra, R. (2013). Bayesian nonparametric inference – why and how. *Bayesian Analysis,* 8, 269–360.

Munkin, M. K., & Trivedi, P. K. (2008). Bayesian analysis of the ordered probit model with endogenous selection. *Journal of Econometrics,* 143, 334-348.

Poirier, D. J., & and Tobias, J. L. (2003). On the predictive distributions of outcome gains in the presence of an unidentified parameter. *Journal of Business and Economic Statistics,* 21, 258-268.

Shiller, R. J. (1984). Smoothness priors and nonlinear regression. *Journal of the American Statistical Association*, 79, 609-615.

Tobias, J. & Chan, J. (2019). An alternate parameterization for Bayesian nonparametric/semiparametric regression. *Advances in Econometrics*, 40, 47-64.

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of American Statistical Association,* 90, 614-618.

Viallefont, V., Richardson, S., & Green, P. J. (2002). Bayesian analysis of Poisson mixtures. *Journal of Nonparametric Statistics,* 14, 181-202.

Villani, M., Kohn, R., & Nott, D. (2012). Generalized smooth finite mixtures. *Journal of Econometrics*, 171, 121-133.

Wedel, M., Desarbo, W. S., Bult, J. R., & Ramaswamy, V. (1993). A latent class Poisson regression model for heterogeneous count data with an application to direct mail. *Journal of Applied Econometrics*, 8, 397-411.

Table 1. Summary of the Data (9818 obs.)

		mean	st. dev.
Utilization			
DRUGS	number of prescription drugs	23.846	25.760
Insurance			
PRIVATE	= 1 if private supplemental insurance	0.575	0.494
Demographic variables			
FAMSIZE	family size	1.914	0.974
AGE	age/10	7.572	0.666
EDUC	years of schooling	11.534	3.332
INCOME	= income/1000	17.912	17.149
FEMALE	= 1 if female	0.597	0.491
BLACK	= 1 if black	0.101	0.301
MARRIED	= 1 if married	0.549	0.498
NOREAST	= 1 if northeast	0.185	0.388
MIDWEST	= 1 if Midwest	0.240	0.427
SOUTH	= 1 if south	0.369	0.483
MSA	=1 if metropolitan stat. area	0.734	0.442
Health variables			
VEGOOD	= 1 if very good health	0.257	0.437
GOOD	= 1 if good health	0.332	0.471
FAIR	= 1 if fair health	0.191	0.393
POOR	= 1 if poor health	0.071	0.258
PHYSLIM	= 1 if physical limitation	0.359	0.480
CHRONIC	= 1 if at least 1 chronic condition	0.824	0.381
CHRONIC1	= 1 if 1 chronic condition	0.291	0.454
CHRONIC2	= 1 if 2 chronic conditions	0.264	0.441
CHRONIC3	= 1 if 3 chronic conditions	0.162	0368
CHRONIC4+	= 1 if 4 or more chronic conditions	0.107	0.309
Year dummies			
YEAR97	=1 if year 1997	0.058	0.234
YEAR98	=1 if year 1998	0.104	0.305
YEAR99	=1 if year 1999	0.092	0.289
YEAR00	=1 if year 2000	0.118	0.323
YEAR01	=1 if year 2001	0.097	0.296
YEAR02	=1 if year 2002	0.186	0.389
YEAR03	=1 if year 2003	0.253	0.435
Exclusion Restriction			
SSIRATIO	= supplemental security income/income	0.571	0.356

Table 2. Marginal Likelihood Values

	Nonparametric IV					Linear IV
	$M(1,2)$	$M(2,1)$	$M(2,2)$	$M(2,3)$	$M(3,2)$	$M(2,2)$
$\log l(y\|\theta^*)$	-46589.00	-46909.00	-46278.00	-46283.00	-46283.00	-46310.00
$\log m(y)$	-46940.29	-47261.55	-46731.41	-46827.84	-46838.90	-46747.25
	(1.28)	(1.47)	(0.91)	(1.13)	(1.54)	(0.85)

Table 3. Posterior Means and Standard Deviations of Component Parameters $\boldsymbol{\gamma}_{12}$ (treated), $\boldsymbol{\gamma}_{22}$ (untreated) and Treatment Equation Parameter $\boldsymbol{\alpha}$

	Treated State Vector γ_{12}		Untreated State Vector γ_{22}		Treatment Equation Vector α	
	mean	std.dev.	mean	std.dev.	mean	std.dev.
CONSTANT	-0.703	0.761	-2.372	0.692		
YEAR97	-0.209	0.249	-0.125	0.269	-0.02	0.053
YEAR98	-0.382	0.193	-0.025	0.223	-0.164	0.044
YEAR99	-0.121	0.183	0.437	0.229	-0.132	0.045
YEAR00	-0.028	0.173	0.018	0.204	-0.196	0.04
YEAR01	-0.188	0.199	-0.137	0.217	-0.215	0.045
YEAR02	0.19	0.174	-0.026	0.193	-0.251	0.038
YEAR03	-0.077	0.172	0.172	0.182	-0.308	0.035
AGE	-0.003	0.087	0.185	0.075	-0.055	0.016
FAMSZE	-0.081	0.08	-0.143	0.056	-0.135	0.012
EDUCYR	0.03	0.02	0.074	0.015	0.042	0.003
FEMALE	0.11	0.103	0.157	0.103	-0.017	0.022
MARRY	0.244	0.15	0.183	0.114	0.362	0.025
NORTHE	-0.051	0.165	-0.131	0.16	0.12	0.033
MWEST	-0.109	0.16	-0.02	0.153	0.262	0.031
SOUTH	0.093	0.156	0.15	0.137	0.177	0.029
PHYLIM	0.207	0.139	0.136	0.141	0.059	0.024
VEGOOD	0.279	0.127	0.213	0.137	0.084	0.033
GOOD	0.446	0.138	0.143	0.132	0.005	0.032
FAIR	0.344	0.191	0.588	0.184	-0.052	0.036
POOR	0.158	0.255	0.123	0.277	-0.087	0.048
MSA	0.181	0.111	-0.178	0.122	0.006	0.024
INCOME	0.005	0.004	0.007	0.003	0.004	0.001
BLACK	-0.152	0.192	-0.413	0.157	-0.287	0.035
HISP	-0.216	0.334	0.161	0.166	-0.515	0.042
CHRONIC					0.086	0.028
CHRONIC1	-0.412	0.155	-1.191	0.25		
CHRONIC2	0.614	0.148	0.322	0.164		
CHRONIC3	1.748	0.249	1.014	0.179		
CHRONIC4+	2.721	0.802	2.87	1.162		
PROB $j = 2$	0.69	0.02	0.532	0.028		
η					0.00072	0.00053

Table 4. Posterior Means and Standard Deviations of Parameters $\boldsymbol{\phi}_{1j}$, $\boldsymbol{\phi}_{2j}$, σ^2_{1j}, σ^2_{2j} by State ($d = 0, 1$) and Components ($j = 1, 2$)

	Treated State				Untreated State			
	Component 1		Component 2		Component 1		Component 2	
	mean	std.dev.	mean	std.dev.	mean	std.dev.	mean	std.dev.
CONSTANT	-0.307	0.997	2.412	0.3	0.243	0.638	3.21	0.316
YEAR97	-0.132	0.331	-0.008	0.075	-0.047	0.227	0.056	0.116
YEAR98	0.1	0.213	0.101	0.071	-0.13	0.173	-0.026	0.099
YEAR99	0.225	0.219	0.096	0.067	-0.347	0.2	-0.005	0.094
YEAR00	0.093	0.207	-0.033	0.066	-0.192	0.171	-0.164	0.09
YEAR01	0.295	0.252	0.293	0.073	-0.044	0.174	0.243	0.102
YEAR02	0.025	0.241	0.147	0.066	-0.171	0.16	0.078	0.086
YEAR03	0.796	0.199	0.295	0.07	-0.146	0.151	0.108	0.084
AGE	0.057	0.108	0.001	0.026	-0.194	0.065	-0.071	0.033
FAMSZE	0.168	0.106	0.049	0.034	-0.137	0.039	-0.027	0.026
EDUCYR	-0.042	0.025	-0.016	0.008	0.015	0.014	-0.001	0.008
FEMALE	0.139	0.128	0.067	0.031	0.176	0.088	-0.032	0.044
MARRY	-0.494	0.185	-0.166	0.073	0.541	0.096	0.088	0.061
NORTHE	0.13	0.216	0.143	0.051	0.462	0.13	0.18	0.066
MWEST	0.114	0.208	0.26	0.058	0.68	0.13	0.469	0.065
SOUTH	0.214	0.204	0.196	0.048	0.466	0.121	0.335	0.056
PHYLIM	-0.017	0.184	0.241	0.035	0.088	0.11	0.276	0.047
VEGOOD	0.03	0.163	-0.036	0.054	0.32	0.126	-0.088	0.072
GOOD	0.112	0.176	0.055	0.051	0.27	0.12	0.022	0.066
FAIR	0.482	0.226	0.322	0.058	0.331	0.164	0.056	0.071
POOR	1.028	0.259	0.605	0.069	0.533	0.212	0.239	0.092
MSA	-0.238	0.129	-0.052	0.033	-0.013	0.102	0.01	0.046
INCOME	-0.008	0.005	-0.003	0.001	0.012	0.003	0.003	0.001
BLACK	-0.065	0.276	0.071	0.075	-0.415	0.121	-0.183	0.077
HISP	0.374	0.489	0.101	0.126	-0.904	0.153	-0.313	0.083
CHRONIC	2.682	0.292	0.726	0.067	3.961	0.217	0.824	0.062
$\delta_{1j}; \delta_{2j}$	-1.557	0.11	-0.306	0.232	1.813	0.089	0.658	0.144
$\sigma^2_{1j}; \sigma^2_{2j}$	0.354	0.082	0.416	0.046	0.348	0.066	0.371	0.043
$\exp(\mu^1_j), \exp(\mu^2_j)$	10.278	0.984	30.625	0.455	13.579	0.877	34.769	0.897

Figure 1. Histograms of DRUGS for All Observations, Medigap and No Medigap Samples.

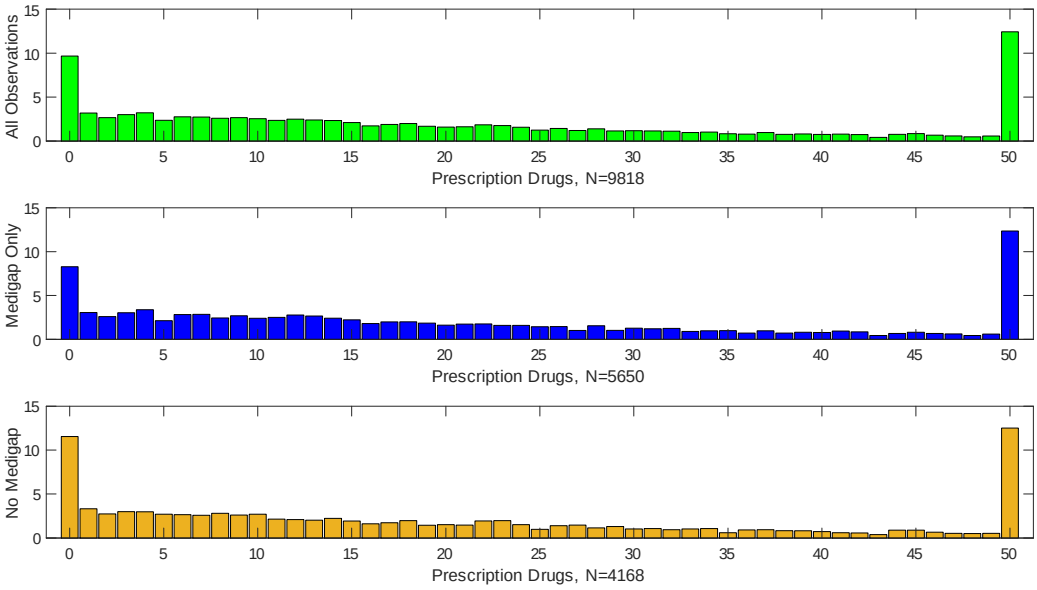

Figure 2. Histograms of DRUGS for Medigap versus No Medigap Samples.

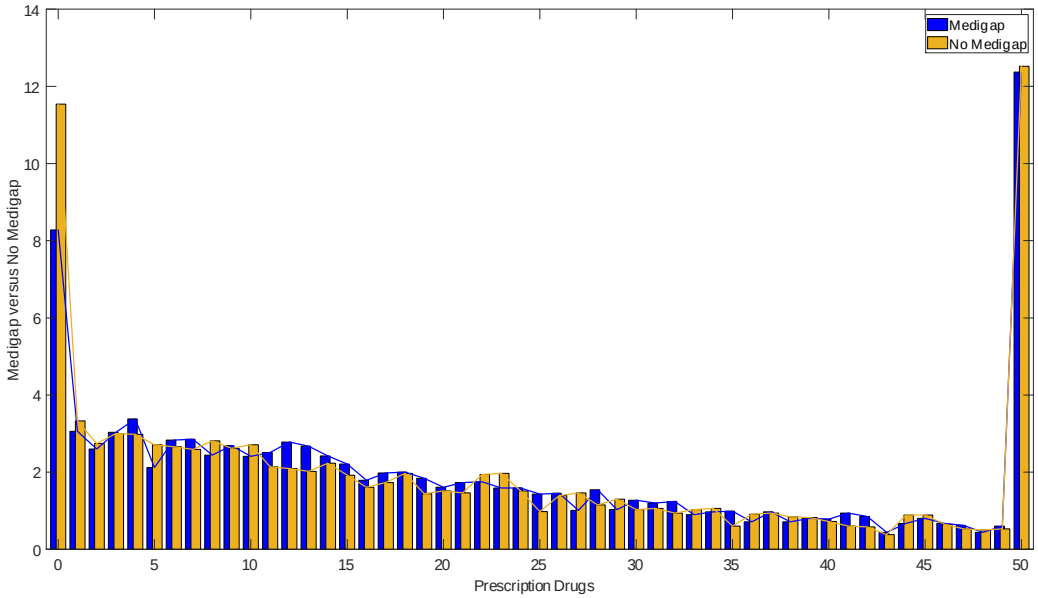

Figure 3. The Effect of SSIRATIO on Medigap: Nonparametric vs. Linear.