econometrics

MDPI

Article

An Exponential Endogenous Switching Regression with Correlated Random Coefficients

Myoung-Jin Keay

Ness School of Management and Economics, South Dakota State University, Brookings, SD 57006, USA; MyoungJin.Keay@sdstate.edu

Abstract: This paper presents a method for estimating the average treatment effects (ATE) of an exponential endogenous switching model where the coefficients of covariates in the structural equation are random and correlated with the binary treatment variable. The estimating equations are derived under some mild identifying assumptions. We find that the ATE is identified, although each coefficient in the structural model may not be. Tests assessing the endogeneity of treatment and for model selection are provided. Monte Carlo simulations show that, in large samples, the proposed estimator has a smaller bias and a larger variance than the methods that do not take the random coefficients into account. This is applied to health insurance data of Oregon.

Keywords: Correlated Random Coefficient; average treatment effect; exponential model; endogenous switching regression

JEL Classification: C15; C25; C34; C35; J13

check for updates

Citation: Keay, Myoung-Jin. 2022. An Exponential Endogenous Switching Regression with Correlated Random Coefficients. *Econometrics* 10: 1. https://doi.org/10.3390/econometrics10010001

Academic Editor: Ryo Okui

Received: 12 October 2021
Accepted: 10 December 2021
Published: 21 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The conventional instrumental variable methods fail to consistently estimate the Average Partial Effects (APE) when the individual heterogeneity enters the model in a non-additive way (Heckman and Vytlacil 1998; Card 2001; Browning and Lechene 2003; Browning and Carro 2007; Imbens 2007). One of the simplest models, which allows for non-additive heterogeneity, is the Correlated Random Coefficient (CRC) model, where the heterogeneity interacts with the covariates to create random coefficients. This is in fact a generalization of the additive heterogeneity models, since the sum of the error and the constant in a linear regression equation is a random intercept. This paper provides the identification and the estimation of the average treatment effect (ATE) for an exponential model that has two regimes induced by an endogenous treatment. The covariates in the model have random coefficients that are correlated with the treatment. The conditional expectation is expressed by an exponential function due to the limited nature of the dependent variable. We also provide a test to assess the endogeneity of the treatment and for model selection. One important feature of this model is that it uncovers another mechanism of the individual heterogeneity working through the heterogeneous and endogenous marginal effects of covariates. This advantage comes from the parametric modelling not attempted by others who used non-parametric approaches, such as Angrist and Imbens (1994, 1995), Abadie (2003) and Heckman and Vytlacil (2005).

The CRC models are closely related, although not exactly the same, to heterogeneous treatment effects models. There are two studies on this subject. The first is by Heckman and Vytlacil (1998) where they considered random, i.e., individual-specific, coefficients on the explanatory variables. The second is due to Angrist and Imbens (1994) under Rubin's counterfactual setting, where the heterogeneous treatment effect, i.e., the individual-specific difference of outcomes in two regimes, appears as the coefficient on the binary regime indicator. These two types of models end up being seemingly very similar in that the individual-specific heterogeneity is a non-constant coefficient. The differences are that

the former is not Rubin's causal model and that it may have random coefficients on a continuous variables. The conventional 2SLS fails to identify what the parameters of interest might be for both models. For the first model, it would be the mean of the random coefficient, and, for the second, the mean of treatment effects. Nevertheless, for the second model, Angrist and Imbens (1994, 1995) showed that the 2SLS estimates another meaningful parameter called local average treatment effects (LATE) or their weighted averages. The identification of LATE was made possible largely due to the counterfactual two-regime setting, which the first model does not allow. The model that I deal with in this article is a synthesis of these two models. It has random coefficients on the continuous covariates as well as the two counterfactual regimes under Rubin's causal model. I will show that the ATE can be identified and estimated.

When the variable of interest is continuous and its coefficient is correlated with the variable itself, the parameter of interest would be the average of the random coefficient. The error term will then include the interactions between a random part of the coefficient and the variable itself. For such cases, the conventional instrumental variable assumptions are not sufficient to consistently estimate the parameters of interests. To solve this problem, Garen (1984) suggests a control function method using some distributional assumptions. Heckman and Vytlacil (1998) propose a two-step plug-in estimator, and Wooldridge (2003) delineates the additionally required conditions for instrumental variables to ensure consistency. More recently, Masten and Torgovitsky (2016) provide the identification result when the first-stage equation also has random coefficients.

The main theme of this paper is the identification and estimation of the ATE in the exponential endogenous switching (ES) regression, where the covariates in the structural equations have CRCs. Terza (1998, 2009) considers the model without CRC. Wooldridge (2015), in a linear setting with a continuous dependent variable, allows the coefficients on all other covariates to be random and correlated with the treatment. In line with this, this study extends Terza (2009)'s exponential ES regression model by allowing the coefficients on the covariates in each regime to be correlated with the binary regime indicator. This is an exponential version of the linear two-regime CRC model. I will show how it can estimate the ATE of health insurance from the number of emergency room visits. Simulations confirm that the exponential CRC model has smaller finite sample bias when the covariates have CRC.

The remainder of this paper is organized as follows. Section 2 discusses the two-regime CRC model with a non-negative dependent variable and provides the estimating equation under some mild assumptions. Section 3 relates the estimating equation with the identification of ATE. Section 4 provides the specification tests for endogeneity and model selection. Section 5 describes the data-generating processes of Monte Carlo simulations and presents the results. Section 6 provides an empirical application that analyses the effect of Oregon's health insurance on the number of emergency room visits. Section 7 presents the concluding remarks.

2. Exponential Regime Switching CRC Model

The variable of interest w is binary, i.e., $w \in \{0, 1\}$, and this calls for an analysis of ATE under a counterfactual framework. In this setting, the first-stage binary selection equation is assumed to follow the probit model, which has been assumed in Heckman (1976) and Terza (2009). For a continuous w, the average partial effects (APE) are identified (Wooldridge 2015), although the average of the random coefficient on w is not. Similarly, I will show that, in the exponential regression model, the ATE is identified while the average of the coefficient on w is not. One can instead use the log transformation on the dependent variable. In this case, however, the dependent variable must not be zero. This can be problematic when it is a count variable, as in the example in Section 6. The exponential conditional expectation is more flexible in this sense and thus is chosen as the preferred model. Let a_w and b_w denote individual heterogeneities that may affect the outcomes y_w. Here, a_w and b_w are the random intercept and the vector of random coefficients at the

regime w. The regime subscript w will often be omitted unless necessary. Let b_j denote the random coefficient on the j-th explanatory variable so that $b = (b_1, b_2, \cdots, b_K)'$. We make the following assumptions:

Assumption 1. *(Mean Independence)*

$$E(y_w|\mathbf{z}, w, a_0, b_0, a_1, b_1) = E(y_w|\boldsymbol{x}, a_w, b_w); \quad w = 0, 1, \tag{1}$$

where the equation describes Regime 1 when $w = 1$ and regime 0 otherwise.

We assume that the conditional expectation is correctly specified. In the estimation later on, the correct specification guarantees the consistency despite an incorrect log-likelihood function (Gourieroux et al. 1984). Once the regime is determined, the regime indicator w is redundant conditional on a_w and b_w, because they summarize all the information about the determined regime. The vector $\mathbf{z}$ denotes all the available exogenous variables that include $\mathbf{x}$. The exogeneity of $\mathbf{z}$ means that it is independent of a_w and b_w. The assumption is that there must be some exogenous variables that are excluded from the conditional expectation function. Additionally, once the unobserved heterogeneity in a particular regime is accounted for, the heterogeneity for the other regime has no additional information. By using the exponential conditional expectation, the model can be written as

$$\begin{aligned} E(y_w|\mathbf{z}, w, a_0, b_0, a_1, b_1) &= E(y_w|\mathbf{x}, a_w, b_w) = \exp(a_w + \mathbf{x}b_w) \\ w &= 1[\mathbf{z}\gamma + v > 0]. \end{aligned} \tag{2}$$

The random intercepts a_w are scalars, $\mathbf{x}$ is $1 \times K$ vector of covariates, and b_w are $K \times 1$ random coefficient vectors. The heterogeneity b_w enters as random coefficients. Although we do not attempt to identify the APEs of covariates per se, their presence hampers the identification of ATE that we identified and estimated. For random coefficients, we will use notation $\sigma^2(x) \equiv \text{var}(x)$ and $\sigma(x, y) \equiv \text{cov}(x, y)$.

Assumption 2. *(Probit) The regime choice equation in (2) follows the probit model, i.e., $v \sim N(0, 1)$, and γ is reparametrized accordingly. Additionally, $\mathbf{z}$ is independent of v.*

Although Terza (1998) initially derives the correction term under the assumption of the multivariate normality between v and the error in the structural equation, Terza (2009) later shows that the normality of v is in fact sufficient for deriving the correction terms. We absorb the dependence by using linear projections and avoid using the multivariate normality. This parametric probit assumes the homogenous instrument effects on the endogenous variables. Heterogeneous instrument effects, as discussed by Masten and Torgovitsky (2016), are not attempted here. Their approach can help relax not only the homogenous instrument effects assumption, but also the distributional assumption on v at the same time. We leave it as a future task.

Let u^b, e^b and $\sigma(u^b, v)$ denote $K \times 1$ vectors, of which the elements are u^{b_j}, e^{b_j} and $\sigma(u^{b_j}, v)$. For example, $\sigma(u^b, v) = [\sigma(u^{b_1}, v), \cdots, \sigma(u^{b_K}, v)]'$. Additionally, let $E(a) = \alpha$ and $E(b) = \beta$. The random coefficients in (1) can be written as

$$a = \alpha + u^a \tag{3}$$

$$b = \beta + u^b. \tag{4}$$

For the errors in the above equations, their orthogonal projections on v are

$$u^a = \sigma(u^a, v) \cdot v + e^a \tag{5}$$

$$u^{b_j} = \sigma(u^{b_j}, v) \cdot v + e^{b_j} \tag{6}$$

for each j. In addition to the orthogonality between v and (e^a, e^b), which is implied by linear projection, we need a slightly stronger assumption, as below.

Assumption 3. *(Conditional Independence)*

(i) $v \perp\!\!\!\perp e^a, e^b$

(ii) $v \perp\!\!\!\perp e^a + \boldsymbol{x}e^b \mid \boldsymbol{z}$

We use the independence of v with e^a and e^b. This is stronger than the orthogonality guaranteed by the projection theorem. Although e^a and e^b are independent with v, their sum may not be independent with v. The second part precludes this. It says that the σ-field of v and that of $e^a + \boldsymbol{x}e^b$ are independent depending on $\boldsymbol{z}$. Here, additionally depending on $w = 1$ is equivalent to conditioning on $v > -\boldsymbol{z}\gamma$. When an arbitrary event in the σ-field of v is independent with the σ-field of $e^a + \boldsymbol{x}e^b$, its intersection with $v > -\boldsymbol{z}\gamma$ must also be independent, because $v > -\boldsymbol{z}\gamma$ is an event in the σ-field of v. Therefore, $v \perp\!\!\!\perp e^a + \boldsymbol{x}e^b \mid \boldsymbol{z}, w$ is implied by Assumption 3(ii). One can now establish the following lemma:

Lemma 1. *Under Assumptions 2 and 3,*

$$E[\exp(u_1^a + \boldsymbol{x}u_1^b)|\boldsymbol{z}, w = 1]$$
$$= \exp\left[\left(\sigma^2(u_1^a) + \sum_{j=1}^{K}\sigma^2(u_1^{b_j})x_j^2 + 2\sum_{j=1}^{K}\sigma(u_1^a, u_1^{b_j})x_j + \sum_{j=1}^{K}\sum_{r\neq j}\sigma(u_1^{b_j}, u_1^{b_r})x_j x_r\right)\Big/2\right]$$
$$\times \frac{\Phi\left(z\gamma + \sigma(u_1^a, v) + \sum_{j=1}^{K}\sigma(u_1^{b_j}, v)x_j\right)}{\Phi(z\gamma)}$$

for Regime 1, and

$$E[\exp(u_0^a + \boldsymbol{x}u_0^b)|\boldsymbol{z}, w = 0]$$
$$= \exp\left[\left(\sigma^2(u_0^a) + \sum_{j=1}^{K}\sigma^2(u_0^{b_j})x_j^2 + 2\sum_{j=1}^{K}\sigma(u_0^a, u_0^{b_j})x_j + \sum_{j=1}^{K}\sum_{r\neq j}\sigma(u_0^{b_j}, u_0^{b_r})x_j x_r\right)\Big/2\right]$$
$$\times \frac{\Phi\left(-z\gamma - \sigma(u_0^a, v) - \sum_{j=1}^{K}\sigma(u_0^{b_j}, v)x_j\right)}{\Phi(-z\gamma)}$$

for Regime 0.

Proof. See Appendix A. □

Let ϵ and v follow the bivariate normal distribution with the correlation ρ. Terza (1998, 2009) directly uses the bivariate normality in order to solve the conditional expectation, i.e., $E[\exp(\epsilon)|\mathbf{z}, v] = \exp\left(\rho\sigma v + \frac{1}{2}\sigma^2(1-\rho^2)\right)$ was derived under the bivariate normal assumption between ϵ and v. In the above derivation, I use the linear projections of u^a and u^b on v. This gives an estimating equation that can be directly used for identifying ATE. The structural parameters, such as β, are not independently identified, but the coefficients on x and x^2 from the estimating equation derived by the linear projection coincide with the coefficients of the estimating equation for ATE.

Let us now derive an estimating equation of the model in (2). Note that by Assumptions 1 and 2,

$$\begin{aligned}
E[y|\mathbf{z}, w, a_0, b_0, a_1, b_1] &= (1-w)E[y_0|\mathbf{z}, w, a_0, b_0, a_1, b_1] + wE[y_1|\mathbf{z}, w, a_0, b_0, a_1, b_1] \\
&= (1-w)E[y_0|\mathbf{x}, a_0, b_0] + wE[y_1|\mathbf{x}, a_1, b_1] \\
&= (1-w)\exp\left(\alpha_0 + u_0^a + \mathbf{x}(\beta_0 + u_0^b)\right) + w\exp\left(\alpha_1 + u_1^a + \mathbf{x}(\beta_1 + u_1^b)\right) \\
E[y|\mathbf{z}, w] &= (1-w)\exp(\alpha_0 + \mathbf{x}\beta_0)E[\exp(u_0^a + \mathbf{x}u_0^b)|\mathbf{z}, w = 0] \\
&\quad + w\exp(\alpha_1 + \mathbf{x}\beta_1)E[\exp(u_1^a + \mathbf{x}u_1^b)|\mathbf{z}, w = 1]
\end{aligned}$$

By Lemma 1,

$$
\begin{aligned}
= \; & (1-w)\exp\left[\left(2\alpha_0+\sigma^2(u_0^a)+\sum_{j=1}^{K}\sigma^2(u_0^{b_j})x_j^2+2\sum_{j=1}^{K}[\beta_0+\sigma(u_0^a,u_0^{b_j})]x_j\right.\right. \\
& \left.\left.+\sum_{j=1}^{K}\sum_{r\neq j}\sigma(u_0^{b_j},u_0^{b_r})x_jx_r\right)\Big/2\right]\times\frac{\Phi\Big(-z\gamma-\sigma(u_0^a,v)-\sum_{j=1}^{K}\sigma(u_0^{b_j},v)x_j\Big)}{\Phi(-z\gamma)} \\
& +w\cdot\exp\left[\left(2\alpha_1+\sigma^2(u_1^a)+\sum_{j=1}^{K}\sigma^2(u_1^{b_j})x_j^2+2\sum_{j=1}^{K}[\beta_1+\sigma(u_1^a,u_1^{b_j})]x_j\right.\right. \\
& \left.\left.+\sum_{j=1}^{K}\sum_{r\neq j}\sigma(u_1^{b_j},u_1^{b_r})x_jx_r\right)\Big/2\right]\times\frac{\Phi\Big(z\gamma+\sigma(u_1^a,v)+\sum_{j=1}^{K}\sigma(u_1^{b_j},v)x_j\Big)}{\Phi(z\gamma)}
\end{aligned}
\tag{7}
$$

The above equation identifies only $\sigma^2(u^{b_j})$, $\sigma(u^a,v)$ and $\sigma(u^{b_j},v)$ separately, and the average of semi-elasticity, i.e., β, is not identified due to the nuisance parameters $\sigma(u^a,u^{b_j})$. The equation can be estimated by a two-step procedure using the probit model for the first-stage selection equation. Given the estimated index $\mathbf{z}\widehat{\gamma}$, the above equation can be estimated either by non-linear least squares or by quasi-maximum likelihood under a Poisson distribution. It can still be estimated simultaneously by a single-step method. See Keay (2018) for a detailed discussion of the estimation procedure and their asymptotic distribution. Since the structural parameters α and β are not separately identified, our interest should lie in the ATEs. This is discussed in the next section.

3. Estimating the Average Treatment Effects

The ATE is defined as $E[y_1-y_0]$. One of the easiest ways to identify this is by using the law of iterated expectation, i.e., $E[y_1-y_0]=EE[y_1-y_0|\mathbf{x}]$, as long as the expectation conditional on x can be derived. To this end, the following lemma is useful:

Lemma 2. *Under Assumption 3, the following result holds:*

$$
E[\exp(u^a+\boldsymbol{x}u^b)|\boldsymbol{x}]=\exp\left[\left(\sigma^2(u^a)+\sum_{j=1}^{K}\sigma^2(u^{b_j})x_j^2+2\sum_{j=1}^{K}\sigma(u^a,u^{b_j})x_j+\sum_{j=1}^{K}\sum_{r\neq j}\sigma(u^{b_j},u^{b_r})x_jx_r\right)\Big/2\right]
$$

Proof. See Appendix A.

The ATE conditional on $\mathbf{x}$ is

$$
E[y_1-y_0|\mathbf{x}]=\exp(\alpha_1+\mathbf{x}\beta_1)E[\exp(u_1^a+\mathbf{x}u_1^b)|\mathbf{x}]-\exp(\alpha_0+\mathbf{x}\beta_0)E[\exp(u_0^a+\mathbf{x}u_0^b)|\mathbf{x}]. \tag{8}
$$

From the above lemma,

$$
\begin{aligned}
& E[y_1-y_0|\mathbf{x}] \\
& =\exp\left[\left(2\alpha_1+\sigma^2(u_1^a)+\sum_{j=1}^{K}\sigma^2(u_1^{b_j})x_j^2+2\sum_{j=1}^{K}[\beta_1+\sigma(u_1^a,u_1^{b_j})]x_j+\sum_{j=1}^{K}\sum_{r\neq j}\sigma(u_1^{b_j},u_1^{b_r})x_jx_r\right)\Big/2\right] \\
& -\exp\left[\left(2\alpha_0+\sigma^2(u_0^a)+\sum_{j=1}^{K}\sigma^2(u_0^{b_j})x_j^2+2\sum_{j=1}^{K}[\beta_0+\sigma(u_0^a,u_0^{b_j})]x_j+\sum_{j=1}^{K}\sum_{r\neq j}\sigma(u_0^{b_j},u_0^{b_r})x_jx_r\right)\Big/2\right].
\end{aligned}
\tag{9}
$$

Note the similarity of this equation with the estimating equation in (7); the arguments of the exponential functions are identical. The structural parameters β are not separately identified, but the whole coefficients are estimated by the procedures applied

to Equation (7). Let $\widehat{E[y_1 - y_0|\mathbf{x}_i]}$ be Equation (9) with the coefficients replaced by the estimates. The ATE estimator is the sample average of $\widehat{E[y_1 - y_0|\mathbf{x}_i]}$ over the values of $\mathbf{x}_i$.

The consistency is guaranteed by Assumption 1. We now establish the asymptotic normality of the ATE. To this end, let us define θ as the vector of the coefficients estimated by the above estimating equation, i.e., $\theta_w = [2\alpha_w + \sigma^2(u_w^a), \sigma^2(u_w^{b_j}), \beta_w + \sigma(u_w^a, u_w^{b_j}), \sigma(u_w^{b_j}, u_w^{b_r})]$. The asymptotic variance of θ is simply $A^{-1}BA^{-1}$, where A and B are the Hessian and information matrices. For the asymptotic distribution of the ATE, we have the following proposition. □

Proposition 1. *Let $g_1(x,\theta) - g_0(x,\theta)$ denote the equation in (9). Let G_1 and G_0 denote the expectation of the derivatives of $g_1(x,\theta)$ and $g_0(x,\theta)$ with respect to θ. Under the usual regularity conditions of a maximum likelihood estimator, the asymptotic distribution of ATE estimator is*

$$\sqrt{N}(\widehat{ATE} - ATE) \to_d N(0, V),$$

where

$$V = E[T]^2 + (G_1 - G_0)A^{-1}BA^{-1}(G_1 - G_0)'$$

and

$$T \equiv g_1(x,\theta) - g_0(x,\theta) - \Big(E[g_1(x,\theta)] - E[g_0(x,\theta)]\Big).$$

Proof. See Appendix A. □

We have seen that the above identification relies on the normality of v. Is there any possible way of relaxing the normality assumption while continuing to use the linear index binary choice setting as in (2)? The estimating equations given in Lemma 1 can be semiparametrically estimated. For $E\Big[\exp\Big((\sigma(u^a, v) + \mathbf{x}\sigma(u^b, v))v\Big)\Big|\ \mathbf{z}, w = 1\Big]$, where $w = 1$ is equivalent to $v > -\mathbf{z}\gamma$, the expectation is conditional on $\mathbf{z}$ and $\mathbf{z}\gamma$, and it can be semiparametrically estimated in terms of these variables by using two-step series or spline methods (Newey 2009). In order to identify the ATE, however, the expectation that it is conditional on $\mathbf{x}$ has to be found as in Equation (9). The fact that $w = 1$ is not in the conditioning set implies that v is integrated over the entire real number. The corresponding value of $\mathbf{z}\gamma$ thus has to be infinity for which no data are available. It becomes clear from this that the role of the normality assumption is to provide information for such an extreme area in the domain. Although Blundell and Powell (2004) provides a semiparametric control function approach, their method is not applicable when the endogenous explanatory variable, the binary regime indicator in this case, is not continuous.

4. Specification Tests

4.1. Tests for Endogeneity of Treatment

In the previous section, we derived the estimating equation Equation (7) and ATE conditional on covariates (9). No restrictions were imposed in these equations, and all the coefficients are presumed to be correlated with the treatment. An estimation of the model without restriction, where the number of identifiable composite parameters is no less than $(K+4)(K+1)/2$, might cause difficulty in numerical optimization. We might be able to estimate a simpler model if some of the random coefficients are not correlated. The Lagrange Multiplier (LM) test can be used for this purpose. The variable addition test (VAT) is also available, which is asymptotically equivalent to LM, but easier to apply. This test constructs a conditional mean by adding appropriately defined variables to create the likelihood function of which the score under restriction is the same as the one used in the LM test (for a detailed proof, see Wooldridge 2014). The actual test was performed by Wald test on the significance of coefficients of the added variables; look at the estimating Equation (9). Inside the exponential function, we have each covariate, the squares of each

covariate and their cross products. Let the vector of these functions of covariates be denoted by $\widetilde{\mathbf{x}}$ and rewrite the estimating equation for Regime 1 as follows:

$$E[y|\mathbf{z}, w = 1] = \exp(\widetilde{\mathbf{x}}\theta_1) \times \frac{\Phi\left(\mathbf{z}\gamma + \theta_2 + \mathbf{x}\theta_3\right)}{\Phi(\mathbf{z}\gamma)}, \tag{10}$$

where $\theta_2 = \sigma(u_1^a, v)$ and $\theta_3 = [\sigma(u_1^{b_1}, v), \cdots, \sigma(u_1^{b_K}, v)]'$. As already mentioned, θ_2 is scalar, θ_3 is a $K \times 1$ vector and θ_1 is a $(K+2)(K+1)/2 \times 1$ vector. The estimating equation for Regime 0 is the same as above, except for the negative sign inside the $\Phi(\cdot)$ in the correction function.

In order to perform VAT, we need to construct a conditional mean function, of which the score is identical to the one from Equation (10) under restriction. Let the restriction or the null hypothesis be

$$H_0 : \theta_2 = \theta_3 = 0.$$

In other words, there are no correlations between the selection error v and any random coefficients in the structural equations. The LM test can be applied and does not require an estimation of a complicated model without restriction. The VAT provides a device that facilitates this LM test by using an auxiliary regression for Regime 1

$$\exp\left(\widetilde{\mathbf{x}}\theta_1 + \theta_2\lambda(\mathbf{z}\gamma) + \theta_3\lambda(\mathbf{z}\gamma)\mathbf{x}\right), \tag{11}$$

where $\lambda(\cdot)$ is the inverse Mill's ratio. The auxiliary regression for Regime 0 will have $-\mathbf{z}\gamma$ instead of $\mathbf{z}\gamma$. It can be easily verified that the likelihood and score functions derived from (11) under restrictions are the same as those from (10). Therefore, the score tests on the significance of coefficients from the above auxiliary regression are equivalent to the LM tests in (10). Thus, the Wald tests for θ_2 and θ_3 in (11) will give a simple and asymptotically equivalent way to test the null without estimating the complicated model without restriction. In the actual test, $\lambda(\mathbf{z}\gamma)$ has to be estimated through the first-stage regression.

4.2. Model Selection Test

What we have considered above is whether or not the random coefficients are correlated with the selection error. Even when the null hypothesis is not rejected, it does not warrant coming back to the exponential ES model by Terza (2009), where all the coefficients are constants. Both the null and alternative hypotheses in the VAT mentioned above assume the presence of random coefficients.

One may want to run a test for model selection between the current model and the one with constant coefficients. In the latter model, all the individual heterogeneity is in the additive error. Since $u^{b_j} = 0$, all the variance and covariance terms involving u^b vanish. From Equation (7), the null hypothesis is the zero restriction on the coefficients of x_j^2 and of interactions $x_j x_r$ inside the exponential functions, and of x_j inside the correction functions.

5. Monte Carlo Simulations

There are three main purposes of the simulation. First, we compare three ATE estimators: the exponential endogenous switching regression with CRC (E-CRC, main topic of this article), the one without CRC (E-nCRC), and the linear version of the endogenous switching regression with CRC (L-CRC) proposed by Wooldridge (2015). This will show the advantage of the E-CRC model for estimating ATE. Second, it will see the impact of violation of probit assumption based on which the estimating equation was derived in the previous section. Third, it will check the reliability of the proposed estimator when the conditional mean is linear instead of being exponential. In the following simulations, the E-CRC and E-nCRC are all estimated by the two-step Poisson Quasi-Maximum Likelihood method.

5.1. Data-Generating Processes

The following data-generating processes (DGP) are inspired by the Oregon experiment discussed in the next section. We designed the DGP so that it resembles its data set. DGPs 1 and 2 are based on the following specifications: $x \sim$ uniform$[15, 50]$ and z is 0 or 1 with $1/2$ probability. Additionally,

$$w = 1[1.4 - 0.05x - 0.3z + v \geq 0] \tag{12}$$

$$E(y_1|x, z, u_1^a, u_1^b) = \exp\left(0.15 + u_1^a + (0.02 + u_1^b)x\right) \tag{13}$$

$$E(y_0|x, z, u_0^a, u_0^b) = \exp\left(0.1 + u_0^a + (0.0059 + u_0^b)x\right), \tag{14}$$

where y_1 and y_0 follow the Poisson distribution with the mean specified above. The above DGP implies that the ATE is in unity.

There are two sessions of simulations for the above population; DGP 1 deals with the ideal case where the errors follow the multivariate normal distribution. The joint distribution of errors $(u_0^a, u_0^b, u_1^a, u_1^b, v)$ for DGP 1 follows the multivariate normal with equicorrelation for any pair of two errors. The variance of v is set as equal to one and those of other errors are all equal to $1/100^2$. The coefficient on x and the error variances might seem too small at first sight. As the coefficient on x is interpreted as the semi-elasticity in exponential models, the values are chosen so that the standard deviation is $1/100$ or 1%. Therefore, in Regime 1, the mean of the coefficient is 2% and its standard deviation is 1%. The values of correlations (ρ) are 0.3 and 0.5. This is the case where the above estimator might perform the best, since the actual DGP conforms to the assumptions on which the estimating equation is based.

An important assumption for Lemma 1 is that v follows a normal distribution, which has brought the normal cumulative distribution function $\Phi(\cdot)$ in the correction terms. In order to check the robustness of the estimator on that assumption, DGP 2 generates errors that follow a skewed distribution. Each error is constructed by $(X - 5)/\sqrt{10}$, where $X \sim \chi^2(5)$, so that it has zero mean and unit variance. Then, it is again divided by 100 to set the variance to the same as that in DGP 1. Each error was constructed to have the specified correlations.[1]

Lastly, DGP 3 uses the same x, z and w, except that x is demeaned, i.e., $x \sim$ uniform $[-17.5, 17.5]$. The dependent variables for two different regimes follow these equations:

$$y_1 = 5 + u_1^a + (0.01 + u_1^b)x + \epsilon_1 \tag{15}$$

$$y_0 = 4 + u_0^a + (0.02 + u_0^b)x + \epsilon_0, \tag{16}$$

where the joint distribution of $(u_0^a, u_0^b, u_1^a, u_1^b, v)$ is the same as DGP 1.[2] The structural errors (ϵ_1, ϵ_0) follow the independent standard normal distribution. This is to compare the relative performances of linear and exponential models when the population is linear.

5.2. Simulation Results

The simulation results for estimating ATE under DGPs 1, 2 and 3 are given in Tables 1–3, respectively. The left and right panels show the results for $\rho = 0.3$ and 0.5 in each table. Each panel lists and compares these three estimators for sample sizes of 5000 and 10,000. Mc. st. dev., RMSE, MAD and IR stand for Monte Carlo standard deviation, root mean squared error, mean absolute deviation, and interdecile range, respectively.

Table 1. ATE estimations under DGP 1.

		$\rho = 0.3$			$\rho = 0.5$		
		L-CRC	**E-nCRC**	**E-CRC**	**L-CRC**	**E-nCRC**	**E-CRC**
n = 5000	bias	0.133	0.024	0.003	0.189	0.056	0.039
	mc. st. dev	0.454	0.485	0.658	0.454	0.460	0.590
	RMSE	0.473	0.486	0.658	0.492	0.463	0.591
	median	0.863	0.953	0.965	0.811	0.942	0.963
	MAD	0.320	0.317	0.371	0.302	0.303	0.309
	IR	1.138	1.197	1.471	1.146	1.117	1.368
n = 10,000	bias	0.137	0.024	0.006	0.190	0.044	0.013
	mc. st. dev	0.318	0.325	0.394	0.328	0.319	0.370
	RMSE	0.347	0.326	0.394	0.379	0.322	0.370
	median	0.879	0.980	0.986	0.815	0.959	0.963
	MAD	0.215	0.219	0.252	0.221	0.208	0.231
	IR	0.797	0.819	0.932	0.829	0.775	0.916

Note: L-CRC, E-nCRC and E-CRC denote the linear ES with CRC, the exponential ES without CRC and the exponential ES with CRC. Bias, mc st dev, RMSE, MAD and IR indicate the absolute value of bias, Monte Carlo standard deviation, root mean squared error, mean absolute deviation and interdecile range. The simulation is run with 500 replications.

Table 2. ATE estimations under DGP 2.

		$\rho = 0.3$			$\rho = 0.5$		
		L-CRC	**E-nCRC**	**E-CRC**	**L-CRC**	**E-nCRC**	**E-CRC**
n = 5000	bias	0.257	0.017	0.188	0.303	0.020	0.262
	mc. st. dev	0.607	0.577	0.964	0.638	0.574	0.932
	RMSE	0.659	0.577	0.982	0.706	0.574	0.968
	median	0.763	0.928	0.998	0.763	0.998	1.125
	MAD	0.385	0.346	0.462	0.442	0.383	0.472
	IR	1.528	1.499	2.060	1.661	1.508	1.810
n = 10,000	bias	0.240	0.046	0.085	0.318	0.019	0.128
	mc. st. dev	0.386	0.373	0.543	0.441	0.384	0.500
	RMSE	0.454	0.375	0.550	0.544	0.384	0.517
	median	0.753	0.923	1.009	0.694	0.967	1.062
	MAD	0.281	0.256	0.309	0.279	0.235	0.302
	IR	1.005	0.933	1.300	1.081	0.951	1.154

Table 3. ATE Estimations under DGP 3.

		$\rho = 0.3$			$\rho = 0.5$		
		L-CRC	**E-nCRC**	**E-CRC**	**L-CRC**	**E-nCRC**	**E-CRC**
n = 5000	bias	0.012	0.016	0.028	0.003	0.021	0.021
	mc. st. dev	0.283	0.290	0.311	0.275	0.289	0.298
	RMSE	0.283	0.290	0.312	0.275	0.290	0.299
	median	1.026	1.011	1.030	1.028	1.007	1.035
	MAD	0.200	0.186	0.193	0.175	0.185	0.200
	IR	0.733	0.716	0.773	0.706	0.701	0.726
n = 10,000	bias	0.020	0.022	0.001	0.021	0.033	0.002
	mc. st. dev	0.219	0.195	0.243	0.214	0.197	0.228
	RMSE	0.220	0.196	0.243	0.215	0.200	0.228
	median	0.987	1.026	1.013	0.988	1.032	1.010
	MAD	0.145	0.129	0.157	0.143	0.135	0.153
	IR	0.552	0.498	0.593	0.529	0.518	0.575

The E-CRC is expected to perform the best when the errors follow a multivariate normal distribution. In Table 1, the biases of E-CRC are smaller than those of L-CRC. The

E-CRC has larger Monte Carlo variances than L-CRC under both the correlations. This might be a sign that the efficiency and the consistency are traded. One can also see that, in Table 1, E-CRC has smaller bias, but is less efficient than E-nCRC. These larger variances are simply due to the complexity of the model. We again face the trade-off between consistency and efficiency.

Table 2 shows the effect of non-normal errors of DGP 2 that may cause some biases for the exponential models because the correction terms were derived under the normality of v. The comparison of L-CRC and E-CRC in Table 1 is also echoed here in Table 2; E-CRC has larger variances, but smaller biases than L-CRC. However, E-CRC has both larger biases and variances than E-nCRC. The simpler model performs relatively better under error misspecification.

DGP 3 in Table 3 checks whether the exponential regression is generally as reliable as the linear one when the dependent variable is non-negative. The overall results from Table 3 do not reveal any substantial difference from Tables 1 and 2. The performances of the linear model under the linear DGP 3 are no better than that of the exponential model in terms of bias; biases from exponential models are in most cases smaller than in linear models. Larger variances of exponential models are also found here, as in DGPs 1 and 2. The true motivation of the DGP 3 simulations is the possibility of incorrect data fitting, and its impact on the ATE estimation. However, such an adverse effect of the exponential regression has not been found by simulations. The exponential CRC model may not be necessarily worse than the linear model even under the linear population. The CRCs generate quadratic terms in the estimating equations. For instance, a linear function $a + bx$ can be better fitted by an exponential function with quadratic terms because $\ln(a + bx)$ can be better approximated by a quadratic equation. Thus, the quadratic terms, which might not be present in a non-CRC setting, help the model be more robust than the linear model. Thus, we expect that the exponential CRC might be well suited for prediction as well as the ATE estimation.

6. An Application: Oregon's Health Insurance Experiment

One of the motivations of expanding Medicaid to those currently not qualified is that it can shift treatment from costly emergency department care to more effective primary care. Health insurance is presumed to reduce the use of emergency departments by encouraging patients to use preventive care. A way to learn about the effect of expanding Medicaid would be to randomly offer coverage to those without it and to compare the outcomes between the treatment and the control groups.

The state of Oregon initiated such an experiment in 2008 through a publicly announced health insurance policy (Finkelstein et al. 2012; Baicker et al. 2004; Taubman et al. 2014). Oregon initially allowed residents without health insurance to register for a lottery, selecting winners and losers. Only the winners were then given a chance to apply for the Oregon version of Medicaid, known as the Oregon Health Plan (OHP). The OHP coverage was not automatic for the winners, however. It was provided only to the applicants who were US citizens or legal immigrants, aged 19–64, and who satisfied some other additional criteria. These selected people had not been previously eligible for Medicaid based on the federal poverty income level and had been without insurance for at least previous six months.

We are mostly interested in the effect of health insurance on the number of emergency room (ER) use. Taubman et al. (2014) use the following regression model

$$vafter_{ih} = \beta_0 + \beta_1 OHP_{ih} + vbefore_{ih}\beta_2 + famsize_{ih}\beta_3 + u_{ih}, \tag{17}$$

where i denotes an individual and h denotes a household. There might be multiple individuals from the same household in the data set. $vafter$, $vbefore$ and $famsize$ are the numbers of ER use after and before OHP and the number of family members in the household. The instrument is binary and indicates whether an individual is a lottery winner. If an individual purchases OHP, OHP = 1, and he or she will be part of Regime 1. If not, they will be part of regime 0.

There are a couple of issues in Taubman et al. 2014. First, they implicitly assume that $vbefore$ is exogenous. Although potentially inaccurate, Keay (2019); Fourkan et al. (2021) find, by a copula regression in a slightly different setting, that the endogeneity of $vbefore$ is of little concern. Second, they run 2SLS without paying enough attention to the fact that $vafter$ and $vbefore$ are censored at 22 and 17. Developing a censored data version of the model is not in the scope of the current research; here, we will alleviate this issue by excluding observations with censored covariates. Reducing the support of covariates does not affect the consistency in general. Even better, it eliminates the source of bias by eliminating wrong covariate values. The observations at censored point 17 must be thrown out. In the data set, 1.2% of observations have a $vbefore$ of over 10. A few large values of $vbefore$ over 10 either make the maximum likelihood method unstable or cause an enormous standard error. For the purpose of a reliable comparison, I will consider the subsamples with $vbefore \leq 9$, 7, and 5, which make up 98.8%, 98.0% and 96.2% of the original data set. Each subsample contains 36%, 29% and 21% of the original number of y-censored points.

We ran regressions of (17) and its exponential regime switching versions. The dependent variable is the count variable, and thus the exponential regression models, such as E-nCRC or E-CRC, would be the alternatives to the conventional linear models. In particular, E-CRC allows for the presence of a random coefficient on $vbefore$. For a person who is more willing to take up the insurance (measured by higher v), their current ER visiting behaviours would be different than those from the past. This may weaken the link between the current and past visits so that the coefficient $b^1_{vbefore}$ is closer to zero. Then, the covariance between them can be negative. If the covariance is non-zero, then the ATE estimator without CRC can be biased. This is the reason why the E-CRC model is estimated.

Table 4 shows the descriptive statistics for the variables in the data set. Table 5 compares OLS, 2SLS, E-nCRC, E-CRC estimates. Panels A, B and C display the results for the subsamples with $vbefore \leq 9$, 7, and 5. On OHP (ATE), the coefficient estimates of OHP from OLS and 2SLS and the ATE estimates of E-nCRC and E-CRC are displayed. R1 and R0 stand for the regimes with $OHP = 1$ and 0, respectively. The covariates coefficients for OLS and 2SLS are all displayed in R1. $b^1_{vbefore}$ is the random coefficient on $vbefore$, where the superscript and subscript indicate the regime and the covariate that this random coefficient is on.

Table 4. Descriptive statistics.

Variable	Mean	Std. Dev.	Min	Max
vafter	0.997	2.410	0	22
vbefore	0.774	1.863	0	17
famsize	1.210	0.409	1	3
OHP	0.241	0.427	0	1
lottary	0.391	0.488	0	1

The 2SLS estimates, the weighted averages of the LATE for each covariate (Angrist and Imbens 1994), are all greater than the OLS across all the subsamples. They all indicate that the effects of insurance are significantly positive. On Panels B and C, the ES estimator gives the ATE not far from LATE. Their standard errors are larger because the model is more complex; the ATE values are all insignificant in all the samples. The 2SLS shows the treatment effect only for compliers, i.e., those who get an insurance because they are eligible. This subpopulation does not include the always takers and never takers. In the data, there are at least 2200 always takers who got insurance although they are not eligible, and at least 5897 never takers who did not get insurance although they are eligible. They make up about 1/3 of the whole observations in the data set. The 2SLS estimate says that the effect of the insurance is positive only for them. E-nCRC and E-CRC indicate that the average treatment effect for the whole population is not statistically different from zero.

This might be because the positive effect of the compliers is diluted by the negative effect of never and always takers. Another reason is that the standard errors are larger. The ATE estimates, unlike other parameters, are affected by the randomness of covariates. The additional term $E[T]^2$ in the asymptotic variance of ATE helps increase the standard errors.

The estimation results show that the ATE (or the coefficient on OHP) estimates are smaller when those with unusually large *vbefore* are eliminated. Although not significant, the ATE estimate for E-CRC in Panel C is even negative. In the data, 2.6% of the individuals visited the ER more than five times. Dropping them changes the magnitude of IV estimates by more than 10%. These indicate that Taubman et al. (2014)'s results are exaggerated by the handful of frequent visitors who may not be from the common distribution. The t-statistics are also lower in the smaller samples, indicating that the evidence for the positive effect of OHP is weaker therein.

A presence of the CRC is detected. The correlations between the coefficient $b^1_{vbefore}$ and the selection error v are significantly negative. The results show the presence of CRC only on *vbefore*, particularly in Regime 1. The CRC on *vbefore* was not taken into account in other models and might have caused biases. Some of those who changed the ER visiting behaviour after obtaining insurance ended up visiting less, which could have helped reduce the ATE estimates.

Table 5. Regression results.

	Panel A. $vbefore \leq 9$				**Panel B.** $vbefore \leq 7$				**Panel C.** $vbefore \leq 5$			
	OLS	**2SLS**	**E-nCRC**	**E-CRC**	**OLS**	**2SLS**	**E-nCRC**	**E-CRC**	**OLS**	**2SLS**	**E-nCRC**	**E-CRC**
$OHP(ATE)$	0.445 ***	0.342 ***	0.329	0.524	0.428 ***	0.330 ***	0.450	0.175	0.427 ***	0.300 ***	0.168	−0.079
	(0.027)	(0.097)	(1.129)	(3.085)	(0.026)	(0.093)	(0.519)	(1.734)	(0.025)	(0.089)	(0.389)	(2.118)
R1: $vbefore$	0.728 ***	0.731 ***	0.272 ***	0.313 *	0.721 ***	0.724 ***	0.311 ***	0.380 *	0.688 ***	0.693 ***	0.390 ***	0.066
	(0.009)	(0.009)	(0.010)	(0.167)	(0.010)	(0.010)	(0.012)	(0.228)	(0.011)	(0.012)	(0.017)	(0.285)
$famsize$	−0.252 ***	−0.249 ***	−0.329 ***	−0.417	−0.251 ***	−0.248 ***	−0.384 ***	−0.372	−0.250 ***	−0.246 ***	−0.406 ***	−2.172
	(0.028)	(0.028)	(0.084)	(1.825)	(0.027)	(0.027)	(0.076)	(1.728)	(0.025)	(0.025)	(0.077)	(1.837)
$cov(b^1_{vbefore}, v)$				−0.240 ***				−0.367 ***				−0.788 *
				(0.064)				(0.089)				(0.449)
$cov(b^1_{famsize}, v)$				−1.111				−1.004				−1.657 *
				(0.897)				(0.794)				(0.875)
R0: $vbefore$			0.333 ***	0.728 ***			0.408 ***	0.586 **			0.507 ***	0.863 ***
			(0.010)	(0.140)			(0.011)	(0.287)			(0.015)	(0.302)
$famsize$			−0.435 ***	−0.089			−0.374 ***	−0.507			−0.418 ***	−2.915 *
			(0.067)	(2.567)			(0.072)	(2.442)			(0.060)	(1.700)
$cov(b^0_{vbefore}, v)$				0.022				−0.058				−0.035
				(0.095)				(0.139)				(0.143)
$cov(b^0_{famsize}, v)$				−0.863				−1.361				−1.523
				(1.065)				(1.360)				(1.190)

Note: *OHP* shows the OLS and 2SLS estimates of the coefficient on *OHP*, and the ATE estimates of E-nCRC and E-CRC. $cov(b^1_{vbefore}, v)$ and $cov(b^1_{famsize}, v)$ indicate the covariances between the selection error and the coefficients on *vbefore* and *famsize* in Regime 1. The numbers report the estimates and standard errors in parenthesis. *, ** and *** indicate the significance at 10%, 5% and 1% levels, respectively. Panels A, B and C display the results for the subsamples with $vbefore \leq 9, 7$ and 5, which make up 98.8%, 98.0% and 96.2% of the original data set.

7. Conclusions

We have so far considered the endogenous switching regression, where there is a random coefficient correlated with the switching variable under the count dependent variable. The count dependent variable requires non-linear modeling using the exponential conditional mean function. Although it is impossible to identify the structural parameters, the identification of ATE has been achieved. The simulations show that this ATE estimator, by the non-linear two-regime CRC model, performs reasonably well at larger sample sizes. It is also found that the exponential model performs well as long as the dependent variable is non-negative, even when the population conditional mean is linear. The empirical application has shown that the CRC model allows us to study the nature of self-selection

more fully. Relaxing the probit assumption would be an obvious extension. By introducing the first-stage equation in a more general form, it would be possible to relax the distribution assumption along with the homogeneous instrument effects at the same time.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/SJG1ED.

Acknowledgments: I am grateful to Jeff Wooldridge, Peter Schmidt, Tim Vogelsang, Todd Elder, and Hira Koul for their support and guidance. I also thank Otavio Bartalotti, Seong Yeon Chang, Hidehiko Ichimura, Seojeong Lee, Valentin Verdier and other seminar participants at Michigan State University, Georgia Southern University, AMES 2014, SETA 2014 at Taipei, Taiwan, and KEA 2014 at Seoul, Korea. I also would like to extend my sincere thanks to the anonymous referees for their helpful comments.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Proof for Lemma 1. Consider Regime 1 only. Regime subscripts are suppressed for notational simplicity. The derivation for Regime 0 is almost identical. Note that

$$\begin{aligned}
E[\exp(u^a + \mathbf{x}u^b)|\mathbf{z}, w=1] &= E\Big[\exp\Big(\sigma(u^a,v)v + e^a + \mathbf{x}\sigma(u^b,v)v + \mathbf{x}e^b\Big)\Big|\,\mathbf{z}, w=1\Big] \\
&= E\Big[\exp\Big((\sigma(u^a,v) + \mathbf{x}\sigma(u^b,v))v\Big)\exp(e^a + \mathbf{x}e^b)\Big|\,\mathbf{z}, w=1\Big] \\
&= E\Big[\exp\Big((\sigma(u^a,v) + \mathbf{x}\sigma(u^b,v))v\Big)\Big|\,\mathbf{z}, w=1\Big] \\
&\qquad \times E[\exp(e^a + \mathbf{x}e^b)|\mathbf{z}, w=1] \\
&= E\Big[\exp\Big((\sigma(u^a,v) + \mathbf{x}\sigma(u^b,v))v\Big)\Big|\,\mathbf{z}, w=1\Big] \\
&\qquad \times E[\exp(e^a + \mathbf{x}e^b)|\mathbf{z}] \\
&= \frac{\Phi\Big(z\gamma + \sigma(u^a,v) + \mathbf{x}\sigma(u^b,v)\Big)}{\Phi(z\gamma)}\exp\left(\frac{(\sigma(u^a,v) + \mathbf{x}\sigma(u^b,v))^2}{2}\right) \\
&\qquad \times E[\exp(e^a + \mathbf{x}e^b)|\mathbf{z}],
\end{aligned}$$

where the third equality comes from $v \perp\!\!\!\perp e^a + \mathbf{x}e^b \mid \mathbf{z}, w$, which is implied by Assumption 3(ii), and the fourth from Assumption 3(i). The last equality is shown by Terza (1998). The argument of the exponential function at the last line can be written as

$$\begin{aligned}
\Big(\sigma(u^a,v) + \mathbf{x}\sigma(u^b,v)\Big)^2 &= \Big(\sigma(u^a,v) + \sum_{j=1}^{K}\sigma(u^{b_j},v)x_j\Big)^2 \\
&= \sigma^2(u^a,v) + 2\sum_{j=1}^{K}\sigma(u^a,v)\cdot\sigma(u^{b_j},v)x_j + \sum_{j=1}^{K}\sigma^2(u^{b_j},v)x_j^2 \\
&\quad + \sum_{j=1}^{K}\sum_{r\neq j}\sigma(u^{b_j},v)\cdot\sigma(u^{b_r},v)x_jx_r,
\end{aligned}$$

where $\sigma^2(\cdot,\cdot)$ is a squared covariance.

Now consider the last term at the last line. Since $e^a + \mathbf{x}e^b$ is a sum of multiple random variables, it can be approximately thought of as following a normal distribution. Then, by using the results on the log-normal distribution [3]

$$E[\exp(e^a + \mathbf{x}e^b)|\mathbf{z}] \approx \exp\left(\text{var}[e^a + \mathbf{x}e^b \mid \mathbf{x}]\Big/2\right)$$

where

$$\begin{aligned}\text{var}[e^a + \mathbf{x}e^b \mid \mathbf{x}] &= \text{var}(e^a) + \text{var}(\mathbf{x}e^b) + 2\text{cov}(e^a, \mathbf{x}e^b) \\ &= \sigma^2(e^a) + \sum_{j=1}^{K}\sigma^2(e^{b_j})x_j^2 + \sum_{j=1}^{K}\sum_{r\neq j}\sigma(e^{b_j}, e^{b_r})x_j x_r + 2\sum_{j=1}^{K}\sigma(e^a, e^{b_j})x_j.\end{aligned}$$

Collecting these terms along with the above second term, we obtain the stated result. □

Proof for Lemma 2. Note that

$$\begin{aligned}E(u^a + \mathbf{x}u^b|\mathbf{x}) &= E(u^a|\mathbf{x}) + E(\mathbf{x}u^b|\mathbf{x}) = 0 \\ \text{var}(u^a + \mathbf{x}u^b|\mathbf{x}) &= \text{var}(u^a|\mathbf{x}) + \text{var}(\mathbf{x}u^b|\mathbf{x}) + 2\text{cov}(u^a, \mathbf{x}u^b|\mathbf{x}) \\ &= \sigma^2(u^a) + \sum_{j=1}^{K}\sigma^2(u^{b_j})x_j^2 + \sum_{j=1}^{K}\sum_{r\neq j}\sigma(u^{b_j}, u^{b_r})x_j x_r + 2\sum_{j=1}^{K}\sigma(u^a, u^{b_j})x_j.\end{aligned}$$

For any fixed value of $\mathbf{x}$, the distribution of $u^a + \mathbf{x}u^b$ is approximately normal due to the central limit theorem. Since the mean and variance of normal random variables are already obtained, the mean of its log-normal variable is trivially found. □

Proof of Proposition 1. [4]

Let the continuous non-linear function of Regime 1 be denoted as $g_1(x, \theta)$ and similar for regime 0. In the current context $g_1(x, \theta_1) = \exp(\alpha_1 + \sigma_1^2/2 + x\beta_1)$.

Note that the structural parameters estimator satisfies

$$\sqrt{N}(\widehat{\theta} - \theta) = -N^{-1/2}\sum A^{-1}s(\theta) = o_p(1)$$

By the mean value theorem, there exists $\widetilde{\theta} \in [\widehat{\theta}, \theta]$ such that the derivatives of $g_1(x, \theta)$ and $g_0(x, \theta)$ evaluated at $\widetilde{\theta}$ satisfy

$$g_1(x, \widehat{\theta}) - g_0(x, \widehat{\theta}) = \Big(g_1(x, \theta) - g_0(x, \theta)\Big) + \Big(\partial g_1(\widetilde{\theta})/\partial\theta - \partial g_0(\widetilde{\theta})/\partial\theta\Big)'(\widehat{\theta} - \theta).$$

Taking the average over the observations,

$$N^{-1/2}\sum\Big(g_1(x, \widehat{\theta}) - g_0(x, \widehat{\theta})\Big) = N^{-1/2}\sum\Big(g_1(x, \theta) - g_0(x, \theta)\Big) + N^{-1}\sum\Big(\partial g_1(\widetilde{\theta})/\partial\theta - \partial g_0(\widetilde{\theta})/\partial\theta\Big)'\sqrt{N}(\widehat{\theta} - \theta)$$

Let

$$G_1 = E\left[\frac{dg_1(x, \theta)}{d\theta}\right], \qquad G_0 = E\left[\frac{dg_0(x, \theta)}{d\theta}\right].$$

Since $\widetilde{\theta} \to_p \theta$, by Lemma 12.1 in the work of Wooldridge (2010)

$$N^{-1}\sum\Big(\partial g_1(\widetilde{\theta})/\partial\theta - \partial g_0(\widetilde{\theta})/\partial\theta\Big) = G_1 - G_0 + o_p(1).$$

Note that $\sqrt{N}(\widehat{\theta} - \theta) = O_p(1)$. Plugging the above, we obtain

$$\begin{aligned} N^{-1/2}\sum\Big(g_1(x,\widehat{\theta})-g_0(x,\widehat{\theta})\Big) &= N^{-1/2}\sum\Big(g_1(x,\theta)-g_0(x,\theta)\Big)+(G_1-G_0)'\sqrt{N}(\widehat{\theta}-\theta)+O_p(1)o_p(1) \\ &= N^{-1/2}\sum\Big(g_1(x,\theta)-g_0(x,\theta)\Big)-N^{-1/2}\sum(G_1-G_0)'A^{-1}s(\theta)+o_p(1) \end{aligned}$$

Then,

$$\begin{aligned} &\sqrt{N}\left(\frac{\sum(g_1(x,\widehat{\theta})-g_0(x,\widehat{\theta}))}{N}\right)-\sqrt{N}E[g_1(x,\theta)-g_0(x,\theta)] \\ &= N^{-1/2}\sum\Big(g_1(x,\theta)-g_0(x,\theta)-\Big(E[g_1(x,\theta)]-E[g_0(x,\theta)]\Big)-(G_1-G_0)'A^{-1}s(\theta)\Big)+o_p(1). \end{aligned}$$

The expectation of the right hand side is equal to zero. Using the Lindberg–Levy Theorem

$$\sqrt{N}(\widehat{ATE}-ATE)\to_d N(0,V),$$

where

$$\begin{aligned} V &= \mathrm{Var}\left[g_1(x,\theta)-g_0(x,\theta)-\Big(E[g_1(x,\theta)]-E[g_0(x,\theta)]\Big)-(G_1-G_0)'A^{-1}s(\theta)\right] \\ &= \mathrm{Var}\left[T-(G_1-G_0)'A^{-1}s(\theta)\right] \\ &= E[T^2]+(G_1-G_0)'A^{-1}BA^{-1}(G_1-G_0)-2\mathrm{Cov}\Big(T,(G_1-G_0)'A^{-1}s(\theta)\Big), \end{aligned}$$

where $T\equiv g_1(x,\theta)-g_0(x,\theta)-\Big(E[g_1(x,\theta)]-E[g_0(x,\theta)]\Big)$.

Note that

$$\begin{aligned} \mathrm{Cov}(T,s(\theta)) &= E(Ts(\theta))-E(T)E(s(\theta)) \\ &= E(Ts(\theta))=E[TE(s(\theta)|x)]=0 \end{aligned}$$

as was shown. □

Notes

1 One can use the copula method to generate a multivariate distribution in which marginal effects are scaled $\chi^2(5)$ distribution. However, the copula method was not used here, since we only want to have pairs of errors with some predetermined correlation without further specifying the dependence structure. Dependence can also be created by using some common standard normal random variables from which two different chi-squared random variables are constructed. We obtained the desired correlation by changing the "parameters" through trial and error. Codes will be provided upon request.

2 The intercept values are sufficiently large to prevent any negative dependent variable. In 10,000 replications, there are at most one or two negative dependent variables. They were dropped and the effect is negligible.

3 Terza (2009) assumes the joint normality of (v,u^a,u^b). On the contrary, my model does not assume the jointicity. By linear projection, $u^a=\rho v+e^a$ and e^a is orthogonal or independent with v. The only requirement is the normality of e^a and e^b. However, the CLT makes the approximation more accurate as the number of covariates increases. My model presupposes the existence of covariates with CRC. Therefore, the more appropriate the model is, the more accurate the approximation is. Finally, according to the simulation results for χ^2 errors in Tables 1 and 2, under large samples, the performances of CRC model is better than the linear model. So, again, when the model is appropriate, it outperforms the preexisting linear method.

4 Terza (1998) derives in his unpublished manuscript the same asymptotic distribution by a different method. I was not aware of this manuscript at the time of writing this proof.

References

Abadie, Alberto. 2003. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics* 113: 231–63. [CrossRef]

Angrist, Joshua, and Guido Imbens. 1994. Identification and Estimation of Local Average Treatment Effects. *Econometrica* 62: 467–75.

Angrist, Joshua D., and Guido W. Imbens. 1995. Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. *Journal of the American Statistical Association* 90: 431–42. [CrossRef]

Baicker, Katherine, Sarah L. Taubman, Heidi L. Allen, Mira Bernstein, Jonathan H. Gruber, Joseph P. Newhouse, Eric C. Schneider, Bill J. Wright, Alan M. Zaslavsky, and Amy N. Finkelstein. 2013. The Oregon Experiment-Effects of Medicaid on Clinical Outcomes. *New England Journal of Medicine* 368: 1713–22. [CrossRef]

Blundell, Richard W., and James L. Powell. 2004. Endogeneity in Semiparametric Binary Response Models. *Review of Economic Studies* 71: 655–79. [CrossRef]

Browning, Martin, and Jesus Carro. 2007. Heterogeneity and microeconometrics modelling. In *Advances in Economics and Econometrics: Theory and Applications III*. Edited by Richard Blundell, Whitney Newey and Torsten Persson. Cambridge: Cambridge University Press.

Browning, Martin, and Valerie Lechene. 2003. Children and Demand: Direct and Non-Direct Effects. *Review of Economics of the Household* 1: 9–31. [CrossRef]

Card, David. 2001. Estimating the Return to Schooling: Progress on Some Persistent Econoemtric Problems. *Econometrica* 69: 1127–60. [CrossRef]

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and Oregon Health Study Group. 2012. The Oregon Health Insurance Experiment: Evidence from the First Year. *Quarterly Journal of Economics* 127: 1057–106. [CrossRef]

Garen, John. 1984. The Returns to Schooling: A Selectivity Bias Approach with a Continuous Choice Variable. *Econometrica* 52: 1199–218. [CrossRef]

Gourieroux, Christian, Alain Monfort, and Alain Trognon. 1984. Pseudo-Maximum Likelihood Methods: Theory. *Econometrica* 52: 681–700. [CrossRef]

Heckman, James J. 1976. The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement* 5: 475–92.

Heckman, James, and Edward Vytlacil. 1998. Instrumental Variables Methods for the Correlated Random Coefficient Model. *Journal of Human Resources* 33: 974–87. [CrossRef]

Heckman, James J., and Edward Vytlacil. 2005. Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica* 73: 669–738. [CrossRef]

Imbens, Guido W. 2007. Nonadditive models with endogenous regressors. In *Advances in Economics and Econometrics: Theory and Applications III*. Edited by Richard Blundell, Whitney Newey and Torsten Persson, Cambridge: Cambridge University Press.

Keay, Myoung-Jin. 2018. *Relative Efficiency of One-Step and Two-Step Control Function Estimators in Parametric Nonlinear Models*. Working Paper. Available online: https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxteW91bmdrZWF5fGd4OjdmOGZjZTEzMTBkMWVhYjE (accessed on 11 October 2021).

Keay, Myoung-Jin. 2019. The Effect of Extended Medicaid on the Emergency Department Visits: An analysis of the 2008 Oregon Health Program. *Journal of Applied Business and Economics* 21: 19–24.

Fourkan, Md, Myoung-Jin Keay, Na Kyeong Lee. 2021. Decreased Emergency Department Use Following Medicaid Expansion: Evidence from Oregon's Health Insurance Experiment. *International Economic Journal* 35: 314–22. [CrossRef]

Masten, Matthew A., and Alexander Torgovitsky. 2016. Indentification of Instrumental Variable Correlated Random Coefficients Models. *Review of Economics and Statistics* 98: 1001–5. [CrossRef]

Newey, Whitney K. 2009. Two-step Series Estimation of Sample Selection Models. *Econometrics Journal* 12: S217–29. [CrossRef]

Taubman, Sarah L., Heidi L. Allen, Bill J. Wright, Katherine Baicker, and Amy N. Finkelstein. 2014. Medicaid Increases Emergency Department Use: Evidence from Oregon's Health Insurance Experiment. *Science* 343: 263–68. [CrossRef]

Terza, Joseph V. 1998. Estimating Count Models with Endogenous Switching: Sample Selection and Endogenous Treatment Effects. *Journal of Econometrics* 84: 129–54. [CrossRef]

Terza, Joseph V. 2009. Parametric nonlinear regression with endogenous switching. *Econometric Reviews* 28: 555–80. [CrossRef]

Wooldridge, Jeffrey M. 2003. Further Results on Instrumental Variables Estimation of Average Treatment Effect in the Correlated Random Coefficient Model. *Economics Letters* 79: 185–91. [CrossRef]

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. Cambridge: The MIT Press.

Wooldridge, Jeffrey M. 2014. Quasi-Maximum Likelihood Estimation and Testing for Nonlinear Models with Endogenous Explanatory Variables. *Journal of Econometrics* 182: 226–34. [CrossRef]

Wooldridge, Jeffrey M. 2015. Control Function Methods in Applied Econometrics. *Journal of Human Resources* 50: 420–45. [CrossRef]