2015

Estimating the Variance of Decomposition Effects

Takuya Hasebe
Sophia University

CUNY GRADUATE CENTER PH.D PROGRAM IN ECONOMICS
WORKING PAPER SERIES

Estimating the Variance of Decomposition Effects

Takuya Hasebe

Working Paper 6

Ph.D. Program in Economics
CUNY Graduate Center
365 Fifth Avenue
New York, NY 10016
April 2015

Estimating the Variance of Decomposition Effects
Takua Hasebe
April 2015
JEL No: C10, J70

ABSTRACT

We derive the asymptotic variance of Blinder-Oaxaca decomposition effects. We show that the delta method approach that builds on the assumption of fixed regressors understates true variability of the decomposition effects when regressors are stochastic. Our proposed variance estimator takes randomness of regressors into consideration. Our approach is applicable to both the linear and nonlinear decompositions, for the latter of which only a bootstrap method is an option. As our derivation follows the general framework of m-estimation, it is straightforward to extend to the cluster-robust variance estimator. We demonstrate the finite-sample performance of our variance estimator with a Monte Carlo study and present a real-data application.

Takuya Hasebe
Faculty of Liberal Arts, Sophia University
7-1 Kioi-cho, Chiyoda-ku
Tokyo 102-8554, Japan
thasebe@sophia.ac.jp

1 Introduction

Since the influential seminar works by Blinder (1973) and Oaxaca (1973), the decomposition method has been used to analyze racial, gender, and intertemporal differences and more. In addition to the original linear model to decompose wages, the method has been extended to nonlinear models to analyze limited dependent variables such as binary and count data outcomes. The decomposition method became a popular tool in empirical studies not only in labor economics but also in other areas such as health economics. Fortin et al. (2011) provides an excellent survey of the decomposition method. Moreover, the recent discussion about the connection to the literature of treatment effects (Fortin et al., 2011; Kline, 2011) makes the decomposition method a even more valuable tool for applied researchers.

Although the decomposition method has been used for a long time, it is relatively recently that statistical inference of the decomposition analysis has been discussed. In early times, results of the decomposition analysis were presented without standard errors. Oaxaca and Ransom (1998) propose the variance estimator derived by the delta method. However, it builds on the implicit assumption of fixed regressors. When regressors are stochastic, which is a more plausible assumption in most empirical studies, the delta method variance tends to overstate statistical significance by ignoring the variability of regressors. Jann (2008) suggests a variance estimator with stochastic regressors for the linear decomposition. Kline (2014) also derives the asymptotic distribution of a variant of the linear decomposition and shows that ignoring of the variability of regressors results in incorrect inference.

The primary contribution of this paper is to derive the asymptotic variance of the nonlinear decomposition, which is also applicable to the linear decomposition. For nonlinear models, as Fortin et al. (2011) suggest, only a bootstrap approach has been a valid option. However, the bootstrap estimation of the variance is often computationally demanding. Therefore, the analytical variance estimator of nonlinear models must be of practical use for

applied researchers. Monte Carlo experiments demonstrate that our proposed variance estimator indeed leads to correct statistical inference. A real-data application also show that our variance estimates are almost identical to the bootstrap estimates.

Secondly, since our derivation of the asymptotic variance is based on the general framework of m-estimation, it is easily extendable to various settings. As an example, we extend our variance estimator to a cluster-robust variance following Cameron et al. (2011). Our Monte Carlo study shows that our variance estimator performs well even in the presence of clustering correlation. In addition, the analytical variance is essential to obtain asymptotic refinement through the bootstrap method for more reliable inference Cameron et al. (2008).

The rest of this paper is organized as follows. In the next section, we introduce the decomposition analysis. Section 3 discusses the estimator of the decomposition effects and derives the asymptotic distribution. ection 4 presents results of a Monte Carlo study, followed by a real-data application in Section 5. Section 6 concludes.

2 Decomposition Analysis

This section introduces the decomposition analysis. Our focus is the decomposition in the mean of outcome. See Fortin et al. (2011) for recent developments of the decomposition beyond the mean.

Let y_i be an outcome of interest and let d_i be an indicator of group such as race and gender, $d_i = 0, 1$, for an observation i, $i = 1, \ldots, N$. The decomposition can be written as

$$E\left[y_{1i}|d_i=1\right]-E\left[y_{0i}|d_i=0\right] = \left\{E\left[y_{1i}|d_i=1\right]-E\left[y_{0i}|d_i=1\right]\right\} + \left\{E\left[y_{0i}|d_i=1\right]-E\left[y_{0i}|d_i=0\right]\right\}, \tag{1}$$

where the subscript indicates a potential outcome. We observe $y_i = d_i y_{1i} + (1 - d_i)y_{0i}$. The decomposition involves counterfactual expectation $E\left[y_{0i}|d_i=1\right]$, which expresses an expected outcome if an individual in one group ($d_i=1$) were treated as if in the other group ($d_i=0$).

By the law of iterated expectation, $E\left[y_{ji}|d_i=k\right] = E\left[E(y_{ji}|x_i, d_i = k)|d_i = k\right]$ for $j=0,1$ and $k = 0, 1$. Furthermore, under the conditional independence assumption, $E\left[y_{ji}|x_i, d_i=k\right] = E\left[y_{ji}|x_i\right]$. This assumption holds for the decomposition analysis since being a particular gender or race is obviously predetermined. Now the expectation conditional on x_i is assumed to be a parametric function of x_i with a parameter vector β_j: $E\left[y_{ji}|x_i\right] = F(x_i;\beta_j)$. For instance, the OLS decomposition specifies $F(x_i;\beta_j) = x_i'\beta_j$. In this case, the first curly bracket in the right-hand of the equation (1) is $E\left[y_{1i}|d_i=1\right] - E\left[y_{0i}|d_i=1\right] = E\left[x_i|d_i=1\right]'(\beta_1 - \beta_0)$, which is the difference due to different effects of observable characteristics. This term is referred to as a coefficient effect. It can also be interpreted as the average treatment effect on the treated under certain conditions.[1] The second bracket is $E\left[y_{0i}|d_i=1\right] - E\left[y_{0i}|d_i=0\right] = \left(E[x_i|d_i=1] - E[x_i|d_i=0]\right)'\beta_0$, which is the difference due to differences in the characteristics, which is referred to as an endowment effect. When the reference group is switched to $d_i=0$, the difference can alternatively be decomposed as

$$E\left[y_{1i}|d_i=1\right] - E\left[y_{0i}|d_i=0\right] = \left\{E\left[y_{1i}|d_i=0\right] - E\left[y_{0i}|d_i=0\right]\right\} + \left\{E\left[y_{1i}|d_i=1\right] - E\left[y_{1i}|d_i=0\right]\right\},$$

of which the first and second curly brackets measure the coefficient and endowment effects, respectively.

Several nonlinear decomposition models are proposed: for example, probit and logit models (Fairle, 2006), Tobit model (Bauer and Sinning, 2010), and count data models (Bauer et al., 2007). Bauer and Sinning (2008) also discuss other nonlinear models. For example, for the probit model, $F(x_i;\beta_j) = \Phi(x_i'\beta_j)$, where $\Phi(\cdot)$ is the cdf of standard normal. For the Tobit model with the outcome left-censored at 0, the conditional expectation function is $x_i'\beta_j\Phi(x_i'\beta_j) + \sigma_j\phi(x_i'\beta_j)$, where $\phi(\cdot)$ is the pdf of standard normal. Note that the conditional expectation involves the parameter σ_j, the standard deviation

[1] See Kline (2011) for the conditions under which this term has a causal interpretation.

of a disturbance term, in addition to the coefficient vector β_i. For count data models, it is occasionally necessary to solve the problem of excess zeros using zero-inflated or hurdle models. The conditional expectation function of the hurdle negative binomial model is $\exp(x_i'\beta_j)/\{(1-(1+\alpha_j\exp(x_i'\beta_j))^{-1/\alpha_j})(1+\exp(z_i'\gamma_j))\}$, where α_j is the dispersion parameter. The regressors x_i affect positive counts of outcome while z_i governs the probability that zero counts occur. These regressors may or may not be identical. Hereafter, we generalize the notation of conditional expectation functions to $F(w_i;\theta_j)$, where w_i is a vector of all regressors and θ_j is a vector of all parameters. See Appendix B for the specified functional forms of $F(w_i;\theta_j)$ for the models considered in this paper.

As shown in the equation (1), the decomposition effects can be expressed as linear combinations of the conditional expectations. Let μ be a vector with four elements, each of which is defined as follows:

$$\mu = \begin{pmatrix} \mu_{11} \\ \mu_{01} \\ \mu_{10} \\ \mu_{00} \end{pmatrix} = \begin{pmatrix} E\left[y_{1i}|d_i=1\right] \\ E\left[y_{0i}|d_i=1\right] \\ E\left[y_{1i}|d_i=0\right] \\ E\left[y_{0i}|d_i=0\right] \end{pmatrix} = \begin{pmatrix} E\left[F(w_i;\theta_1)|d_i=1\right] \\ E\left[F(w_i;\theta_0)|d_i=1\right] \\ E\left[F(w_i;\theta_1)|d_i=0\right] \\ E\left[F(w_i;\theta_0)|d_i=0\right] \end{pmatrix} \tag{2}$$

Then, the coefficient effect is $\mu_{11}-\mu_{01}$. In matrix notation, it can be written as $R_c\mu$, where $R_c=(1,-1,0,0)$. Given the variance of $\widehat{\mu}$, $V(\widehat{\mu})$, the variance of the coefficient effect is $R_cV(\widehat{\mu})R_c'$. Likewise, the variance of the endowment effect, $\mu_{01}-\mu_{00}$, is computed as $R_eV(\widehat{\mu})R_e'$ by setting $R_e=(0,1,0,-1)'$, and the variances of those effects with switched references can also be obtained by modifying R_c and R_e. Therefore, estimating the variance of the decomposition effects is reduced to the estimation of variance of $\widehat{\mu}$.

The estimation of μ is straightforward. We estimate θ_j from relevant samples and compute the conditional expectation functions with relevant estimates and samples. However,

estimating its variance is not as straightforward. Oaxaca and Ransom (1998) discuss the delta method approach under the implicit assumption of fixed regressors. This approach accounts for the variability of $\widehat{\theta}_j$. When regressors x_i are stochastic, however, it is inappropriate. To think of this point concretely, consider the coefficient effect in the linear decomposition. As shown above, it is $E[x_i|d_i=1]'(\beta_1-\beta_0)$. In order to estimate this effect, we need to estimate $E[x_i|d_i=1]$ as well as β_0 and β_1. While the delta method approach takes into account the fact that β_0 and β_1 are estimated, it does not take into account the fact that $E[x_i|d_i=1]$ is estimated.

Jann (2008) discusses the variance estimator that accounts for the variability of regressors for the linear decomposition. However, as shown below, his approach cannot extend to the nonlinear decompositions directly. Kline (2014) also derives the variance of the linear decomposition, but only the coefficient effect. For the nonlinear decompositions, the bootstrap inference is an option for applied researchers as Fortin et al. (2011) suggest.[2] Although it is useful, the bootstrap approach is computationally intensive, especially when a model is highly nonlinear and/or a sample size is large. An analytical variance estimator is computationally much less intensive. Moreover, an analytical variance estimator is essential in order to obtain asymptotic refinement through the bootstrap methods (Horowitz, 2001).

In the next section, we describe the estimation of μ and derive its asymptotic variance. The derivation starts by noting that the estimation of μ involves sequential steps. Then, we follow the derivation of the variance of sequential m-estimation. The framework of the m-estimation enables us to extend to the cluster-robust variance easily.

[2] For example, Bauer et al. (2007) and Bauer and Sinning (2008, 2010) report the bootstrap standard errors.

3 Estimation

3.1 Estimation of μ

The estimation of μ involves two steps sequentially. The first step estimates the parameters of the conditional expectation functions, $\theta = (\theta_1', \theta_0')'$ by OLS or MLE (or generally m-estimation). At the second step, we estimate μ using the estimated parameters $\widehat{\theta}$. When deriving the asymptotic variance of $\widehat{\mu}$, it is necessary to take into account the fact that θ is estimated at the first step.

At the first step, θ is estimated by solving the equations:

$$N^{-1} \sum_{i=1}^{N} h_{\theta i}(\theta) = 0, \tag{3}$$

where $h_{\theta i}(\theta)$ is a vector defined as

$$h_{\theta i}(\theta) = \begin{pmatrix} d_i s_1(y_i, w_i; \theta_1)/\tau_1 \\ \\ (1 - d_i) s_0(y_i, w_i; \theta_0)/\tau_0 \end{pmatrix},$$

where τ_1 and τ_0 are $\Pr(d_i = 1)$ and $\Pr(d_i = 0)$, respectively.[3] For OLS, $s_j(y_i, w_i, \theta_j) = x_i(y_i - x_i'\beta_j)$, and for MLE, $s_j(y_i, w_i, \theta_j) = \partial \ln L_i(\theta_j)/\partial \theta_j$, where $\ln L_i(\theta_j)$ is the contribution to the log likelihood by an observation i. The equation (3) is the first order conditions for an m-estimation of θ.[4] This step is equivalent to estimating θ_j with a corresponding sample separately.

[3] Precisely speaking, the probabilities τ_1 and τ_0 are also estimated by $\widehat{\tau}_1 = N^{-1} \sum_{i=1}^{N} d_i$ and $\widehat{\tau}_0 = N^{-1} \sum_{i=1}^{N} (1 - d_i)$. However, it is not necessary to account for the variability from the estimation of these probabilities. See Appendix A for details.

[4] Clearly, we can interpret our estimation procedure as an estimating equation estimator.

The second step estimates μ by solving the sample counterpart of the equation (2):

$$N^{-1}\sum_{i=1}^{N} h_{\mu i}(\mu, \widehat{\theta}) = 0,$$

where $h_{\mu i}(\mu, \theta)$ is defined as

$$h_{\mu i}(\mu, \theta) = \begin{pmatrix} d_i(F(w_i;\theta_1) - \mu_{11})/\tau_1 \\ d_i(F(w_i,\theta_0) - \mu_{01})/\tau_1 \\ (1-d_i)(F(w_i;\theta_1) - \mu_{10})/\tau_0 \\ (1-d_i)(F(w_i;\theta_0) - \mu_{00})/\tau_0 \end{pmatrix}.$$

It is equivalent to computing the conditional expectations with relevant samples and estimated parameters. The fact that θ is estimated at the first step needs to be taken into account in deriving the asymptotic variance of $\widehat{\mu}$.

Proposition 1 *Under the regular conditions, we have the asymptotic distribution of* $\widehat{\mu}$ *as follows:*

$$\sqrt{N}(\widehat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, V(\widehat{\mu})).$$

The asymptotic variance $V(\widehat{\mu})$ *is*

$$V(\widehat{\mu}) = S_{\mu\mu} + G_{\mu\theta} V(\widehat{\theta}) G_{\mu\theta}', \tag{4}$$

where $V(\widehat{\theta})$ is the asymptotic variance of $\widehat{\theta}$, and

$$G_{\mu\theta} = \lim N^{-1} \sum_{i=1}^{N} E\left[\partial h_{\mu i}(\mu, \theta)/\partial \theta'\right]$$

$$= \lim N^{-1} \sum_{i=1}^{N} E \begin{bmatrix} d_i(\partial F(w_i;\theta_1)/\partial \theta_1')/\tau_1 & 0 \\ 0 & d_i(\partial F(w_i;\theta_0)/\partial \theta_0')/\tau_1 \\ (1-d_i)(\partial F(w_i;\theta_1)/\partial \theta_1')/\tau_0 & 0 \\ 0 & (1-d_i)(\partial F(w_i;\theta_0)/\partial \theta_0')/\tau_0 \end{bmatrix}$$

and

$$S_{\mu\mu} = \lim N^{-1} \sum_{i=1}^{N} E\left[h_{\mu i}(\mu, \theta) h_{\mu i}(\mu, \theta)'\right].$$

See Appendix A for the proof.

The term $G_{\mu\theta} V(\widehat{\theta}) G_{\mu\theta}{}'$ corresponds to the delta method approach. The matrix $S_{\mu\mu}$ represents variability due to variation in w_i, which would arise even if the true value of θ were known. Since $S_{\mu\mu}$ is a positive definite matrix, the variance estimator based on the delta method approach understates the true variance. Put it differently, ignoring the variation in w_i results in underestimation of $V(\widehat{\mu})$.

We estimate $V(\widehat{\mu})$ by

$$\widehat{V}(\widehat{\mu}) = N^{-1} \sum_{i=1}^{N} \widehat{h}_{\mu i} \widehat{h}_{\mu i}{}' + \left(N^{-1} \sum_{i=1}^{N} \partial h_{\mu i}/\partial \theta'|_{\theta=\widehat{\theta}} \right) \widehat{V}(\widehat{\theta}) \left(N^{-1} \sum_{i=1}^{N} \partial h_{\mu i}/\partial \theta'|_{\theta=\widehat{\theta}} \right)',$$

where $\widehat{h}_{\mu i} = h_{\mu i}(\widehat{\theta})$.

One of the concerns regarding statistical inference that empirical researchers often face is correlation within cluster. It is well-known that ignoring clustering tends to underestimate standard errors, and consequently it leads to overstate statistical significance (Moulton, 1986). A panel data structure necessitates the clustering correlation as well (Arrelano, 1987; Liang and Zeger, 1988). An advantage of seeing the estimation of μ as m-estimation is

the extension to the cluster robust estimator is straightforward by following Cameron et al. (2011), which even makes it possible to control for multiway clustering. For example, suppose that G denotes the number of clusters, and there are N_g observations within each cluster: $N = \sum_g^G N_g$. Then, the one-way clustering robust variance estimator is

$$\begin{aligned}\widehat{V}_c(\widehat{\mu}) =& N^{-1} \sum_{g=1}^{G} \left(\sum_{i=1}^{N_g} \widehat{h}_{\mu i} \right) \left(\sum_{i=1}^{N_g} \widehat{h}_{\mu i} \right)' + \\ & \left(N^{-1} \sum_{g=1}^{G} \sum_{i=1}^{N_g} \partial h_{\mu i} / \partial \theta' |_{\theta=\widehat{\theta}} \right) \widehat{V}_c(\widehat{\theta}) \left(N^{-1} \sum_{g=1}^{G} \sum_{i=1}^{N_g} \partial h_{\mu i} / \partial \theta' |_{\theta=\widehat{\theta}} \right)'\end{aligned}$$

where $\widehat{V}_c(\widehat{\theta})$ is the cluster robust variance of θ.[5]

3.2 Alternative Estimation

This subsection discusses an alternative to μ and its asymptotic variance. As mentioned previously, Jann (2008) derives the standard errors of the linear decomposition allowing stochastic regressors. The derivation is based on the fact that the estimation of μ_{jk} is the product of two vectors for the linear decomposition. Suppose $\overline{x}_k$ is the vector of sample means of the regressors of the group of $d_i = k$, that is, an estimate of $E[x_i | d_i = k]$, and $\widehat{\beta}_j$ is the estimated coefficient vector of the group of $d_i = j$. Then, $\widehat{\mu}_{jk} = \overline{x}_k{}' \widehat{\beta}_j$, and the variance of $\widehat{\mu}_{jk}$ is $\overline{x}_k' V(\widehat{\beta}_k) \overline{x}_k + \widehat{\beta}_j' V(\overline{x}_k) \widehat{\beta}_j + \texttt{trace}(V(\overline{x}_k) V(\widehat{\beta}_j))$ although the last term asymptotically vanishes (Jann, 2008). However, this derivation cannot be extended to nonlinear models since $\widehat{\mu}_{jk}$ is not the function of the product of the two vectors, $\overline{w}_i$ and $\widehat{\theta}_j$, but the average of $F(w_i; \widehat{\theta}_j)$ over the sample with $d_i = k$.

[5] As a finite-sample adjustment, we multiply $\widehat{V}_c(\widehat{\mu})$ by $G/(G-1)$ in estimating the variances in the following sections.

To explore more, define

$$\widetilde{\mu} = \begin{pmatrix} \widetilde{\mu}_{11} \\ \widetilde{\mu}_{01} \\ \widetilde{\mu}_{10} \\ \widetilde{\mu}_{00} \end{pmatrix} = \begin{pmatrix} F(\mu_{w_1}; \theta_1) \\ F(\mu_{w_1}; \theta_0) \\ F(\mu_{w_0}; \theta_1) \\ F(\mu_{w_0}; \theta_0), \end{pmatrix}$$

where $\mu_{w_j} = E[w_i | d_i = j]$. Obviously, $\widetilde{\mu}$ is not identical to μ unless $F(\cdot)$ is a linear function. We estimate $\widetilde{\mu}_{jk}$ by $F(\overline{w}_k; \widehat{\theta}_j)$. As $\widehat{\theta}$ and $\overline{w} = (\overline{w}_1', \overline{w}_0')'$ are asymptotically independent, the asymptotic variance of $\widehat{\widetilde{\mu}}$ is

$$V(\widehat{\widetilde{\mu}}) = \widetilde{G}_{\widetilde{\mu}\mu_w} V(\overline{w}) \widetilde{G}_{\widetilde{\mu}\mu_w}' + \widetilde{G}_{\widetilde{\mu}\theta} V(\widehat{\theta}) \widetilde{G}_{\widetilde{\mu}\theta}'$$

where $\widetilde{G}_{\widetilde{\mu}\mu_w} = \partial\widetilde{\mu}/\partial\mu_w$ and $\widetilde{G}_{\widetilde{\mu}\theta} = \partial\widetilde{\mu}/\partial\theta$. This expression has parallel structure to $V(\widehat{\mu})$. The first term arises from the fact the regressors are stochastic while the second term expresses the variability that comes from $\widehat{\theta}$. However, for the nonlinear models, the asymptotic distributions of $\widehat{\mu}$ and $\widehat{\widetilde{\mu}}$ are not equivalent. Thus, it is not appropriate to use $V(\widehat{\widetilde{\mu}})$ to make an inference on μ. For comparison, we estimate $V(\widehat{\widetilde{\mu}})$ in addition to $V(\widehat{\mu})$ in the Monte Carlo simulations and the real-data application in the following sections.

4 Monte Carlo Simulation

In this section we conduct the Monte Carlo simulation study to see the performance of the variance estimators discussed in the preceding sections. Besides the linear decomposition based on OLS, we consider several nonlinear models: Specifically, probit and logit for a binary outcome, Tobit for a censored outcome, and Poisson and Negative Binomial (NB) for a count data outcome. For the OLS, probit, and Tobit models, the data generating process

(DGP) is $y_{ji}^* = \beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \varepsilon_{ji}$, where $\varepsilon_{ji} \sim \mathcal{N}(0,1)$. Then, $y_{ji} = y_{ji}^*$ for the OLS, $y_{ji} = 1(y_{ji}^* > 0)$ for probit, and $y_{ji} = y_{ji}^* \times 1(y_{ji}^* > 0)$ for Tobit, where $1(\cdot)$ is an indicator function. For the logit model, $y_{ji} = 1(\beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \varepsilon_{ji} > 0)$, where ε_{ji} follows the logistic distribution. For the Poisson and Negative Binomial models, y_{ji} is generated to have a mean equal to $\exp(\beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i})$. For the Negative Binomial model, the dispersion parameter α_j is set to 0.5. We set the coefficients and DGPs of regressors the same for each group d_i: $\beta_{0j} = 0.5$, $\beta_{1j} = 1$, and $\beta_{2j} = -0.5$ for $j = 0, 1$, and two regressors are distributed as $x_{1i} \sim \mathcal{N}(0,1)$ and $x_{2i} \sim \chi^2_{10}$, the latter of which is subtracted by 10 and divided by $\sqrt{20}$ to have mean 0 and variance 1. Hence, the true values of coefficient and endowment effects are zero for all experiments. This is for a practical reason that we cannot compute the true values analytically unless the DGPs are the same across groups for the nonlinear models.

We consider the sample sizes $N = 1,000$ and $5,000$. Instead of fixing the size of each group, we allow it to vary at each replication. Specifically, the value of d_i is assigned as $d_i = 1(u_i + \nu_i > d^*)$, where $u_i \sim \text{uniform}(0,1)$ and $\nu_i \sim \mathcal{N}(0, 0.01)$. The latter adds slightly more variability to d_i. The threshold value d^* controls the proportional size of each group. That is, d^* and $(1 - d^*)$ are the probabilities of $d_i = 0$ and $d_i = 1$, respectively. We consider five values of d^*: $d^* \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. We run 10,000 replications for each setting.

Table 1 reports the benchmark results when $d^* = 0.5$. Columns (1) and (2) report the standard deviations of Monte Carlo replicates of decomposition effects based on μ and $\widetilde{\mu}$, respectively. As expected, the sampling distributions differ between μ and $\widetilde{\mu}$ for the nonlinear models. Especially, the endowment effects of Poisson and NB models show large differences. The endowment effects based on μ, i.e., $R_e\mu$, exhibit twice as large variation as that on $\widetilde{\mu}$, i.e., $R_e\widetilde{\mu}$, and these differences do not vanish even with larger samples.

Columns (3)-(5) are the averages of estimated standard errors over 10,000 replicates. By comparing column (1) and column (3), we can see that the averages of our proposed standard errors that account for the variability of regressors are very close to the Monte Carlo

Table 1: The Monte Carlo Results: Benchmark

		Std. Dev.[a]		Average of Std. Err.[b]			Rejection Probability [c]		
		μ	$\widetilde{\mu}$	s.e.$(\widehat{\mu})$	s.e.$(\widehat{\mu})_d$	s.e.$(\widehat{\widetilde{\mu}})$	s.e.$(\widehat{\mu})$	s.e.$(\widehat{\mu})_d$	s.e.$(\widehat{\widetilde{\mu}})$
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$N = 1,000$									
OLS	E	0.0707	0.0707	0.0708	0.0035	0.0708	0.0491	0.9492	0.0491
	C	0.0630	0.0630	0.0633	0.0632	0.0633	0.0497	0.0501	0.0497
Probit	E	0.0195	0.0287	0.0195	0.0014	0.0286	0.0509	0.8988	0.0040
	C	0.0248	0.0375	0.0249	0.0249	0.0376	0.0506	0.0513	0.0024
Logit	E	0.0145	0.0182	0.0144	0.0016	0.0180	0.0497	0.8683	0.0135
	C	0.0283	0.0357	0.0283	0.0282	0.0357	0.0520	0.0523	0.0153
Tobit	E	0.0381	0.0354	0.0380	0.0026	0.0355	0.0504	0.9323	0.0690
	C	0.0394	0.0407	0.0407	0.0406	0.0459	0.0448	0.0456	0.0222
Poisson	E	0.1734	0.0707	0.1719	0.0108	0.0708	0.0439	0.9411	0.4201
	C	0.0864	0.0703	0.0866	0.0856	0.0701	0.0497	0.0531	0.1101
NB	E	0.1762	0.0707	0.1744	0.0213	0.0708	0.0397	0.9059	0.4239
	C	0.1599	0.0892	0.1594	0.1575	0.0880	0.0449	0.0522	0.2719
$N = 5,000$									
OLS	E	0.0315	0.0315	0.0316	0.0007	0.0316	0.0484	0.9783	0.0484
	C	0.0283	0.0283	0.0283	0.0283	0.0283	0.0504	0.0504	0.0504
Probit	E	0.0086	0.0126	0.0087	0.0003	0.0126	0.0491	0.9511	0.0038
	C	0.0111	0.0167	0.0111	0.0111	0.0167	0.0498	0.0499	0.0031
Logit	E	0.0064	0.0080	0.0064	0.0003	0.0079	0.0490	0.9471	0.0157
	C	0.0127	0.0159	0.0126	0.0126	0.0159	0.0487	0.0488	0.0158
Tobit	E	0.0169	0.0157	0.0169	0.0005	0.0158	0.0493	0.9705	0.0665
	C	0.0176	0.0181	0.0182	0.0182	0.0206	0.0416	0.0419	0.0218
Poisson	E	0.0774	0.0313	0.0773	0.0022	0.0316	0.0509	0.9705	0.4182
	C	0.0385	0.0316	0.0384	0.0383	0.0313	0.0491	0.0496	0.1123
NB	E	0.0780	0.0315	0.0776	0.0044	0.0316	0.0490	0.9581	0.4217
	C	0.0705	0.0395	0.0706	0.0704	0.0394	0.0471	0.0483	0.2712

[a] The standard deviations of 10,000 replicates. Endowment (E) and Coefficient (C) effects are computed with $d_i = 1$ as a reference group. That is, E=$R_e\mu$ and C=$R_c\mu$ (column (1)) and E=$R_e\widetilde{\mu}$ and C=$R_c\widetilde{\mu}$ (column (2)), where $R_e = (0, 1, 0, -1)$ and $R_c = (1, -1, 0, 0)$. The results with $d_i = 0$ as a reference group are similar and omitted here. The results are available upon request.

[b] The averages of 10,000 replicates of estimated standard errors. The standard errors are s.e.$(\widehat{\mu}) = \sqrt{R.\widehat{V}(\widehat{\mu})R.'/N}$, s.e.$(\widehat{\mu})_d = \sqrt{R.(\widehat{G}_{\mu\theta}\widehat{V}(\widehat{\theta})\widehat{G}_{\mu\theta}')R.'/N}$, and s.e.$(\widehat{\widetilde{\mu}}) = \sqrt{R.\widehat{V}(\widehat{\widetilde{\mu}})R.'/N}$ with $R. = R_e$ or R_c correspondingly.

[c] The relative frequencies that the null hypothesis $R.\mu = 0$ is rejected at the 5% significance level. The test statistics are calculated with the corresponding standard errors.

standard deviations. On the other hand, the delta method standard errors, which is shown in column (4), of the endowment effect (E) are considerably smaller than the Monte Carlo standard deviations as a result of not considering the variability of regressors. However, note that the delta method standard errors of the coefficient effect (C) are comparable with the Monte Carlo standard deviations. To see why, consider the OLS model. As clear from equation (4), the difference between our proposed variance estimator and the delta method variance estimator arises due to $S_{\mu\mu}$. It can be shown that for the endowment effect, $R_e S_{\mu\mu} R_e{}' = \beta_0{}'\text{Var}(x_1)\beta_0 + \beta_0{}'\text{Var}(x_0)\beta_0$, where $\text{Var}(x_j)$ is the variance of x_i conditional on $d_i = j$. Since these two terms are positive, our proposed variance will always be larger than the delta method variance. On the other hand, for the coefficient effect, $R_c S_{\mu\mu} R_c{}' = \beta_1{}'\text{Var}(x_1)\beta_1 + \beta_0{}'\text{Var}(x_1)\beta_0 - 2\beta_1{}'\text{Var}(x_1)\beta_0$. When $\beta_1 = \beta_0$, the last term completely cancel out the first two terms. This implies that when $\widehat{\beta}_1$ and $\widehat{\beta}_0$ are close to each other, our proposed variance estimator will also get close to the delta method variance estimator. The same argument applies to the nonlinear models.

Columns (6)-(8) report the relative frequencies of the rejection of the null hypothesis that E=$R_e\mu$=0 or C=$R_c\mu = 0$ at the 5% significance level. Since the true values of E and C are zero in our setting, the relative frequencies measure a size of the test. As column (6) shows, there are little size distortions when the test statistics are computed with our proposed variance estimator except for the tests for the Poisson and NB models, which are under-sized when the sample size is small. Column (7) shows that the delta method variance leads to severe size distortions when the endowment effects are examined. Column (8) shows that computing the test statistics based on $V(\widetilde{\mu})$ results in severe size distortions. It is because that the variance $V(\widetilde{\mu})$ does not represent the variability of μ properly.[6]

Table 2 reports the rejection probabilities that based on $\widehat{V}(\widehat{\mu})$ for different values of d^*.

[6] Although not reported here, when the null hypothesis is $R_e\widetilde{\mu}$=0 or $R_c\widetilde{\mu} = 0$, the tests based on $V(\widetilde{\mu})$ perform well.

Table 2: Rejection Probabilities with Different Threshold Value d^*

		Threshold Value d^*				
		0.5	0.4	0.3	0.2	0.1
		(1)	(2)	(3)	(4)	(5)
$N = 1,000$						
OLS	E	0.0491	0.0462	0.0500	0.0492	0.0507
	C	0.0497	0.0489	0.0518	0.0524	0.0524
Probit	E	0.0509	0.0465	0.0505	0.0504	0.0529
	C	0.0506	0.0504	0.0526	0.0527	0.0545
Logit	E	0.0497	0.0484	0.0456	0.0492	0.0499
	C	0.0520	0.0516	0.0511	0.0502	0.0514
Tobit	E	0.0504	0.0490	0.0507	0.0514	0.0557
	C	0.0448	0.0428	0.0409	0.0420	0.0459
Poisson	E	0.0439	0.0471	0.0518	0.0609	0.0753
	C	0.0497	0.0483	0.0487	0.0482	0.0533
NB	E	0.0397	0.0443	0.0459	0.0584	0.0734
	C	0.0449	0.0441	0.0454	0.0447	0.0461
$N = 5,000$						
OLS	E	0.0484	0.0513	0.0494	0.0496	0.0511
	C	0.0504	0.0512	0.0502	0.0499	0.0524
Probit	E	0.0491	0.0513	0.0495	0.0485	0.0495
	C	0.0498	0.0491	0.0490	0.0493	0.0505
Logit	E	0.0490	0.0528	0.0550	0.0531	0.0515
	C	0.0487	0.0496	0.0491	0.0498	0.0460
Tobit	E	0.0493	0.0512	0.0499	0.0533	0.0483
	C	0.0416	0.0451	0.0465	0.0457	0.0481
Poisson	E	0.0509	0.0482	0.0480	0.0501	0.0576
	C	0.0491	0.0504	0.0516	0.0554	0.0529
NB	E	0.0490	0.0464	0.0498	0.0526	0.0567
	C	0.0471	0.0485	0.0471	0.0494	0.0474

Note: See the notes in 1.

As d^* gets smaller and smaller, the sizes become more and more distorted, in particular, for the Poisson and NB models. The size distortions tend to be less severe when $N = 5,000$ than when $N = 1,000$.

In addition to the case where each observation is independent of one another, we also consider the case where observations are correlated with clusters. To do so, we restrict the regressor x_2 to vary only at a cluster level and add cluster-specific error components. More

Table 3: The Monte Carlo Results: Clustering

		Std. Dev.	Average of Std. Err.		Rejection Probability	
		μ	no clustering	clustering	no clustering	clustering
		(1)	(2)	(3)	(4)	(5)
OLS						
$G = 100$	E	0.0403	0.0354	0.0403	0.0858	0.0496
	C	0.0425	0.0315	0.0414	0.1424	0.0523
$G = 50$	E	0.0576	0.0501	0.0570	0.0872	0.0557
	C	0.0600	0.0443	0.0574	0.1464	0.0614
$G = 25$	E	0.0823	0.0712	0.0808	0.0864	0.0569
	C	0.0831	0.0620	0.0784	0.1455	0.0699
Probit						
$G = 100$	E	0.0110	0.0097	0.0110	0.0857	0.0528
	C	0.0145	0.0124	0.0144	0.0931	0.0546
$G = 50$	E	0.0155	0.0138	0.0156	0.0802	0.0513
	C	0.0205	0.0175	0.0201	0.0957	0.0574
$G = 25$	E	0.0222	0.0197	0.0221	0.0797	0.0551
	C	0.0290	0.0247	0.0275	0.0971	0.0716
Tobit						
$G = 100$	E	0.0215	0.0197	0.0215	0.0703	0.0499
	C	0.0128	0.0104	0.0127	0.1094	0.0502
$G = 50$	E	0.0307	0.0279	0.0303	0.0764	0.0583
	C	0.0180	0.0146	0.0175	0.1098	0.0582
$G = 25$	E	0.0438	0.0395	0.0426	0.0740	0.0610
	C	0.0249	0.0205	0.0239	0.1077	0.0621

Note: See the notes in 1. This table report the results based on μ.

specifically, we consider the following DGP: $y_i^* = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2g} + \sqrt{0.5}(\varepsilon_i + \varepsilon_g)$, where $\varepsilon_i \sim \mathcal{N}(0,1)$ and $\varepsilon_g \sim \mathcal{N}(0,1)$ for $g = 1, \ldots, G$. In this exercise, we consider only OLS, probit, and Tobit models. As before, $y_{ji} = y_{ji}^*$ for the OLS, $y_{ji} = 1(y_{ji}^* > 0)$ for probit, and $y_{ji} = y_{ji}^* \times 1(y_{ji}^* > 0)$ for Tobit, and there is no difference in DGP by group. Besides, we also add the clustering correlation to the group assignment process: $d_i = 1(u_i + \nu_g > d_*)$, where $\nu_g \sim \mathcal{N}(0, 0.01)$, following Kline (2014). We set the number of clusters as $G = 25, 50$, and 100. There are 40 observations in each cluster. That is, $N_g = 40$ and $N = 40 \times G$.

Table 3 reports the results. As columns (2) and (4) show, the standard errors that

ignore clustering underestimate the true variability of the decomposition effects and lead to over-rejection of the true null hypothesis. When the clustering correlation is taken into account, our proposed variance estimator is able to estimate the true variability. However, when the number of clusters is smaller, the cluster-robust variance estimator becomes biased downward. This result is consistent with the findings in the previous literature. It is because the cluster-robust variance builds on the assumption that G goes to infinity. However, as the analytical form of the cluster-robust variance estimator is presented, it is feasible to correct bias by asymptotic refinement using the bootstrap method (Cameron et al., 2008).

5 Real-Data Application

This section illustrates a real-data application. The data set used for this application is from the RAND Health Insurance Experiment (HIE).[7] We decompose the gender differences of various outcomes: an observation i is female if $d_i = 1$. Specifically, the outcomes of interest are annual individual health expenditures, a binary choice of whether the expenditure is positive, and the number of outpatient visits. The expenditures in logarithms given positive expenditures are decomposed by OLS, and the binary choice is decomposed by the probit and logit models. The Tobit model is also used for the annual expenditure (not in log) to capture zero expenditures. The number of outpatient visits is count data, and thus, we employ the Poisson and Negative Binomial (NB) model. Furthermore, we also consider the hurdle Poisson and NB models and zero-inflated Poisson and NB models since zero outpatient visits account for a considerable portion of the sample (around 31%). For the process governing zero counts, we use the logit model with the same regressors as the process for positive counts. That is, $x_i = z_i$. We use the same regressors for all the outcomes. See Appendix C for the definitions and summary statistics of these variables.

[7] The data extract is downloaded from `http://cameron.econ.ucdavis.edu/mmabook/mmaprograms.html`.

Table 4: Real-Data Application: HIE data

			Benchmark[a]				Clustering[b]	
		Est.	s.e.$(\widehat{\mu})$	s.e.$(\widehat{\mu})_d$	s.e.$(\widehat{\widehat{\mu}})$	s.e.$(\widehat{\mu})_b$[c]	s.e.$(\widehat{\mu})$	s.e.$(\widehat{\mu})_b$[c]
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
OLS	Endowment	0.151	0.011	0.007	0.011	0.011	0.019	0.019
	Coefficient	0.182	0.023	0.023	0.023	0.021	0.029	0.028
Probit	Endowment	0.010	0.002	0.001	0.003	0.002	0.004	0.004
	Coefficient	0.068	0.006	0.006	0.006	0.006	0.008	0.008
Logit	Endowment	0.010	0.002	0.001	0.003	0.002	0.004	0.004
	Coefficient	0.067	0.006	0.006	0.006	0.006	0.008	0.008
Tobit	Endowment	19.290	2.757	2.086	2.437	2.600	4.402	4.581
	Coefficient	6.023	48.808	48.804	47.504	51.344	49.678	49.455
Poisson	Endowment	0.271	0.032	0.025	0.025	0.034	0.055	0.057
	Coefficient	0.559	0.060	0.060	0.057	0.062	0.091	0.091
NB	Endowment	0.288	0.033	0.024	0.025	0.034	0.056	0.054
	Coefficient	0.556	0.061	0.061	0.057	0.063	0.093	0.091
ZIP	Endowment	0.286	0.032	0.025	0.022	0.034	0.055	0.057
	Coefficient	0.544	0.060	0.060	0.073	0.062	0.091	0.090
ZINB	Endowment	0.288	0.032	0.024	0.024	0.033	0.054	0.054
	Coefficient	0.546	0.060	0.060	0.087	0.063	0.092	0.090
HP	Endowment	0.275	0.031	0.024	0.024	0.033	0.053	0.055
	Coefficient	0.533	0.056	0.055	0.054	0.058	0.085	0.084
HNB	Endowment	0.280	0.031	0.023	0.025	0.033	0.053	0.054
	Coefficient	0.541	0.061	0.060	0.061	0.062	0.092	0.090

[a] The benchmark standard errors treat all observations as independent.
[b] Clustering at an individual level. There are 5,908 unique individuals in the data.
[c] The bootstrap standard errors are based on 200 replications.
[d] A sample drawn at each replication is at the cluster (individual) level.

In the benchmark computations of standard errors, we assume that each observation is independent of one another. Besides, as the data have a panel structure, that is, multiple observations per individual, we also compute the standard errors controlling for clustering at an individual level. In addition to the various ways of estimating standard errors discussed above, we also compute the bootstrap standard errors for comparison.

Table 4 summarizes the results. Our proposed standard errors and bootstrap standard errors are comparable in all the models, so are they even in controlling for clustering. This fact

validates our proposed variance estimator since the bootstrap approach is widely accepted in the applied literature. Although the time elapsed to conduct the bootstrap resampling is not measured, it is quite time-consuming, especially, for highly nonlinear cases such as zero-inflated Poisson and NB models. Of course, the analytical standard errors are computationally less intensive. Computational easiness is valuable to applied researchers. However, as noted above, the bootstrap approach is still useful along with the analytical variance in order to obtain asymptotic refinement. As expected, the delta method approach underestimates the standard errors of the endowment effect compared to the proposed and bootstrap estimators by ignoring the variability of the regressors. We can also see that the standard errors based on $\widetilde{\mu}$ do not coincide with the bootstrap standard errors.

6 Conclusion

This paper derives the asymptotic variance of the decomposition effects that are applicable to both linear and nonlinear cases. Our proposed estimator is an useful alternative to the bootstrap approach, which is the mostly used variance estimator in the applied literature of the nonlinear decomposition. We confirm the validity of our proposed variance estimator with the Monte Carlo simulations and the real-data application.

Our derivation of the asymptotic variance is in general settings, employing the framework of m-estimation. It makes it easy to extend the variance estimator to control for clustering correlation. our approach is also straightforward for further extensions. This section briefly mentions to several possible extensions.

First, we illustrate the decomposition using OLS and MLE since the literature has exclusively used these estimation methods. Our approach is clearly applicable to a nonlinear least squares model since it is one of m-estimators. It is also possible to extend to the decomposition based on generalized method of moments (GMM). Therefore, we are able to accommodate a variety of models.

Second, the decomposition may have additional terms besides the endowment and coefficient effects described in the paper. For example, the "threefold" decomposition (Daymont and Andrisani, 1984) is often applied. In this case, the additional term is simply a combination of the elements of μ like the other two effects, and therefore, it is possible to estimate the variance in the same way as other two effects by setting R properly. Also, the decomposition may also involve the $F(\cdot)$ evaluated with the parameters other than θ_1 or θ_0. For example, the parameters are estimated from a pooled sample or a weighted average of θ_1 and θ_0. We are able to apply the proposed approach by modifying the moment conditions at the first step and/or the second step.

Third, while the previous sections cover aggregate decompositions, the decomposition analysis often determines the contribution of each regressor to the endowment and coefficient effects (the "detailed" decomposition). Because of its linearity, the estimation of the detailed decomposition and its variance is straightforward for the OLS decomposition. We can simply divide the conditional expectations in (2) into the contribution of each regressor. For the nonlinear decomposition, there is no unified approach for the detailed decomposition.[8] However, in principle, we can modify our approach so that we can estimate the asymptotic variance of the detailed decomposition for the nonlinear models.

The capability of these extensions values our proposed variance estimator further.

Acknowledgment

The author thanks Wim Vijverberg for his helpful comments.

[8] See, for example, Yun (2004) and Fairle (2006).

References

Arrelano, M., 1987. Computing robust standard errors for within-group estimators. Oxford Bulletin of Economics and Statistics 49 (4), 431–434.

Bauer, T., Ghlmann, S., Sinning, M., 2007. Gender differences in smoking behavior. Health Economics 16 (9), 895–909.

Bauer, T., Sinning, M., 2008. An extension of the Blinder-Oaxaca decomposition to nonlinear models. Advances in Statistical Analysis 92 (2), 197–206.

Bauer, T., Sinning, M., 2010. Blinder-Oaxaca decomposition for tobit models. Applied Economics 42 (12), 1569–1575.

Blinder, A. S., 1973. Wage discrimination: Reduced form and structural estimates. The Journal of Human Resources 8 (4), 436–455.

Cameron, A., Trivedi, P., 2005. Microeconometrics: Methods and Applications. Cambridge University Press.

Cameron, A. C., Gelbach, J. B., Miller, D. L., 2008. Bootstrap-based improvements for inference with clustered errors. The Review of Economics and Statistics 90 (3), 414–427.

Cameron, A. C., Gelbach, J. B., Miller, D. L., 2011. Robust inference with multiway clustering. Journal of Business & Economic Statistics 29 (2), 238–249.

Daymont, T. N., Andrisani, P. J., 1984. Job preferences, college major, and the gender gap in earning. Journal of Human Resources 19 (3), 408 – 428.

Fairle, R. W., 2006. An extension of the Blinder-Oaxaca decomposition technique to logit and probit models. Journal of Economic and Social Measurement 30 (4), 305–316.

Fortin, N., Lemieux, T., Firpo, S., 2011. Chapter 1 - Decomposition Methods in Economics. Vol. 4, Part A of Handbook of Labor Economics. Elsevier, pp. 1 – 102.

Horowitz, J. L., 2001. The Bootstrap. Vol. V. Elsevier Science B.V., Ch. 52, pp. 3159–3228.

Jann, B., 2008. The Blinder-Oaxaca decomposition for linear regression models. The Stata Journal 8 (4), 453–479.

Kline, P., 2011. Oaxaca-Blinder as a reweighting estimator. The American Economic Review 101 (3), 532–537.

Kline, P., 2014. A note on variance estimation for the Oaxaca estimator of average treatment effects. Economics Letters 122 (3), 428 – 431.

Liang, K.-Y., Zeger, S. L., 1988. Longitudianl data analysis using generalized linear models. Biometrika 73 (1), 13–22.

Moulton, B. R., 1986. Random group effects and the precision of regression estimates. Journal of Econometrics 32 (3), 385–397.

Murphy, K. M., Topel, R. H., 1985. Estimation and inference in two-step econometric models. Journal of Business & Economic Statistics 3 (4), 370–379.

Newey, W. K., 1984. A method of moments interpretation of sequential estimators. Economics Letters 14 (23), 201 – 206.

Newey, W. K., McFadden, D., 1994. Large Sample Estimation and Hypothesis Testing. Vol. IV. Elsevier Science B.V., Ch. 36, pp. 2111–2245.

Oaxaca, R. L., 1973. Male-female wage differentials in urban labor markets. International Economic Review 14 (3), 693–709.

Oaxaca, R. L., Ransom, M. R., 1998. Calculation of approximate variances for wage decomposition differentials. Journal of Economic and Social Measurement 24 (1), 55 – 61.

Pagan, A., 1986. Two stage and related estimators and their applications. Review of Economic Studies 53 (4), 517 – 538.

Yun, M.-S., 2004. Decomposing differences in the first moment. Economics Letters 82 (2), 275 – 280.

Appendices

A Proof of Proposition 1

The derivation of asymptotic variance of $\widehat{\mu}$ is based on the sequential two-step estimation by Newey (1984). Murphy and Topel (1985) and Pagan (1986) also derive similar results, and Newey and McFadden (1994) and Cameron and Trivedi (2005) illustrate the derivation in a clear fashion. Let $\delta = (\theta', \mu')'$. Then, δ can be estimated by solving the equations (3) and (3.1) simultaneously. The consistency of δ requires the population moment condition that $E(h_{\theta i}(\theta)', h_{\mu i}(\mu, \theta))' = 0$. Under the regular conditions, the asymptotic distribution is

$$\sqrt{N}(\widehat{\delta} - \delta) \xrightarrow{d} \mathcal{N}(0, G^{-1} S (G^{-1})'),$$

where

$$G = \lim N^{-1} \sum_{i=1}^{N} E \begin{pmatrix} \partial h_{\theta i}(\theta)/\partial \theta' & \partial h_{\theta i}(\theta)/\partial \mu' \\ \partial h_{\mu i}(\mu, \theta)/\partial \theta' & \partial h_{\mu i}(\mu, \theta)/\partial \mu' \end{pmatrix} = \begin{pmatrix} G_{\theta\theta} & G_{\theta\mu} \\ G_{\mu\theta} & G_{\theta\theta} \end{pmatrix}$$

and

$$S = \lim N^{-1} \sum_{i=1}^{N} E \begin{pmatrix} h_{\theta i}(\theta) h_{\theta i}(\theta)' & h_{\theta i}(\theta) h_{\mu i}(\theta)' \\ h_{\mu i}(\mu, \theta) h_{\theta i}(\theta)' & h_{\mu i}(\mu, \theta) h_{\mu i}(\mu, \theta)' \end{pmatrix} = \begin{pmatrix} S_{\theta\theta} & S_{\theta\mu} \\ S_{\mu\theta} & S_{\mu\mu} \end{pmatrix}$$

Since $E\left[\partial h_{\theta i}(\theta)/\partial \mu'\right] = G_{\theta\mu} = 0$, the inverse of G is

$$G^{-1} = \begin{pmatrix} G_{\theta\theta}^{-1} & 0 \\ -G_{\mu\mu}^{-1} G_{\mu\theta} G_{\theta\theta} & G_{\mu\mu}^{-1} \end{pmatrix}.$$

Therefore, we can obtain the asymptotic variances of $\widehat{\theta}$ and $\widehat{\mu}$:

$$V(\widehat{\theta}) = G_{\theta\theta}^{-1} S_{\theta\theta} G_{\theta\theta}^{-1}$$

and

$$V(\widehat{\mu}) = G_{\mu\mu}^{-1} \left\{ S_{\mu\mu} + G_{\mu\theta} G_{\theta\theta}^{-1} S_{\theta\theta} G_{\theta\theta}^{-1} G_{\mu\theta}{}' - G_{\mu\theta} G_{\theta\theta}^{-1} S_{\theta\mu} - S_{\mu\theta} G_{\theta\theta}^{-1} G_{\mu\theta}{}' \right\} G_{\mu\mu}^{-1}. \tag{A.1}$$

In our context, this expression can be simplified. First, $G_{\mu\mu}$ is simply a 4×4 identity matrix with a negative sign. Second, $S_{\theta\mu} = S_{\mu\theta}{}' = 0$. To see this, note that $E(h_{\theta i} h_{\mu i}{}') = E\left[E(h_{\theta i} h_{\mu i}{}' | w_i)\right] = E\left[E(h_{\theta i} | w_i) h_{\mu i}{}'\right]$ by the law of iterated expectation and $h_{\mu i}$ is a function of w_i. Provided that w_i is exogenous, $E(h_{\theta i} | w_i) = 0$, and thus $S_{\theta\mu} = \lim N^{-1} \sum_{i=1}^{N} E(h_{\theta i} h_{\mu i}{}') = 0$. The term $G_{\theta\theta}^{-1} S_{\theta\theta} G_{\theta\theta}^{-1}$ is the asymptotic variance of $\widehat{\theta}$, $V(\widehat{\theta})$. The simplification results in the equation (4). Under the assumption of homoskedasticity or the information matrix equality, $V(\widehat{\theta})$ can be simplified further. We do not make such assumptions in the Monte Carlo simulation and the real-data application in this paper.

The expression (A.1) shows that when $G_{\mu\theta} = \lim N^{-1} \sum_{i=1}^{N} E(\partial h_{\mu i} / \partial \theta) \neq 0$, as in the case of our study, it is necessary to account for the variability of $\widehat{\theta}$ in the second step. Looking at the opposite way, it reveals why we do not need to take the variability of $\widehat{\tau}$ in estimating θ and μ. It is easy to verify that $E(\partial h_{\theta i} / \partial \tau) = 0$ and $E(\partial h_{\mu i} / \partial \tau) = 0$, where $\tau = (\tau_1, \tau_0)'$. Therefore, the variability of $\widehat{\tau}$ does not influence the asymptotic variance of $\widehat{\theta}$ and $\widehat{\mu}$.

B Conditional Expectation Functions

C Variable Definitions and Summary Statistics

Table B.1: Functional Form

Model	$F(w;\theta)$
OLS	$x'\beta$
Probit	$\Phi(x'\beta)$
Logit	$\exp(x'\beta)/(1+\exp(x'\beta))$
Tobit	$x'\beta\Phi(x'\beta)+\sigma\phi(x'\beta)$
Poisson	$\exp(x'\beta)$
Negative Binomial (NB)	$\exp(x'\beta)$
Zero-inflated Poisson [a]	$\exp(x'\beta)/(1+\exp(z'\gamma))$
Zero-inflated NB [a]	$\exp(x'\beta)/(1+\exp(z'\gamma))$
Hurdle Poisson [a]	$\exp(x'\beta)/\{(1-\exp(-\exp(x'\beta)))(1+\exp(z'\gamma))\}$
Hurdle NB [a]	$\exp(x'\beta)/\{(1-(1+\alpha\exp(x'\beta))^{-1/\alpha})(1+\exp(z'\gamma))\}$

[a] The regime that leads to a zero outcome is specified by a logit model. That is, $\Pr(y_i=0|z_i)=\exp(z'\gamma)/(1+\exp(z'\gamma))$.

Table C.2: Health expenditure [a]

		MALE		FEMALE	
Number of Obs.		9,751		10,435	
Variables	Definition	Mean	Std. Dev.	Mean	Std. Dev.
MED	Annual medical expenditures in constant dollars excluding dental and outpatient mental	141.607	729.469	199.607	666.615
LNMED	log(MED)	3.928	1.445	4.262	1.501
DMED	1 if medical expenditures > 0	0.739	0.439	0.817	0.387
MDU	number of outpatient visits to a medical doctor	2.432	4.038	3.262	4.867
LC	ln(coinsurance+1) with $0 \leq$ rate ≤ 100	2.377	2.041	2.390	2.042
IDP	1 if individual deductible plan	0.255	0.436	0.265	0.441
LPI	log(annual participation incentive payment) or 0 if no payment	4.732	2.704	4.687	2.691
FMDE	log(medical deductible expenditure) if IDP=1 and MDE>1 or 0 otherwise.	4.043	3.490	4.019	3.454
PHYSLIM	1 if physical limitation	0.099	0.291	0.147	0.347
NDISEASE	number of chronic diseases	9.826	5.865	12.570	7.221
HLTHG	1 if good health	0.336	0.472	0.386	0.487
HLTHF	1 if fair health	0.067	0.250	0.087	0.282
HLTHP	1 if poor health	0.010	0.101	0.019	0.137
LINC	log of family income (in dollars)	8.761	1.195	8.659	1.256
LFAM	log of family size	1.276	0.530	1.223	0.546
EDUCDEC	education of household head (in years)	12.068	2.971	11.872	2.639
AGE	age	24.786	16.663	26.589	16.819
CHILD	1 if age is less than 18	0.430	0.495	0.375	0.484
BLACK	1 if black	0.167	0.371	0.196	0.393

[a] Source: Derived from the dataset used in Cameron and Trivedi (2005);

[b] The numbers of observations with nonzero MED are 7,210 for male and 8,523 for female, respectively.