A NOTE ON VARIANCE ESTIMATION FOR THE OAXACA ESTIMATOR OF
AVERAGE TREATMENT EFFECTS

Patrick M. Kline

A Note on Variance Estimation for the Oaxaca Estimator of Average Treatment Effects
Patrick M. Kline
NBER Working Paper No. 19784
January 2014
JEL No. C01,J31,J7

ABSTRACT

We derive the limiting distribution of the Oaxaca estimator of average treatment effects studied by
Kline (2011). A consistent estimator of the asymptotic variance is proposed that makes use of standard
regression routines. It is shown that ignoring uncertainty in group means will tend to lead to an overstatement
of the asymptotic standard errors. Monte Carlo experiments examine the finite sample performance
of competing approaches to inference.

Patrick M. Kline
Department of Economics
UC, Berkeley
508-1 Evans Hall #3880
Berkeley, CA 94720
and NBER
pkline@econ.berkeley.edu

1 Introduction

In a seminal contribution, Oaxaca (1973) proposed estimation of counterfactual means by applying the regression coefficients of one demographic group to the mean covariates of another.[1] Kline (2011) showed that a variant of Oaxaca's technique corresponds to a "doubly robust" estimator of average treatment effects – estimation is consistent if either mean untreated outcomes or the odds of treatment are linear in the covariates. An attractive feature of the Oaxaca estimator of treatment effects is that it relies on simple regression based methods that are easy to implement and adapt to the needs of a particular application. Several recent studies apply the Oaxaca estimator of treatment effects to program evaluation problems (e.g. Angrist and Rokkanen, 2012; Busso, Gregory, and Kline, 2013; Kline and Moretti, forthcoming), however no formal discussion of the estimator's asymptotic properties has yet been provided. This note derives the limiting distribution of the Oaxaca estimator of average treatment effects and proposes a simple computational approach to variance estimation. It is shown that ignoring the variability of the mean values of covariates will tend to lead to an overstatement of asymptotic standard errors.[2] Monte Carlo results are provided to illustrate the relative performance of competing approaches to inference.

2 Setup

The data are a triple $\{Y_i, X_i, D_i\}_{i=1}^N$ where Y_i is a scalar outcome, X_i a $K \times 1$ vector of covariates which is assumed to include an intercept, and D_i a scalar indicator for treatment status with one indicating treated. We assume throughout that all variables have finite second moments and that $E[X_i X_i'|D_i = 0]$ is full rank.

Our parameter of interest is:

$$\theta \equiv \mu_y^1 - \mu_x^{1\prime} \beta^0$$

where $\mu_y^1 \equiv E[Y_i|D_i = 1]$, $\mu_x^1 \equiv E[X_i|D_i = 1]$, and $\beta^0 \equiv E[X_i X_i'|D_i = 0]^{-1} E[X_i Y_i|D_i = 0]$. To relate this to Oaxaca (1973)'s original framework, note that if one takes Y_i is log wages and D_i as an indicator for being female then this quantity corresponds to the difference between women's actual mean wages and their mean predicted wage based upon the regression coefficients governing the population of men. Quantities such as θ have traditionally been used to detect labor market discrimination but can more generally be thought of as measures of average causal effects. Kline (2011) discusses the conditions under which θ identifies the average treatment effect on the treated – a standard estimand in program evaluation (Heckman and Robb, 1985).

The sample analogue of θ is:

$$\hat{\theta} \equiv \hat{\mu}_y^1 - \hat{\mu}_x^{1\prime} \hat{\beta}^0$$

where $\hat{\mu}_y^1 \equiv \frac{\frac{1}{N}\sum_i D_i Y_i}{\frac{1}{N}\sum_i D_i}$, $\hat{\mu}_x^1 \equiv \frac{\frac{1}{N}\sum_i D_i X_i}{\frac{1}{N}\sum_i D_i}$, $\hat{\beta}^0 \equiv \left(\frac{1}{N}\sum_i (1 - D_i) X_i X_i'\right)^{-1}\left(\frac{1}{N}\sum_i (1 - D_i) X_i Y_i\right)$. By the continuous mapping theorem, $\hat{\theta} \xrightarrow{p} \theta$.

We assume that any dependence across observations is weak enough that a central limit theorem holds:

$$\left(\hat{\mu}_y^1, \hat{\mu}_x^{1\prime}, \hat{\beta}^0\right)' \overset{a}{\sim} N\left(\left(\mu_y^1, \mu_x^{1\prime}, \beta^0\right)', V/N\right),$$

where $V \equiv \begin{bmatrix} V_y & & \\ V_{xy} & V_x & \\ V_{\beta y} & V_{\beta x} & V_\beta \end{bmatrix}$ for V_y a scalar, V_{xy} a $K \times 1$ vector of covariances, V_x a $K \times K$ variance matrix, $V_{\beta y}$ a $K \times 1$ vector of covariances, $V_{\beta x}$ a $K \times K$ covariance matrix, and V_β a $K \times K$ variance matrix. Hence, by the Delta method:

---

[1] A large subsequent literature on decomposition methods reviewed by Fortin, Lemieux, and Firpo (2011) extends Oaxaca's original contribution, focusing most recently on identification and estimation of distributional counterfactuals.

[2] In a related result, Wooldridge (2002, 2007) shows the conditions under which variance estimates ignoring sampling uncertainty in an estimated propensity score will tend to yield conservative inference. The present analysis considers the effects of ignoring sampling uncertainty in the mean value of covariates which corresponds to one component of the Oaxaca estimator's implicit propensity score estimate.

$$\sqrt{N}\left(\hat{\theta}-\theta\right)=\sqrt{N}\left[\left(\hat{\mu}_y^1-\mu_y^1\right)-\mu_x^{1\prime}\left(\hat{\beta}^0-\beta^0\right)-\left(\hat{\mu}_x^1-\mu_x^1\right)'\beta^0\right]+o_p\left(1\right)$$

Asymptotically, there are three sources of uncertainty: i) variability in the treated mean outcome $\left(\hat{\mu}_y^1-\mu_y^1\right)$, ii) variability in the regression coefficients among the untreated sample $\left(\hat{\beta}^0-\beta^0\right)$, and iii) variability in the covariate means in the treatment group $\left(\hat{\mu}_x^1-\mu_x^1\right)$. In general, these three sources of variability may be correlated, which gives rise to the following asymptotic variance expression:

$$\sqrt{N}\left(\hat{\theta}-\theta\right)\overset{d}{\to}N\left(0,V_\theta\right),$$

where $V_\theta\equiv V_y+\mu_x^{1\prime}V_\beta\mu_x^1+\beta^{0\prime}V_x\beta^0-2\mu_x^{1\prime}V_{\beta y}-2\beta^{0\prime}V_{xy}-2\beta^{0\prime}V_{\beta x}\mu_x^1$.

Because the estimator is asymptotically linear, bootstrap techniques can be applied (see Theorem 23.5 in Van Der Vaart, 2000). However, the bootstrap can be computationally expensive relative to analytical estimates. Furthermore, it is sometimes of interest to obtain analytical variance estimates for use in obtaining an asymptotic refinement via the bootstrap.

3 Ignoring uncertainty in the group means

It is common (e.g., Oaxaca and Ransom, 1994, 1998) for researchers to ignore the variability in the group means $\left(\hat{\mu}_x^1-\mu_x^1\right)$, which implies dropping the terms $\beta^{0\prime}V_x\beta^0-2\beta^{0\prime}V_{xy}-2\beta^{0\prime}V_{\beta x}\mu_x^1$ from V_θ.[3] We show now that this will tend to lead to an *over*-estimate of the asymptotic sampling error. To see this, note first that typically $V_{\beta x}=0$ because $\hat{\mu}_x^{1\prime}$ and $\hat{\beta}^0$ are estimated on separate samples that are independent of one another.[4] Suppose that this is the case. Define $\beta^1\equiv E\left[X_iX_i'|D_i=1\right]^{-1}E\left[X_iY_i|D_i=1\right]$ as the regression coefficient among the treated units. If the data are independent across observations, then we can write $V_{xy}=V_x\beta^1$.[5] Hence $\beta^{0\prime}V_{xy}=\beta^{0\prime}V_x\beta^1$. Thus, for $\beta^0\approx\beta^1$, we have $\beta^{0\prime}V_x\beta^0-2\beta^{0\prime}V_{xy}=-\beta^{0\prime}V_x\beta^0$ which implies the sampling variance that ignores the variability in the group means will be too large whenever the regression coefficients in the treated and untreated samples are close to one another – which will typically be the case.

Thus, it is possible to conduct conservative inference based on plugin estimators of $V^{naive}\equiv V_y+\mu_x^{1\prime}V_\beta\mu_x^1-2\mu_x^{1\prime}V_{\beta y}$. This is easily accomplished in standard regression packages by running a regression of Y_i on D_i and the elements of $(1-D_i)X_i$ without a constant. Heuristically, the standard error on the coefficient accompanying D_i provides an estimate of V_y, the covariance matrix for the coefficients accompanying the elements of $(1-D_i)X_i$ provide estimates of V_β, and the covariance between these two sets of coefficients provide estimates of $V_{\beta y}$. Note that in the absence of clustering or cross-sectional dependence $V_{\beta y}=0$. Finally, one simply plugs in sample means as estimates of μ_x^1 to obtain the composite variance estimate $\hat{V}^{naive}$.

4 Full variance estimation

To derive the full asymptotic variance, note that:

---

[3]It is important to note that there is a distinction between ignoring uncertainty in the group means $\hat{\mu}_x$ and in conditioning on the underlying covariates X_i themselves. If one wishes to conduct inference conditional on an experimental design $\{D_i,X_i\}_{i=1}^N$ it is necessary to estimate $V_{\theta|x}\equiv V_{y|x}+\mu_x^{1\prime}V_{\beta|x}\mu_x^1-2\mu_x^{1\prime}V_{\beta y|x}$ where $V_{y|x}$, $V_{\beta|x}$, and $V_{\beta y|x}$ give conditional asymptotic variances. In general $V_y\geq V_{y|x}$ and so one will again tend to overestimate the conditional asymptotic variance by simply ignoring variability in the group means $\hat{\mu}_x$. See Abadie, Imbens, and Zheng (2011) for further discussion.

[4]They may however be correlated if they are in the same clusters or share some other form of cross-sectional dependence. However, for $V_{\beta x}\neq 0$ one also needs for the regression model for the control sample to be misspecified so that $E\left[Y_i-X_i'\beta^0|X_i,D_i=0\right]\neq 0$.

[5]This follows because with *iid* data $V_x=\frac{E[X_iX_i'|D_i=1]}{E[D_i]}-\mu_x\mu_x^{1\prime}$ while $V_{xy}=\frac{E[X_iY_i|D_i=1]}{E[D_i]}-\mu_x^1E[Y_i|D_i=1]=\frac{E[X_iX_i'|D_i=1]\beta^1}{E[D_i]}-\mu_x^1\mu_x^{1\prime}\beta^1$.

$$\hat{\theta} = \frac{\frac{1}{N}\sum_i D_i\left(Y_i - X_i'\beta^0\right) - \frac{1}{N}\sum_i D_i X_i'\left(\hat{\beta}^0 - \beta^0\right)}{\frac{1}{N}\sum_i D_i}$$

$$= \theta + \frac{\frac{1}{N}\sum_i D_i\left(Y_i - X_i'\beta^0 - \theta\right)}{\frac{1}{N}\sum_i D_i} - \frac{\frac{1}{N}\sum_i D_i X_i'}{\frac{1}{N}\sum_i D_i}\left(\hat{\beta}^0 - \beta^0\right)$$

$$= \theta + \frac{\frac{1}{N}\sum_i D_i\left(Y_i - X_i'\beta^0 - \theta\right)}{\frac{1}{N}\sum_i D_i} - \frac{\frac{1}{N}\sum_i D_i X_i'}{\frac{1}{N}\sum_i D_i}\left(\frac{1}{N}\sum_j (1-D_j) X_j X_j'\right)^{-1}\frac{1}{N}\sum_j (1-D_j) X_j\left(Y_j - X_j'\beta^0\right).$$

By assumption, a central limit theorem applies, so that:

$$\left[\begin{array}{c}\frac{1}{\sqrt{N}}\sum_i D_i\left(Y_i - X_i'\beta^0 - \theta\right)\\ \frac{1}{\sqrt{N}}\sum_j (1-D_j) X_j\left(Y_j - X_j'\beta^0\right)\end{array}\right] \xrightarrow{d} N\left(0, \begin{array}{cc} V_{d(y-x'\beta)} & \\ V_{(1-d)x(y-x'b),d(y-x'\beta)} & V_{(1-d)x(y-x'b)}\end{array}\right),$$

where,

$$V_{d(y-x'\beta)} \equiv \lim_{N\to\infty} E\left[\left(\frac{1}{\sqrt{N}}\sum_i D_i\left(Y_i - X_i'\beta^0 - \theta\right)\right)^2\right]$$

$$= E[D_i]^2\left(V_y + \beta^{0\prime}V_x\beta^0 - 2\beta^{0\prime}V_{xy}\right),$$

$$V_{(1-d)x(y-x'b)} \equiv \lim_{N\to\infty} E\left[\left(\frac{1}{\sqrt{N}}\sum_i (1-D_i) X_i\left(Y_i - X_i'\beta^0\right)\right)\left(\frac{1}{\sqrt{N}}\sum_i (1-D_i) X_i\left(Y_i - X_i'\beta^0\right)\right)'\right]$$

$$= E\left[(1-D_i) X_i X_i'\right] V_\beta E\left[(1-D_i) X_i X_i'\right],$$

$$V_{(1-d)x(y-x'b),d(y-x'\beta)} \equiv \lim_{N\to\infty} E\left[\left(\frac{1}{\sqrt{N}}\sum_j (1-D_j) X_j\left(Y_j - X_j'\beta^0\right)\right)\left(\frac{1}{\sqrt{N}}\sum_i D_i\left(Y_i - X_i'\beta^0 - \theta\right)\right)\right]$$

$$= E[D_i]^2 E\left[(1-D_i) X_i X_i'\right]\left(V_{\beta y} - \beta^{0\prime}V_{\beta x}\right).$$

Therefore, by Slutzky's theorem,

$$\sqrt{N}\left(\hat{\theta} - \theta\right) = E[D_i]^{-1}\frac{1}{\sqrt{N}}\sum_i D_i\left(Y_i - X_i'\beta^0 - \theta\right)$$

$$-E[X_i|D_i = 1]' E[X_i X_i'|D_i = 0]^{-1}\frac{1}{\sqrt{N}}\sum_j (1-D_j) X_j\left(Y_j - X_j'\beta^0\right) + o_p(1).$$

Hence, the asymptotic variance can be written:

$$V_\theta = E[D_i]^{-2} V_{d(y-x'\beta)} \tag{1}$$

$$+\mu_x^{1\prime} E\left[(1-D_i) X_i X_i'\right]^{-1} V_{(1-d)x(y-x'b)} E\left[(1-D_i) X_i X_i'\right]^{-1}\mu_x^1 \tag{2}$$

$$-2\mu_x^{1\prime} E\left[X_i X_i'|D_i = 0\right]^{-1} V_{(1-d)x(y-x'b),d(y-x'\beta)}. \tag{3}$$

One can estimate V_θ using appropriate sample analogues. For example, if the data are independent and identically distributed across observations then, $\hat{V}_{d(y-x'\beta)}^{iid} \equiv \frac{1}{N}\sum_i D_i\left(Y_i - X_i'\hat{\beta}^0 - \hat{\theta}\right)^2 \xrightarrow{p} V_{d(y-x'\beta)}$, $\hat{V}_{d(y-x'\beta)}^{iid} \equiv$

$\frac{1}{N} \sum_i (1 - D_i) X_i \left(Y_i - X_i'\hat{\beta}^0\right)^2 \xrightarrow{p} V_{d(y-x'\beta)}$, and $V_{d(y-x'\beta),(1-d)x(y-x'b)}^{iid} = 0$. Hence, a consistent estimator of V_θ is given by:

$$\hat{V}_\theta^{iid} \equiv \frac{\hat{V}_{d(y-x'\beta)}^{iid}}{\left(\frac{1}{N}\sum D_i\right)^2} + \hat{\mu}_x^{1\prime} \left[\frac{1}{N} \sum_i (1 - D_i) X_i X_i'\right]^{-1} \hat{V}_{(1-d)x(y-x'b)}^{iid} \left[\frac{1}{N} \sum_i (1 - D_i) X_i X_i'\right]^{-1} \hat{\mu}_x^1.$$

If the data instead exhibit dependence within (but not between) clusters then, letting $c(i)$ denote the cluster corresponding to observation i, variance estimators that are consistent for fixed cluster sizes as the number of clusters approaches infinity are given by:

$$\hat{V}_{d(y-x'\beta)}^{cluster} \equiv \frac{1}{N} \sum_{c'} \left(\sum_{i:c(i)=c'} D_i \left(Y_i - X_i'\hat{\beta}^0 - \hat{\theta}\right)\right)^2 \xrightarrow{p} V_{d(y-x'\beta)},$$

$$\hat{V}_{(1-d)x(y-x'b)}^{cluster} = \frac{1}{N} \sum_{c'} \left(\sum_{i:c(i)=c'} (1 - D_i) X_i \left(Y_i - X_i'\hat{\beta}^0\right)\right) \left(\sum_{i:c(i)=c'} (1 - D_i) X_i \left(Y_i - X_i'\hat{\beta}^0\right)\right)' \xrightarrow{p} V_{(1-d)x(y-x'b)},$$

$$\hat{V}_{(1-d)x(y-x'b),d(y-x'\beta)}^{cluster} \equiv \frac{1}{N} \sum_{c'} \left(\sum_{j:c(j)=c'} (1 - D_j) X_j \left(Y_j - X_j'\hat{\beta}^0\right)\right) \left(\sum_{i:c(i)=c'} D_i \left(Y_i - X_i'\hat{\beta}^0 - \hat{\theta}\right)\right)$$
$$\xrightarrow{p} V_{(1-d)x(y-x'b),d(y-x'\beta)}.$$

provided that standard regularity conditions hold (Newey and McFadden, 1994). Thus, in this case, a consistent estimator of V_θ is given by:

$$\hat{V}_\theta^{cluster} \equiv \frac{\hat{V}_{d(y-x'\beta)}^{cluster}}{\left(\frac{1}{N}\sum D_i\right)^2} + \hat{\mu}_x^{1\prime} \left[\frac{1}{N} \sum_i (1 - D_i) X_i X_i'\right]^{-1} \hat{V}_{(1-d)x(y-x'b)}^{cluster} \left[\frac{1}{N} \sum_i (1 - D_i) X_i X_i'\right]^{-1} \hat{\mu}_x^1.$$
$$- 2\hat{\mu}_x^{1\prime} \left[\frac{1}{N} \sum_i (1 - D_i) X_i X_i'\right]^{-1} \hat{V}_{(1-d)x(y-x'b),d(y-x'\beta)}^{cluster}.$$

5 A computational trick

The above calculations are rather tedious. Fortunately, standard regression software can be used to compute the necessary variance estimates in a simple way. The following regression based procedure will provide an estimate of V_θ:

1) Estimate $\hat{\beta}^0$ via OLS in the untreated sample.

2) Form a new variable $Y_i^* = D_i \left(Y_i - X_i'\hat{\beta}^0\right) + (1 - D_i) Y_i$.

3) Regress Y_i^* on D_i and the elements of $(1 - D_i) X_i$ without a constant using an appropriate variance estimation technique (e.g. clustering).

4) The coefficient on D_i provides the Oaxaca point estimate $\hat{\theta}$ with (in an abuse of notation) corresponding variance estimate $\hat{V}_{\hat{\theta}}$ which (given appropriate regularity conditions) will be consistent for the term in (1). The coefficients on the elements of $(1 - D_i) X_i$ provide estimates of $\hat{\beta}^0$ with corresponding variance estimate $\hat{V}_{\hat{\beta}}$ which will be consistent for the term $E\left[(1 - D_i)X_iX_i'\right]^{-1} V_{(1-d)x(y-x'b)} E\left[(1 - D_i)X_iX_i'\right]^{-1}$ in (2). The regression also supplies an estimate of $\hat{V}_{\hat{\beta}\hat{\theta}}$ which will be zero unless some form of cross-sectional dependence is allowed. $\hat{V}_{\hat{\beta}\hat{\theta}}$ will be consistent for the term $E\left[X_iX_i'|D_i = 0\right]^{-1} V_{(1-d)x(y-x'b),d(y-x'\beta)}$ in (3).

5) The full variance estimate is computed as $\hat{V}_{\hat{\theta}} + \hat{\mu}_x^{1\prime} \hat{V}_{\hat{\beta}} \hat{\mu}_x^1 - 2\hat{\mu}_x^{1\prime} \hat{V}_{\hat{\beta}\hat{\theta}}$ which will be consistent for V_θ.

6 Monte Carlo

We now consider the performance of the proposed Oaxaca variance estimator in a simple Monte Carlo simulation design. The data generating process is given by:

$$Y_i = 2 + (1 - D_i)(2X_i) + D_i(3X_i) + \eta_{c(i)}^{(1)} + \varepsilon_i$$
$$D_i = 1\left[\eta_{c(i)}^{(2)} + v_i > 0\right],$$

where $v_i \sim N(0,1)$, and ε_i, $\eta_c^{(1)}$, and $\eta_c^{(2)}$ are each drawn from independent students-t distributions with six degrees of freedom. This design ensures a within cluster correlation in both the treatment D_i and the composite regression error $\eta_{c(i)}^1 + \varepsilon_i$, which necessitates cluster robust inference. Because treatment varies within cluster, it is possible for $V_{\beta y} \neq 0$. As noted by Chesher (1995), Monte Carlo experiments utilizing symmetrically distributed covariates are likely to give overly optimistic results. To avoid this problem, we set $X_i = 4\left(X_i^* - \frac{2}{7}\right) + D_i$ where $X_i^* \sim Beta(2,5)$, which yields a regressor with a substantial amount of skew and a reasonable amount of predictive power.[6] In this design, the differences between β^0 and β^1 and the fact that $\mu_x^1 = 1$ imply the population value of θ equals one.

Table I reports simulation results for several different sample sizes. In all cases each cluster has exactly 10 observations. Columns (3) and (4) use the regression based procedure of Section 5 to compute variance estimates, while columns (5) and (6) use the procedure of Section 3 which ignores the uncertainty in the sample means $\hat{\mu}_x^1$.[7] As column (1) indicates, the Oaxaca point estimates $\hat{\theta}$ are essentially unbiased in this design. Unsurprisingly, the standard error estimates summarized in column (3) slightly underestimate the true variability of the estimator when few clusters are present, leading to mild over-rejection of hypothesis tests. With 100 clusters, the standard error estimates yield nearly exact coverage. As expected, ignoring uncertainty in the mean value of the covariate increases the estimated standard errors, which leads to under-rejection. In this case, the problem is relatively mild, with a 6-7% over-estimate of the true asymptotic standard error. Interestingly, although ignoring uncertainty in $\hat{\mu}_x^1$ asymptotically inflates the standard error estimate, it does not always do so in finite samples, as column (7) makes clear.

			Accounting for Uncertainty in $\hat{\mu}_x^1$		Ignoring Uncertainty in $\hat{\mu}_x^1$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Clusters	Mean $\hat{\theta}$	Std. Dev. $\hat{\theta}$	Mean $\sqrt{\frac{\hat{V}_\theta}{N}}$	False Reject Prob.	Mean $\sqrt{\frac{\hat{V}_\theta}{N}}$	False Reject Prob.	Prob. Ignoring Lowers $\sqrt{\frac{\hat{V}_\theta}{N}}$
25	0.9952	0.4374	0.4157	0.0651	0.4463	0.0472	0.0895
50	0.9997	0.3066	0.3003	0.0557	0.3214	0.0410	0.0255
100	1.0002	0.2161	0.2152	0.0494	0.2300	0.0355	0.0025
200	0.9995	0.1541	0.1529	0.0520	0.1632	0.0382	0.0002

Table I

Notes: Statistics computed using 10,000 simulation draws. "False Reject Prob." refers to fraction of simulations in which a Wald test rejected the hypothesis that $\theta = 1$ at the 5% level. "Prob. Ignoring Lowers $\sqrt{\frac{\hat{V}_\theta}{N}}$" reports fraction of simulations in which standard error accounting for uncertainty in $\hat{\mu}_x^1$ exceeded standard error ignoring uncertainty in $\hat{\mu}_x^1$.

---

[6]The population R^2 in the regression of Y_i on X_i in the $D_i = 0$ sample is approximately 35%.

[7]Variance estimates were computed in Stata 12.1 using the "cluster" variance estimation routine which uses a small sample degrees of freedom adjustment of the form $\frac{C}{C-1}$ where C is the number of clusters. Estimates ignoring the variance in the sample means still allow for correlation between $\hat{\beta}$ and $\hat{\mu}_y^1$ – that is, they take the form $\hat{V}_y + \hat{\mu}_x^{1\prime}\hat{V}_\beta\hat{\mu}_x^1 - 2\hat{\mu}_x^{1\prime}\hat{V}_{\beta y}$. All code used in this paper can be found online at http://emlab.berkeley.edu/~pkline/.

7 Conclusion

The Oaxaca estimator of average treatment effects studied by Kline (2011) is a simple alternative to matching and inverse probability weighting methods in evaluation problems involving selection on observables. This note derived the estimator's limiting distribution. A simple computational approach to standard error calculation was proposed that makes use of conventional statistical routines. Naive routines that ignore variability in the mean values of the covariates are likely to yield conservative inferences.

References

Abadie, A., Imbens, G. W., & Zheng, F. (2011). Robust Inference for Misspecified Models Conditional on Covariates (No. w17442). National Bureau of Economic Research.

Angrist, J., & Rokkanen, M. (2012). Wanna Get Away? RD Identification Away from the Cutoff (No. w18662). National Bureau of Economic Research.

Busso, M., Gregory, J., & Kline, P. (2013). Assessing the Incidence and Efficiency of a Prominent Place Based Policy. The American Economic Review, 103(2), 897-947.

Chesher, A. (1995). A mirror image invariance for M-estimators. Econometrica, 63(1), 207-211.

Fortin, N., Lemieux, T., & Firpo, S. (2011). Decomposition methods in economics. Handbook of labor economics, 4, 1-102. Chicago

Heckman, James J. and Robb, R. 1985. Alternative methods for evaluating the impact of interventions. In J. Heckman & B. Singer (Eds.), Longitudinal Analysis of Labor Market Data (pp. 156-245). New York: Cambridge University Press.

Kline, P. (2011). Oaxaca-Blinder as a reweighting estimator. American Economic Review, 101(3), 532-37.

Kline, P. & Moretti, E. (forthcoming). Local economic development, agglomeration economies, and the big push: 100 years of evidence from the tennessee valley authority. Quarterly Journal of Economics.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. International economic review, 14(3), 693-709.

Oaxaca, R. L., & Ransom, M. R. (1994). On discrimination and the decomposition of wage differentials. Journal of econometrics, 61(1), 5-21.

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

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. Handbook of econometrics, 4, 2111-2245.

Wooldridge, J.M. (2002.) Inverse probability weighted M-estimation for sample selection, attrition, and stratification. Portuguese Economic Journal 1, 117–139.

Wooldridge, J.M. (2007.) Inverse probability weighted estimation for general missing data problems. Journal of Econometrics. 141(2), 1281-1301.

Van der Vaart, A. W. (2000). Asymptotic statistics (Vol. 3). Cambridge university press.