

Nizalova, Olena Y.; Murtazashvili, Irina

**Working Paper**

## Exogenous treatment and endogenous factors: Vanishing of omitted variable bias on the interaction term

IZA Discussion Papers, No. 6282

**Provided in Cooperation with:**

IZA – Institute of Labor Economics

*Suggested Citation:* Nizalova, Olena Y.; Murtazashvili, Irina (2012) : Exogenous treatment and endogenous factors: Vanishing of omitted variable bias on the interaction term, IZA Discussion Papers, No. 6282, Institute for the Study of Labor (IZA), Bonn, <https://nbn-resolving.de/urn:nbn:de:101:1-201203064127>

This Version is available at:

<https://hdl.handle.net/10419/58438>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

IZA DP No. 6282

**Exogenous Treatment and Endogenous Factors:  
Vanishing of Omitted Variable Bias on the  
Interaction Term**

Olena Nizalova  
Irina Murtazashvili

January 2012

# **Exogenous Treatment and Endogenous Factors: Vanishing of Omitted Variable Bias on the Interaction Term**

**Olena Nizalova**

*Kyiv School of Economics / Kyiv Economics Institute  
and IZA*

**Irina Murtazashvili**

*University of Pittsburgh*

Discussion Paper No. 6282  
January 2012

IZA

P.O. Box 7240  
53072 Bonn  
Germany

Phone: +49-228-3894-0  
Fax: +49-228-3894-180  
E-mail: [iza@iza.org](mailto:iza@iza.org)

Any opinions expressed here are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit organization supported by Deutsche Post Foundation. The center is associated with the University of Bonn and offers a stimulating research environment through its international network, workshops and conferences, data service, project support, research visits and doctoral program. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

## **ABSTRACT**

### **Exogenous Treatment and Endogenous Factors: Vanishing of Omitted Variable Bias on the Interaction Term**

Whether interested in the differential impact of a particular factor in various institutional settings or in the heterogeneous effect of policy or random experiment, the empirical researcher confronts a problem if the factor of interest is correlated with an omitted variable. This paper considers circumstances under which the estimate of the mentioned effect is consistent. We find that if the source of heterogeneity and omitted variable are jointly independent of policy or treatment, then the OLS estimate on the interaction term between the treatment and endogenous factor turns out to be consistent.

JEL Classification: C21, C93

Keywords: treatment effect, heterogeneity, policy evaluation, random experiments, omitted variable bias

Corresponding author:

Olena Nizalova  
Kyiv School of Economics / Kyiv Economics Institute  
13 Yakira Str  
Kyiv, 04119  
Ukraine  
E-mail: [nizalova@kse.org.ua](mailto:nizalova@kse.org.ua)

# 1 Introduction

Significant increase in the evaluations of random and natural experiments throughout economics fields is raising the question of whether it is possible to obtain a consistent estimate of the heterogeneous treatment effect if the heterogeneity is occurring along the lines of a factor which is correlated with some omitted variable(s). The textbook approach to econometric modeling suggests that the exclusion of a relevant regressor correlated with included explanatory variables will result in an omitted variable bias. But are there circumstances when the exclusion of a relevant variable is of not such a severe consequence? It turns out that this situation is indeed possible and quite common in applied works. If all the regressors but the exogenous regressor and the interaction term between this exogenous regressor and an endogenous covariate are jointly independent of the exogenous regressor of interest, the OLS estimate of interaction term's coefficient is consistent. While a special case, it is very common in applied studies and is of huge relevance for policy analysis. Here are several recent examples.

Glewwe, Kremer, and Moulin (2009) in the evaluation of a randomized trial of a free textbook provision in rural Kenya find that only students with good past test scores benefit from the intervention. Banerjee et al. (2007) also consider the previous test scores as a source of heterogeneity of an impact of a remedial education program in India and find that the largest gains are experienced by children with lowest test-scores. Similarly, Banerjee et al. (2010) estimate the impact of a randomized introduction of microcredit in a new market depending on the presence of existing business at the time of the program or on the propensity to become business owners. Clearly, the past test scores and current business ownership (propensity to become a business owner) are correlated with unobserved variables (for example, ability), raising the question of whether the estimates can be trusted.

However, the only study that mentions the issue of heterogeneity along the lines a factor correlated with omitted variables is Blank (1991) which evaluates the American Economic Review experiment of a double-blind refereeing process held over the period 1987-1989 and resulted in 1,498 papers with completed referee reports. Double-blind or single-blind review has been assigned randomly. The evaluation suggests that the double-blind procedure is stricter: lower acceptance rate and more critical referee reports. However, the emphasis of the paper was not on the overall effect of the double-blind refereeing, but rather on the heterogeneous treatment impact along several dimensions - gender and university rank being the most important. Clearly, gender is likely to be correlated with other important factors unobserved in

the experiment - age and experience in the profession, which affect acceptance rate. Likewise, being in a higher ranked university maybe the result of the overall higher unobserved productivity. The coefficients on interaction terms turn out to be statistically insignificant, suggesting no benefits of double-blind refereeing to either women or authors from lower-ranked universities. But can this finding be trusted? The author states that the coefficients on the interaction terms "should be robust to the inclusion of any other variables in the model, since they come from two experimental samples that are identical in all other characteristics". However, with respect to the main effects of gender and the university rank, the author claims that "it is not clear how to interpret the coefficients on these variables, because they are contaminated by excluded variables" (Blank 1991, p. 1054-55). We are to prove this explicitly in the paper - the consistency of the estimates of the heterogeneous impact of random treatment/ exogenous policy when the heterogeneity occurs along the lines of a factor correlated with the omitted variable(s).

## 2 Econometric Result in the Context of the AER Experiment

For concreteness, let us talk about the example of the AER experiment in Blank (1991). To start with, we consider only one dimension of heterogeneity - the rank of the university. A simplified relation between the acceptance rates and assignment to the blind review group is described as:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon^*, \quad (1)$$

where  $x_2 = x_3 \cdot x_4$  and  $\varepsilon^* = \varepsilon + \beta_5 x_5$ . Here,  $x_3$  specifies the university rank<sup>1</sup>,  $x_4$  is an indicator of the double-blind treatment,  $x_5$  is the unobserved individual-specific effect, and  $\varepsilon$  is the idiosyncratic error. The question is whether the effect of the double-blind reviewing affects the acceptance rates differently depending on the university rank, i.e. whether  $\beta_2$  is statistically different from zero. However, the unobserved personality traits are correlated with the university rank. Standard econometric wisdom suggests that in a cross-sectional setting the estimates of all the parameters will be inconsistent since  $\text{Corr}(x_3, x_5) \neq 0$ . But is this indeed the case?

The probability limit of  $\widehat{\beta}_2$  is derived in the Appendix in formulas (7) and (8), which show that the "bias" term in  $\widehat{\beta}_2$  equals  $\beta_5 Q_2$ , where we can express  $Q_2$  as

$$Q_2 = \text{plim} \frac{\sigma_5}{\sigma_2} \cdot \frac{r_{25}(1 - r_{34}^2) + r_{35}(r_{24}r_{34} - r_{23}) + r_{45}(r_{23}r_{34} - r_{24})}{1 - r_{23}^2 - r_{24}^2 - r_{34}^2 + 2r_{23}r_{24}r_{34}}. \quad (2)$$

---

<sup>1</sup>Although the university rank is represented by a set of indicators in Blank (1991), we use one variable,  $x_3$ .

Here,  $\sigma_j$  is the sample standard deviation of  $x_j$ ,  $j = 2, 3, 4, 5$ , and  $r_{kl}$  is the sample correlation between  $x_k$  and  $x_l$ ,  $k = 2, 3, 4$  and  $l = 3, 4, 5$ .  $Q_2$  is the plim of the regression coefficient on  $x_2$  in the “auxiliary” regression of the excluded variable,  $x_5$ , on the included variables,  $x_2$ ,  $x_3$  and  $x_4$ . Equation (2) is derived under the usual assumptions, which impose no restrictions on the relationships among any variables. However, we can assume zero correlations among some of the variables when studying the heterogeneous treatment effect, assuming for example that the university rank and author productivity are jointly independent of the assignment process to the double-blind reviewing procedure. Then we obtain

$$Q_2 = \text{plim} \frac{\sigma_5}{\sigma_2} \cdot \frac{r_{25} - r_{23}r_{35}}{1 - r_{23}^2 - r_{24}^2}. \quad (3)$$

Further, note that independence of  $x_4$  and  $(x_3, x_5)$  implies that (1)  $x_4$  is independent of  $x_3$ , and (2)  $x_4$  is independent of  $x_5$  conditional on  $x_3$ . The first condition guarantees  $r_{23} = \sigma_3 \cdot \frac{E(x_4)}{\text{S.D.}(x_3 \cdot x_4)}$ , where  $E(x_4)$  is the expected value of  $x_4$  and  $\text{S.D.}(x_3 \cdot x_4)$  is the standard deviation of  $x_3 \cdot x_4$ . The two conditions together insure  $r_{25} = r_{23}r_{35}$ . Thus, when  $x_4$  and  $(x_3, x_5)$  are independent,  $\frac{r_{25} - r_{23}r_{35}}{1 - r_{23}^2 - r_{24}^2} = 0$  and the “bias” term disappears:  $\text{plim}(\hat{\beta}_2) = \beta_2$ . This implies that the coefficient estimate of  $x_3 \cdot x_4$  is consistent under independence of  $x_4$  and  $(x_3, x_5)$ , which is actually stronger than necessary to guarantee this result. For consistency, it would be sufficient to have either  $f(x_3|x_5, x_4) = f(x_3|x_5)$  or  $f(x_5|x_3, x_4) = f(x_5|x_3)$  in combination with  $x_4$  being independent of either  $x_5$  or  $x_3$ , respectively.

Let us revisit the Blank’s (1991) study. The question of interest there is estimating the differences in the effect of the double-blind reviewing procedure for different groups of researchers. The author is after the coefficient estimate of the interaction term between the university rank and variable identifying the sample randomly assigned to the double-blind reviewing. While there are valid reasons to suspect that the university rank is correlated with the unobservables (say, productivity of the author), this treatment is independent of the university rank as well as productivity of the authors once rank is accounted for. These two independences guarantee that the OLS estimates of the interaction terms between university rank and treatment dummies are consistent as we show above.

### 3 Inclusion of Additional Explanatory Variables

In this section Monte Carlo simulations are employed to extend theoretical findings from the previous section to the cases when other regressors are added to Equation (1). The number of replications is 1000.

The data generating process (DGP) employed is:

$$y_i = 1 + 2r_i \cdot d_i + 3r_i + 4d_i + 5f_i + 6s_i + 7n_i + 8c_i + u_i. \quad (4)$$

Here,  $r_i$  (university rank) and  $u_i$  (idiosyncratic error) are generated as independent Normal  $(0, 1)$ . The unobserved heterogeneity,  $c_i$ , is generated as  $c_i = \lambda r_i + e_i^c$ , where  $e_i^c \sim \text{Normal}(0, 1)$ . The exogenous treatment,  $d_i$ , is generated as Bernoulli  $(0.5)$ .

We consider three possibilities for additional regressors: (1) regressors independent of  $c_i$  (gender of the referee), (2) regressors correlated with  $d_i$  but uncorrelated with  $c_i^2$ , (3) regressors with non-zero simple correlation with  $c_i$  (gender of the author). Case (1) is represented by  $f_i \sim \text{Bernoulli}(0.5)$ . Case (2) is represented by  $s_i = 0.5d_i - 1 + e_i^s$ , where  $e_i^s \sim \text{Discrete Uniform}(0, 3)$ . For case (3), we consider two DGPs for  $n_i$  – the rank of the school granting doctorate to the author<sup>3</sup>: (A)  $n_i = 0.5r_i + e_i^n$ , and (B)  $n_i = 0.5c_i + e_i^n$ , where  $e_i^n \sim \text{Normal}(0, 1)$ . These two DGPs result in non-zero simple correlation between  $c_i$  and  $n_i$ . However, the partial correlation between  $n_i$  and  $c_i$ , i.e. correlation net of the effect of the other included regressors (in particular,  $r_i$ ), is zero for DGP (A), while it is clearly not for DGP (B).

Tables 1 and 2 present simulation results. We consider two regressions: with six ( $c_i$  is excluded) and seven ( $c_i$  is included) regressors (in addition to the constant term). When  $\lambda = 0$ ,  $\text{Corr}(r_i, c_i) = 0$  and OLS estimation delivers consistent estimates of all model parameters for both regressions considered, as long as  $c_i$  is partially uncorrelated with all of the additional included regressors. Indeed, when  $\lambda = 0$ ,  $\hat{\beta}_2$ ,  $\hat{\beta}_4$  and  $\hat{\beta}_7$  are always consistent when  $n_i \sim \text{DGP (A)}$ . As expected, even when  $\lambda = 0$ ,  $\hat{\beta}_7$  is inconsistent when  $n_i \sim \text{DGP (B)}$ , since the partial correlation between  $c_i$  and  $n_i$  is not zero in that case. Contrary,  $\hat{\beta}_3$  from the model with six regressors is consistent only when  $\lambda = 0$  regardless of  $N$  and DGP for  $n_i$ . The fact that  $\text{Corr}(s_i, d_i) \neq 0$  has no effect on any of the OLS estimates in all cases, since these variables are independent of  $c_i$ . Similarly,  $\hat{\beta}_5$  is always consistent.

Clearly, when seven regressors are included all estimates are consistent. More importantly, when only six regressors are used,  $\hat{\beta}_2$  and  $\hat{\beta}_4$  are consistent, while the consistency of  $\hat{\beta}_3$  and  $\hat{\beta}_7$  depends on the (partial) correlations  $\text{Corr}(r_i, c_i)$  and  $\text{Corr}(n_i, c_i)$ , respectively. The simulation findings are unambiguous: when the partial correlation between the unobserved heterogeneity and some included regressor is different from zero, the OLS slope estimate of that included regressor is the only estimate which is inconsistent, and its

<sup>2</sup>It is difficult to think of such a regressor in the AER experiment, but generally it is possible to have such variables.

<sup>3</sup>In case of multiple authors, this can be measured by the highest rank of the schools granting doctorate among all co-authors.



bias does not disappear as  $N \rightarrow \infty$ .

## 4 Conclusions

Increasing interest in the heterogeneity of the impact in policy evaluation and random experiment settings leads to a question of whether the estimates are consistent when the source of heterogeneity is correlated with some omitted variable(s). This paper presents the conditions under which it is possible to arrive at a consistent OLS estimate of the mentioned effect. We discuss relevant applications and illustrate our findings with some simulation evidence.

## **Acknowledgments**

This paper has benefited from helpful comments and suggestions of Tom Coupé, Soiliou Namoro, Jean-Francois Richard, Peter Schmidt, and Jeffrey Wooldridge.

## **References**

Banerjee, A. V., E. Duflo, R. Glennerster and C. Kinnan, 2010. The Miracle of Microfinance? Evidence from a Randomized Evaluation. Accessed at <http://econ-www.mit.edu/files/5993> on 09/02/2011.

Banerjee, A. V., S. Cole, E. Duflo and L. Linden, 2007. Remediating Education: Evidence from Two Randomized Experiments in India. *The Quarterly Journal of Economics* 122:3, 1235-1264.

Blank, Rebecca M., 1991. The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review. *American Economic Review* 81:5, 1041-1067.

Glewwe, P., M. Kremer and S. Moulin., 2009. Many Children Left Behind? Textbooks and Test Scores in Kenya. *American Economic Journal: Applied Economics* 1:1, 112-135.

Green, W.H., 2003. *Econometric Analysis*. 5th ed. Englewood Cliffs, NJ: Prentice Hall.

Table 1: OLS Estimation Results for  $(\beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7)' = (2, 3, 4, 5, 6, 7)'$  and  $\lambda = 0.5$ .

# of Regressors:	6	7	6	7	6	7	6	7
	$N = 100$		$N = 1000$		$N = 100$		$N = 1000$	
	(A): $n_i = 0.5r_i + e_i^n$				(B): $n_i = 0.5c_i + e_i^n$			
(1) $\widehat{\beta}_2$	1.959	1.978	2.010	2.003	1.962	1.978	2.013	2.003
SE( $\widehat{\beta}_2$ )	(1.677)	(0.210)	(0.512)	(0.064)	(1.502)	(0.210)	(0.459)	(0.064)
(2) $\widehat{\beta}_3$	7.007	3.004	6.994	3.000	6.228	3.007	6.192	2.999
SE( $\widehat{\beta}_3$ )	(1.253)	(0.165)	(0.384)	(0.050)	(1.071)	(0.157)	(0.329)	(0.048)
(3) $\widehat{\beta}_4$	4.051	4.006	3.980	3.996	4.058	4.006	3.990	3.996
SE( $\widehat{\beta}_4$ )	(1.688)	(0.211)	(0.524)	(0.065)	(1.512)	(0.211)	(0.470)	(0.065)
(4) $\widehat{\beta}_5$	4.952	5.005	5.033	5.001	4.967	5.005	5.032	5.001
SE( $\widehat{\beta}_5$ )	(1.652)	(0.206)	(0.511)	(0.063)	(1.479)	(0.206)	(0.458)	(0.063)
(5) $\widehat{\beta}_6$	6.000	6.001	5.998	5.998	6.003	6.001	5.993	5.998
SE( $\widehat{\beta}_6$ )	(0.739)	(0.092)	(0.229)	(0.028)	(0.661)	(0.092)	(0.205)	(0.028)
(6) $\widehat{\beta}_7$	7.028	7.007	6.993	6.999	10.193	7.007	10.190	6.999
SE( $\widehat{\beta}_7$ )	(0.833)	(0.104)	(0.256)	(0.032)	(0.666)	(0.104)	(0.205)	(0.032)
(7) RMSE( $\widehat{\beta}_2$ )	1.700	0.210	0.505	0.064	1.552	0.210	0.450	0.064
(8) SD( $\widehat{\beta}_2$ )	1.701	0.209	0.506	0.064	1.552	0.209	0.450	0.064
(9) LQ( $\widehat{\beta}_2$ )	0.795	1.836	1.674	1.961	0.967	1.836	1.693	1.961
(10) Median( $\widehat{\beta}_2$ )	1.958	1.972	2.008	2.002	1.982	1.972	2.000	2.002
(11) UQ( $\widehat{\beta}_2$ )	3.197	2.108	2.356	2.044	3.032	2.108	2.319	2.044

Notes: Odd columns report results for the estimating equation with six regressors, while even columns – for the estimating equation with all seven regressors. Rows (1) through (6) contain means of OLS slope estimates and their corresponding standard errors from 1000 replications. Rows (7) through (11) contain the root mean squared error (RMSE), standard deviation (SD), lower quartile (LQ), median, and upper quartile (UQ) for  $\widehat{\beta}_2$  – our main coefficient of interest – from 1000 replications. Also, the first four columns report the results when  $n_i$  is generated according to DGP (A), while the last four columns – according to DGP (B).

Table 2: OLS Estimation Results for  $(\beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7)' = (2, 3, 4, 5, 6, 7)'$  and  $\lambda = 0$ .

# of Regressors:	6	7	6	7	6	7	6	7
	$N = 100$		$N = 1000$		$N = 100$		$N = 1000$	
	(A): $n_i = 0.5r_i + e_i^n$				(B): $n_i = 0.5c_i + e_i^n$			
(1) $\hat{\beta}_2$	1.959	1.978	2.010	2.003	1.962	1.978	2.013	2.003
SE( $\hat{\beta}_2$ )	(1.677)	(0.210)	(0.512)	(0.064)	(1.502)	(0.210)	(0.459)	(0.064)
(2) $\hat{\beta}_3$	3.007	3.007	2.994	3.000	3.026	3.010	2.990	2.999
SE( $\hat{\beta}_3$ )	(1.253)	(0.156)	(0.384)	(0.048)	(1.058)	(0.148)	(0.325)	(0.045)
(3) $\hat{\beta}_4$	4.051	4.006	3.980	3.996	4.058	4.006	3.990	3.996
SE( $\hat{\beta}_4$ )	(1.688)	(0.211)	(0.524)	(0.065)	(1.512)	(0.211)	(0.470)	(0.065)
(4) $\hat{\beta}_5$	4.952	5.005	5.033	5.001	4.967	5.005	5.032	5.001
SE( $\hat{\beta}_5$ )	(1.652)	(0.206)	(0.511)	(0.063)	(1.479)	(0.206)	(0.458)	(0.063)
(5) $\hat{\beta}_6$	6.000	6.001	5.998	5.998	6.003	6.001	5.993	5.998
SE( $\hat{\beta}_6$ )	(0.739)	(0.092)	(0.229)	(0.028)	(0.661)	(0.092)	(0.205)	(0.028)
(6) $\hat{\beta}_7$	7.028	7.007	6.993	6.999	10.193	7.007	10.190	6.999
SE( $\hat{\beta}_7$ )	(0.833)	(0.104)	(0.256)	(0.032)	(0.666)	(0.104)	(0.205)	(0.032)
(6) RMSE( $\hat{\beta}_2$ )	1.700	0.210	0.505	0.064	1.552	0.210	0.450	0.064
(7) SD( $\hat{\beta}_2$ )	1.701	0.209	0.506	0.064	1.552	0.209	0.450	0.064
(8) LQ( $\hat{\beta}_2$ )	0.795	1.836	1.674	1.961	0.967	1.836	1.693	1.961
(9) Median( $\hat{\beta}_2$ )	1.958	1.972	2.008	2.002	1.982	1.972	2.000	2.002
(10) UQ( $\hat{\beta}_2$ )	3.197	2.108	2.356	2.044	3.032	2.108	2.319	2.044

Notes: Odd columns report results for the estimating equation with six regressors, while even columns – for the estimating equation with all seven regressors. Rows (1) through (6) contain means of OLS slope estimates and their corresponding standard errors from 1000 replications. Rows (7) through (11) contain the root mean squared error (RMSE), standard deviation (SD), lower quartile (LQ), median, and upper quartile (UQ) for  $\hat{\beta}_2$  – our main coefficient of interest – from 1000 replications. Also, the first four columns report the results when  $n_i$  is generated according to DGP (A), while the last four columns – according to DGP (B).

## Appendix

We can use the popular econometric textbook by Green (2003) to get the following general result. Suppose the correct specification of the regression model is  $\mathbf{y} = \mathbf{i}\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{X}_3\beta_3 + \varepsilon$ , where  $\mathbf{i}$  is a vector of ones. Premultiplying this equation by matrix  $\mathbf{M}_1 = \mathbf{I} - \mathbf{i}(\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}$ , where  $\mathbf{I}$  is an  $n \times n$  identity matrix, yields a demeaned version of the original model:

$$\mathbf{M}_1\mathbf{y} = \tilde{\mathbf{X}}_2\beta_2 + \tilde{\mathbf{X}}_3\beta_3 + \mathbf{M}_1\varepsilon, \quad (5)$$

where  $\tilde{\mathbf{X}}_2$  and  $\tilde{\mathbf{X}}_3$  are mean-differenced  $\mathbf{X}_2$  and  $\mathbf{X}_3$ .<sup>4</sup> Further, suppose we do not observe  $\mathbf{X}_3$  and estimate  $\mathbf{M}_1\mathbf{y} = \tilde{\mathbf{X}}_2\beta_2 + \varepsilon^*$ , where  $\varepsilon^* = \tilde{\mathbf{X}}_3\beta_3 + \mathbf{M}_1\varepsilon$  and  $\mathbf{M}_1\varepsilon$  is a vector of mean-differenced errors. Then, under the usual assumptions, we modify the omitted variable formula from Green (2003) to report the probability limit of  $\hat{\beta}_2$ :

$$\text{plim}(\hat{\beta}_2) = \beta_2 + \mathbf{Q} \cdot \beta_3, \quad (6)$$

where  $\mathbf{Q} = \text{plim}(\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_3$  is the probability limit of the matrix of regression coefficients from the auxiliary regressions of the excluded mean-differenced variables,  $\tilde{\mathbf{X}}_3$ , on the included mean-differenced variables,  $\tilde{\mathbf{X}}_2$ .

We want to apply general theoretical result (6) to obtain  $\hat{\beta}_2$  in our context. Thus, we can rewrite Equation (6) as

$$\text{plim}(\hat{\beta}_2) = \beta_2 + \beta_5 Q_2, \quad (7)$$

where scalar  $Q_2$  is the first row of a  $3 \times 1$  column  $\mathbf{Q}$ :

$$\mathbf{Q} = \text{plim}(\sigma_5(\Lambda R \Lambda)^{-1} \Lambda \omega) = \text{plim}(\sigma_5 \Lambda^{-1} R^{-1} \omega), \quad (8)$$

where  $\Lambda = \begin{pmatrix} \sigma_2 & 0 & 0 \\ 0 & \sigma_3 & 0 \\ 0 & 0 & \sigma_4 \end{pmatrix}$ ,  $R = \begin{pmatrix} 1 & r_{23} & r_{24} \\ r_{23} & 1 & r_{34} \\ r_{24} & r_{34} & 1 \end{pmatrix}$  and  $\omega = \begin{pmatrix} r_{25} \\ r_{35} \\ r_{45} \end{pmatrix}$ . Here,  $\sigma_j$  is the sample standard deviation of  $x_j$ ,  $j = 2, 3, 4, 5$ , and  $r_{kl}$  is the sample correlation between  $x_k$  and  $x_l$ ,  $k = 2, 3, 4$  and  $l = 3, 4, 5$ .

---

<sup>4</sup>Note that we are not able to estimate the intercept  $\beta_0$  from the demeaned model.