On the Effect of Imputation on the 2SLS Variance

Helmut Farbmacher[a,b,*]
[a]Department of Economics, University of Munich, Germany
[b]Munich Center for the Economics of Aging, Max Planck Society, Germany
Alexander Kann[c]
[c]Department of Economics, University of Mannheim, Germany

April 11, 2019

Abstract

Endogeneity and missing data are common issues in empirical research. We investigate how they jointly affect inference on causal parameters. Conventional methods to estimate the variance, which treat the imputed data as if it had been observed in the first place, are not reliable. We derive the asymptotic variance and propose a heteroskedasticity robust variance estimator of two-stage least squares, which takes into account the imputation. Monte Carlo simulations support our theoretical findings.

Key Words: endogeneity, instrumental variables, imputation, variance estimation

1 Introduction

Missing data is a common issue in empirical studies in economics and social sciences. A popular method to handle this issue is the complete case approach, which excludes incomplete observations from the analysis. Among others, an alternative approach is regression imputation, which uses the complete observations to fill in the missing values. The imputed data is then used as if it had been observed in the first place.

While the imputation of an exogenous regressor is a well-researched topic (e.g., Little, 1992), little is known about the imputation of an endogenous regressor. Two-stage least squares (2SLS) estimation is a way to deal with the endogeneity of a regressor. McDonough and Millimet (2017) discussed the bias of 2SLS with imputation. Their analysis was mainly based on the seminal work of Nagar (1959) on the finite-sample bias of 2SLS. However, they did not discuss the estimation of the variance, which is challenging, as we show in this study. 2SLS inference is affected by imputation, and hence the conventional variance estimator, which ignores the fact that there has been an imputation, is only valid if we are interested in testing whether the parameter of the endogenous regressors is zero. Standard errors and confidence intervals based

*farbmacher@econ.lmu.de

on the conventional variance are invalid. We obtain the asymptotic distribution of 2SLS after regression imputation, and derive a heteroskedasticity robust variance estimator. This allows constructing valid standard errors, confidence intervals and conducting tests.

We focus on settings where the endogenous regressor is missing at random, the number of instruments is fixed as the sample size grows, and the imputation method is regression imputation. The missing at random setting is essentially a selection based on observables (Wooldridge, 2007), and covers many important applications (for instance, Wooldridge, 2007; Graham *et al.*, 2012; Chaudhuri *et al.*, 2018, applications with missing exogenous variables). To simplify the comparison with the conventional variance estimator, we assume missing completely at random in parts of this study.

We illustrate our theoretical results using Monte Carlo simulations. In the simulation results we focus on the key parameters in the interplay between the 2SLS estimation and the regression imputation. These key parameters are the fraction of missing data, which can be observed easily, and the direction of the OLS bias, which is unobserved, but very often researchers have strong prior beliefs about it.

In the next section, we describe the setup, including the model, the structure of the missing data, and a brief description of the regression imputation. In Section 3, we derive the asymptotic variance and propose a variance estimator which takes into account the use of imputation and which is robust to general forms of heteroskedasticity. Section 4 presents the results of the Monte Carlo simulations and Section 5 presents a summary of the paper.

2 Setup

2.1 Model

Consider the standard simultaneous equation model

$$\begin{aligned} y_i &= x_i\beta + u_i \\ x_i &= Z_{i.}'\pi + v_i, \end{aligned} \tag{1}$$

with dependent variable (y_i), endogenous regressor (x_i), and a set of L instruments $(Z_{i.})$.[1] The parameter of interest is β. The two-stage least squares (2SLS) estimator is a way to deal with the endogeneity of x:

$$\widehat{\beta} = (x'P_Z x)^{-1} x'P_Z y, \tag{2}$$

[1] To simplify notation, we omit exogenous control variables.

where $P_Z = Z(Z'Z)^{-1}Z'$. Throughout the paper we assume that 2SLS1–2SLS4 are fulfilled in model (1), which assures consistency and valid inference of the 2SLS estimator.

2SLS 1. $plim\left(\frac{Z'u}{n}\right) = E[Z_{i.}u_i] = 0$, $plim\left(\frac{Z'v}{n}\right) = E[Z_{i.}v_i] = 0$.

2SLS 2. $plim\left(\frac{Z'x}{n}\right) = E[Z_{i.}x_i] = Q_{Zx} \neq 0$.

2SLS 3. $plim\left(\frac{Z'Z}{n}\right) = E[Z_{i.}Z'_{i.}] = Q_{ZZ}$, *with* Q_{ZZ} *a finite and full rank matrix.*

2SLS 4. *Observations are identically and independently distributed.*

2SLS 5. *Errors are homoskedastatic,* $E[uu'|Z] = \sigma_u^2 I_n$, $E[vv'|Z] = \sigma_v^2 I_n$, $E[vu'|Z] = \sigma_{uv} I_n$, *while* σ_u^2, σ_v^2 *and* σ_{uv} *are finite.*

2.2 Missing data

The analysis is, however, complicated by the fact that the data for the endogenous regressor, x, are missing for some observations, causing them to be incomplete. Therefore, $\widehat{\beta}$ as defined above cannot be calculated with the data at hand. The subscript '1' will indicate that data is missing, that is, y_0, x_0, Z_0 will denote the complete observations and y_1, x_1, Z_1 the observations with missing values for x. Let $\widehat{p}$ be the fraction of missing data in the endogenous regressor and assume

2SLS 6. $\widehat{p} = \frac{n_1}{n} \to p < 1$ *as* $n \to \infty$,

which implies that not only the number of incomplete observations (n_1) but also the number of complete observations (n_0) increases as n increases. Without loss of generality we may assume that the first n_0 observations of the matrix Z and the vectors x and y are complete observations and the remaining n_1 observations are incomplete.

The literature on missing data distinguishes three types of missing-data mechanisms: missing completely at random (MCAR), missing at random (MAR)[2], and not missing at random (NMAR). While MCAR is the easiest to deal with, real data is often NMAR or MAR. The structure of the missing data is called ignorable if the data is either MCAR or MAR. Since dealing with NMAR is very different from dealing with the other two types, we focus on missing (completely) at random in this exposition, and use the following additional assumption.

2SLS 7. *The structure of the missing data is ignorable.*

[2]That is, the structure of the missing data depends only on the observables.

2.3 Regression imputation

Regression imputation (RI), which is a two-step procedure, can be applied as a tool to fill in the missing values. In the first step the complete observations are used to regress the endogenous variable (x_0) on the imputation variables to obtain the imputation parameters. These parameters are equal to the first stage estimates from the complete case approach $\left(\widehat{\pi}_{CC} = (Z_0'Z_0)^{-1}Z_0'x_0\right)$ if the imputation method incorporates the instruments.[3] In the second step these estimates are employed to impute the missing values for x_1 by multiplying the imputation variables for the incomplete observations with the parameters obtained from the first step. The imputed variable is then used in the 2SLS estimation as if it had been observed in the first place.[4]
The model in (1) can be restated in terms of $\widetilde{x}$ by adding an imputation error:

$$\begin{aligned} y &= \widetilde{x}\beta + u - e\beta = \widetilde{x}\beta + \widetilde{u} \\ \widetilde{x} &= Z\pi + v + e = Z\pi + \widetilde{v} \\ \widetilde{x} &= x + e = \begin{pmatrix} x_0 \\ x_1 \end{pmatrix} + \begin{pmatrix} 0 \\ e_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ \widetilde{x}_1 \end{pmatrix}. \end{aligned} \tag{3}$$

The imputation error e_1 and the composite errors of the imputed model ($\widetilde{v}$ & $\widetilde{u}$) are defined by

$$\begin{aligned} e_1 &= Z_1(\widehat{\pi}_{CC} - \pi) - v_1 = Z_1(Z_0'Z_0)^{-1}Z_0'v_0 - v_1 \\ \widetilde{v} &= \begin{pmatrix} v_0 \\ \widetilde{v}_1 \end{pmatrix} = \begin{pmatrix} v_0 \\ v_1 + e_1 \end{pmatrix} = \begin{pmatrix} v_0 \\ Z_1(Z_0'Z_0)^{-1}Z_0'v_0 \end{pmatrix} \\ \widetilde{u} &= \begin{pmatrix} u_0 \\ \widetilde{u}_1 \end{pmatrix} = \begin{pmatrix} u_0 \\ u_1 - e_1\beta \end{pmatrix} = \begin{pmatrix} u_0 \\ u_1 + v_1\beta - Z_1(Z_0'Z_0)^{-1}Z_0'v_0\beta \end{pmatrix}. \end{aligned} \tag{4}$$

While the 2SLS estimator with regression imputation remains consistent if 2SLS 1 to 3 and 6 to 7 are fulfilled, the inference is affected by the imputation. The independence across observations is violated in the imputed model since the complete data has been used to impute the incomplete observations. In the next section, we derive a variance estimator which takes this into account.

3 Estimation and Inference

We consider

$$\widehat{\beta}_{RI} = (\widetilde{x}'P_Z\widetilde{x})^{-1}\widetilde{x}'P_Z y, \tag{5}$$

[3] McDonough and Millimet (2017) show in Monte Carlo simulations that imputation methods that incorporate the instruments produce the smallest finite sample bias of 2SLS estimation.

[4] Clearly, the observations with missing values cannot add any information to the estimation of the first stage parameters. Hence, the relevant F-statistic for the first stage parameters should be based on the complete case observations.

and derive its limiting distribution under heteroskedasticity in the next proposition.

Proposition 1. *Suppose that 2SLS 1–4 and 2SLS 6–7 hold, and that* $E\left[y_i^4\right] < \infty$, $E\left[\widetilde{x}_i^4\right] < \infty$, *and* $E\left[\|Z_{i.}\|^4\right] < \infty$, *then we have*

$$\sqrt{n}(\widehat{\beta}_{RI} - \beta) \xrightarrow{d} N(0, V_{\widehat{\beta}_{RI}}) \tag{6}$$

with the asymptotic variance given by

$$V_{\widehat{\beta}_{RI}} = \left(Q_{xZ}Q_{ZZ}^{-1}Q_{Zx}\right)^{-1} Q_{xZ}Q_{ZZ}^{-1}\Omega\, Q_{ZZ}^{-1}Q_{Zx}\left(Q_{xZ}Q_{ZZ}^{-1}Q_{Zx}\right)^{-1}$$

where

$$\begin{aligned}\Omega =& E[u_i^2 Z_{i.}Z_{i.}'] - 2pE[u_{0i}v_{0i}Z_{0i.}Z_{0i.}']Q_{Z_0Z_0}^{-1}Q_{Z_1Z_1}\beta \\ &+ \frac{p^2}{1-p}Q_{Z_1Z_1}Q_{Z_0Z_0}^{-1}E[v_{0i}^2 Z_{0i.}Z_{0i.}']Q_{Z_0Z_0}^{-1}Q_{Z_1Z_1}\beta^2 \\ &+ 2pE[u_{1i}v_{1i}Z_{1i.}Z_{1i.}']\beta + pE[v_{1i}^2 Z_{1i.}Z_{1i.}']\beta^2\end{aligned}$$

Proof. See Appendix A.1. □

The first term of Ω corresponds to the standard asymptotic variance of 2SLS without missing data. The remaining terms can be attributed to regression imputation. If no data is missing ($p = 0$) or $\beta = 0$, Ω collapses to the standard asymptotic variance of 2SLS. The latter simplification (i.e., $\beta = 0$) is a well-known result in the literature about generated regressors (see, for example, Murphy and Topel, 1985). It allows valid inference based on the conventional variance estimator if, and only if, we are interested in tests of $\beta = 0$, which often is of major interest in applications. However, it does not justify the construction of standard errors or confidence intervals.

The estimation of Ω is challenging since we cannot obtain reliable residual estimates for the imputed observations u_1 and v_1. Hence, the first, fourth and fifth terms of Ω cannot be simply estimated by using the corresponding residuals. However, it is possible to circumvent this issue by using the error of the imputed model ($\widetilde{u}_1$), for which residuals ($\widehat{\widetilde{u}}_1$) can be obtained. The variance estimator given in the following proposition is consistent under general forms of heteroskedasticity.

Proposition 2. *Suppose that 2SLS 1–4 and 2SLS 6–7 hold, and that* $E\left[y_i^4\right] < \infty$, $E\left[\widetilde{x}_i^4\right] < \infty$, *and* $E\left[\|Z_{i.}\|^4\right] < \infty$, *then we have*

$$\widehat{V}_{\widehat{\beta}_{RI}} = \left(\widetilde{x}'P_Z\widetilde{x}\right)^{-1}\widetilde{x}'Z(Z'Z)^{-1}\widehat{W}_{RI}(Z'Z)^{-1}Z'\widetilde{x}\left(\widetilde{x}'P_Z\widetilde{x}\right)^{-1}$$

is a consistent estimator of the asymptotic variance ($n\widehat{V}_{\widehat{\beta}_{RI}} \xrightarrow{p} V_{\widehat{\beta}_{RI}}$) with

$$\begin{aligned}\widehat{W}_{RI} = &\left(\sum_{i=1}^{n} \widehat{\widetilde{u}}_i^2 Z_{i.} Z_{i.}'\right) - 2\left(\sum_{i=1}^{n_0} \widehat{\widetilde{u}}_i \widehat{\widetilde{v}}_i Z_{i.} Z_{i.}'\right)\left(\sum_{i=1}^{n_0} Z_{i.} Z_{i.}'\right)^{-1}\left(\sum_{i=n_0+1}^{n} Z_{i.} Z_{i.}'\right)\widehat{\beta}_{RI} \\ &+ \left[\left(\sum_{i=n_0+1}^{n} Z_{i.} Z_{i.}'\right)\left(\sum_{i=1}^{n_0} Z_{i.} Z_{i.}'\right)^{-1}\left(\sum_{i=1}^{n_0} \widehat{\widetilde{v}}_i^2 Z_{i.} Z_{i.}'\right)\left(\sum_{i=1}^{n_0} Z_{i.} Z_{i.}'\right)^{-1}\left(\sum_{i=n_0+1}^{n} Z_{i.} Z_{i.}'\right)\right. \\ &\left. - \sum_{i=n_0+1}^{n} Z_{i.} Z_{i.}'\left(\sum_{i=1}^{n_0} Z_{i.} Z_{i.}'\right)^{-1}\left(\sum_{i=1}^{n_0} \widehat{\widetilde{v}}_i^2 Z_{i.} Z_{i.}'\right)\left(\sum_{i=1}^{n_0} Z_{i.} Z_{i.}'\right)^{-1} Z_{i.} Z_{i.}'\right]\widehat{\beta}_{RI}^2\end{aligned}$$

where $\widehat{\widetilde{u}}_i = y_i - \widetilde{x}_i\widehat{\beta}_{RI}$ and $\widehat{\widetilde{v}}_i = \widetilde{x}_i - Z_{i.}'\widehat{\pi}_{CC}$

Proof. See Appendix A.2. □

In the following we compare the conventional variance estimator, which ignores the fact that there has been an imputation, with the true asymptotic variance. To simplify the comparison, we assume homoskedasticity and MCAR. The asymptotic variance is then stated in the following corollary.

Corollary 1. *Under the conditions of Proposition 1, 2SLS 5 and MCAR, the asymptotic variance is given by*

$$V_{\widehat{\beta}_{RI}}^{hom} = \left(Q_{xZ}Q_{ZZ}^{-1}Q_{Zx}\right)^{-1}\sigma_u^2 + \left(Q_{xZ}Q_{ZZ}^{-1}Q_{Zx}\right)^{-1}\left(\frac{p}{1-p}\right)\sigma_v^2\beta^2.$$

Proof. See Appendix A.3. □

The conventional variance estimator calculated by standard statistical software is given by

$$\widehat{V}_{\widehat{\beta}_{RI}}^{naive} = \left(\widetilde{x}'P_Z\widetilde{x}\right)^{-1}\widehat{\sigma}_{\widetilde{u}}^2, \quad with\ \widehat{\sigma}_{\widetilde{u}}^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{\widetilde{u}}_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widetilde{x}_i\widehat{\beta}_{RI})^2. \tag{7}$$

and its limit under homoskedasticity and MCAR is

$$n\widehat{V}_{\widehat{\beta}_{RI}}^{naive} \xrightarrow{p} \left(Q_{xZ}Q_{ZZ}^{-1}Q_{Zx}\right)^{-1}\sigma_u^2 + \left(Q_{xZ}Q_{ZZ}^{-1}Q_{Zx}\right)^{-1} p\left(2\sigma_{uv}\beta + \sigma_v^2\beta^2\right). \tag{8}$$

Comparing the limit of the conventional estimator with the true asymptotic variance in Corollary 1, we can see that it does not reflect the true variance of $\widehat{\beta}_{RI}$. For instance, while the asymptotic variance always increases with p, the conventional estimator could even decrease with p if $-2\sigma_{uv}\beta > \sigma_v^2\beta^2$. In this case, its limit is smaller than the asymptotic variance without any missing data problem, which is clearly counterintuitive. Moreover, the degree of endogeneity (σ_{uv}) erroneously affects the conventional limit, whereas it has indeed no effect on the true

asymptotic variance.

4 Monte Carlo Simulation

We illustrate our theoretical findings with Monte Carlo simulations. To implement the missing data problem, we first mimic the model in (1) using the data generating process described below, and then randomly delete the value x_i with probability p. The missing-data mechanism is thus MCAR. For each of the n observations, $Z_{i.}$ and v_i are drawn from

$$Z_{i.} \sim N\left(0_L, \frac{1}{L} I_L\right) \; ; \quad v_i \sim N(0,1),$$

where L denotes the number of instruments. We follow Hausman *et al.* (2012) and Chao *et al.* (2014) and define the structural error as

$$u_i = \sigma_{uv} v_i + \sqrt{\frac{1-\sigma_{uv}^2}{\phi + (0.86^2)}}(\phi \epsilon_{1i} + 0.86\epsilon_{2i}), \;\; \epsilon_{1i} \sim N(0, Z_{i.}' Z_{i.}), \; \epsilon_{2i} \sim N(0, 0.86^2), \; \phi = 5,$$

where ϕ defines the strength of the heteroskedasticity. We set $\beta = 0.5$, $L = 3$, $n = 1000$, the number of Monte Carlo repetitions $R = 5000$, and the step size that we use to alter p in the different simulations to $\Delta p = 0.005$. As derived in the previous section, the sign of the endogeneity is crucial. Hence, we show the results for $\sigma_{uv} = 0.3$ and $\sigma_{uv} = -0.3$. We set $\pi = \sqrt{\frac{FL}{n}}$. Note that we divide by n and not by n_0. Therefore, F defines the first-stage F-statistic in the entire sample containing both complete and incomplete observations, and the first-stage F-statistic in the complete case sample gradually decreases with increasing p. We set $F = 100$. That is, the complete case F-statistic is around 100 for $p = 1$ and around 20 for $p = 0.8$.

Figure 1 compares our consistent and the conventional standard errors with the observed root mean squared error (RMSE) obtained from the Monte Carlo simulations. It shows that the conventional estimator cannot properly describe the standard error of the 2SLS estimator with regression imputation. The probability p has a linear effect on the conventional estimator (see Eq. 8) while it has a nonlinear effect on the true asymptotic variance, as shown in Corollary 1. Moreover, as expected, the conventional standard error can even decrease as p increases ($-2\sigma_{uv}\beta > \sigma_v^2\beta^2$ in the right graph of Figure 1). The reason why McDonough and Millimet (2017), who did not derive a consistent variance estimator for regression imputation, have acceptable results in their simulations is solely due to the values they chose for the parameters ($\sigma_{uv} > 0$, $\beta > 0$, low p) and cannot be generalized to the entire parameter space. Our variance estimator ($\widehat{V}_{\widehat{\beta}_{RI}}$) performs well in both settings. Additionally, Figure 2 shows how often the null hypothesis ($H_0 : \beta = 0.5$) is rejected at the 5% nominal level. Again, our proposed variance

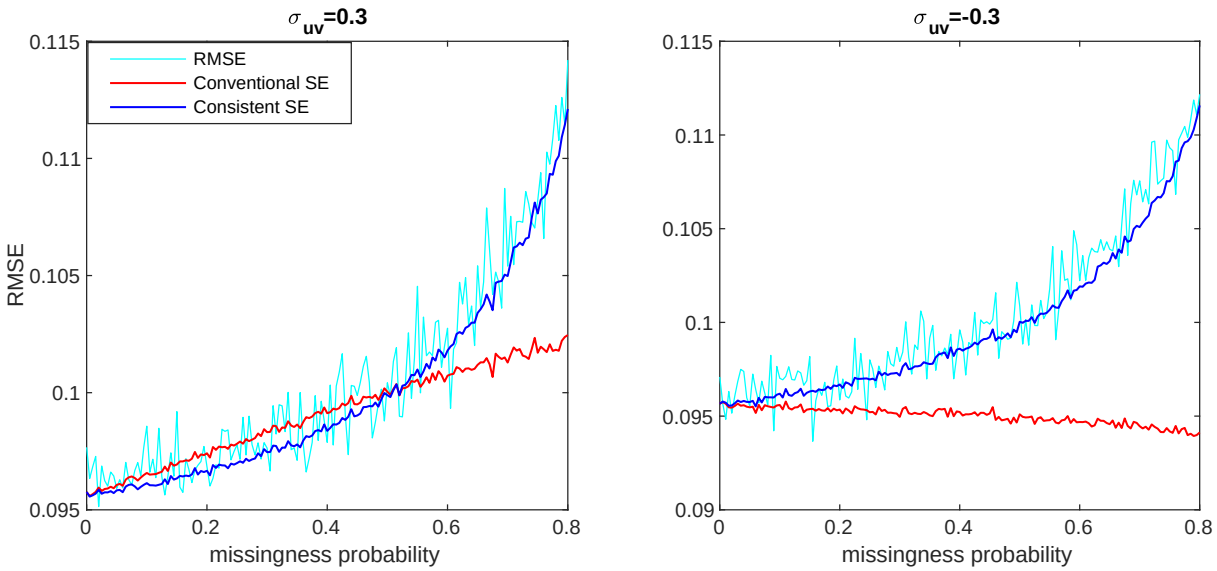

Figure 1: Consistent vs naive variance estimator (RMSE)

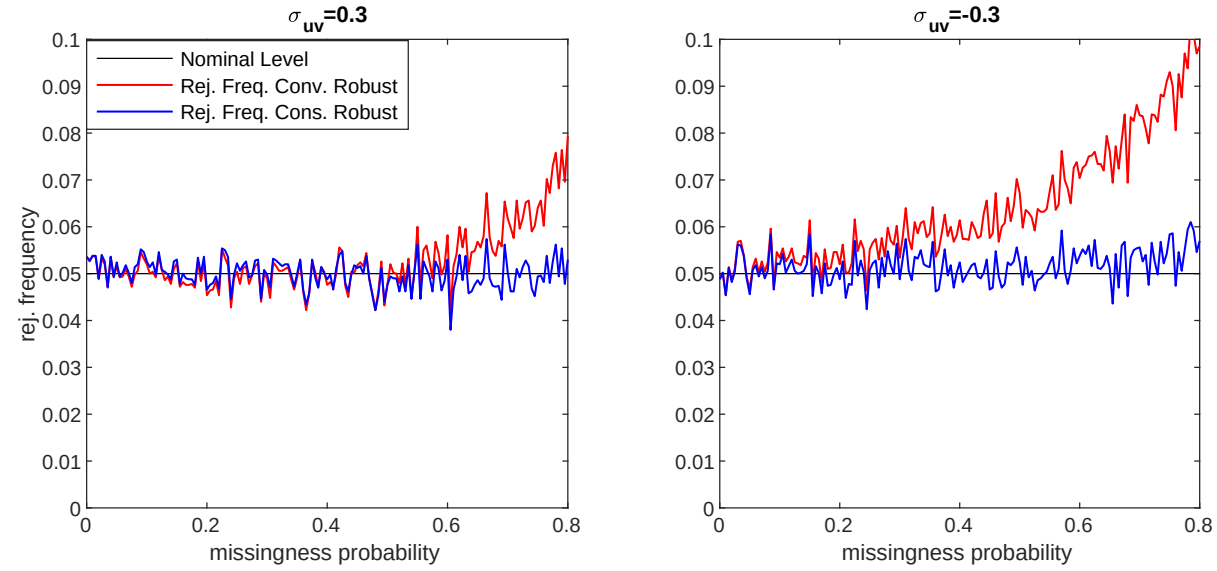

Figure 2: Consistent vs naive variance estimator (Rejection Frequency)

estimator performs better than the conventional one – particularly if $\sigma_{uv} < 0$.

5 Conclusion

We have investigated how two issues, which are likely to be present in many empirical studies, affect the estimation of causal effects, namely missing data and the endogeneity of the regressors. If researchers use an instrumental variables regression after single imputation of missing

values for an endogenous regressor, they have to be aware that conventional methods to estimate the variance fail to take into account the fact that there was an imputation. The asymptotic variance of 2SLS is affected by the imputation, implying that conventional methods cannot be used to construct standard errors, confidence intervals, or carry out tests. We have derived a heteroskedastic variance estimator which takes the imputation into account and is consistent. Monte Carlo simulations show that our estimator performs well while the conventional variance estimator does not.

References

Chao JC, Hausman JA, Newey WK, Swanson NR, Woutersen T, 2014. Testing overidentifying restrictions with many instruments and heteroskedasticity. *Journal of Econometrics* **178**: 15–21

Chaudhuri S, Frazier DT, Ranault E, 2018. Indirect Inference with endogenously missing exogenous variables. *Journal of Econometrics* **205**(1): 55–75

Graham BS, Pinto C, Egel D, 2012. Inverse Probability Tilting for Moment Condition Models with Missing Data. *Review of Economic Studies* **79**(3): 1053–1079

Hausman JA, Newey WK, Woutersen T, Chao JC, Swanson NR, 2012. Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics* **3**(2): 211–255

Little RJA, 1992. Regression With Missing X's: A Review. *Journal of the American Statistical Association* **87**(420): 1227–1237

McDonough IK, Millimet DL, 2017. Missing data, imputation, and endogeneity. *Journal of Econometrics* **199**(2): 141 – 155

Murphy KM, Topel RH, 1985. Estimation and Inference in Two-Step Econometric Models. *Journal of Business and Economic Statistics* **3**: 370–379

Nagar AL, 1959. The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations. *Econometrica* **27**(4): 575–595

Wooldridge JM, 2007. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics* **141**(2): 1281–1301

A Appendix

A.1 Proof of Proposition 1

We begin from

$$\sqrt{n}(\widehat{\beta}_{RI} - \beta) = \left(\left(\frac{\widetilde{x}'Z}{n} \right) \left(\frac{Z'Z}{n} \right)^{-1} \left(\frac{Z'\widetilde{x}}{n} \right) \right)^{-1} \left(\left(\frac{\widetilde{x}'Z}{n} \right) \left(\frac{Z'Z}{n} \right)^{-1} \left(\frac{Z'\widetilde{u}}{\sqrt{n}} \right) \right). \tag{9}$$

From 2SLS 2 and 2SLS 3, we have $\widetilde{x}'P_Z\widetilde{x}/n \xrightarrow{p} Q_{xZ}Q_{ZZ}^{-1}Q_{Zx}$, $\widetilde{x}'Z/n \xrightarrow{p} Q_{xZ}$ and $Z'Z/n \xrightarrow{p} Q_{ZZ}$. The same holds for the complete and incomplete part of the sample due to 2SLS 6, e.g. $\frac{1}{n_0}\sum_{i=1}^{n_0} Z_{i.}Z_{i.}' = Z_0'Z_0/n_0 \xrightarrow{p} Q_{Z_0Z_0}$. Hence, it remains to determine how $Z'\widetilde{u}/\sqrt{n}$ behaves as $n \xrightarrow{p} \infty$.

$$\begin{aligned}
\frac{Z'\widetilde{u}}{\sqrt{n}} =& \frac{Z_0'u_0}{\sqrt{n}} + \frac{Z_1'\widetilde{u}_1}{\sqrt{n}} \\
=& \frac{Z_0'u_0}{\sqrt{n}} + \frac{Z_1'u_1}{\sqrt{n}} + \frac{Z_1'v_1}{\sqrt{n}}\beta - \frac{n_1}{n_0}\frac{Z_1'Z_1}{n_1}\left(\frac{Z_0'Z_0}{n_0}\right)^{-1}\frac{Z_0'v_0}{\sqrt{n}}\beta \\
=& \frac{\sqrt{n_0}}{\sqrt{n}}\left(\frac{Z_0'u_0}{\sqrt{n_0}} - \frac{n_1}{n_0}\frac{Z_1'Z_1}{n_1}\left(\frac{Z_0'Z_0}{n_0}\right)^{-1}\frac{Z_0'v_0}{\sqrt{n_0}}\beta \right) + \frac{\sqrt{n_1}}{\sqrt{n}}\left(\frac{Z_1'u_1}{\sqrt{n_1}} + \frac{Z_1'v_1}{\sqrt{n_1}}\beta \right) \\
=& \frac{\sqrt{n_0}}{\sqrt{n}}\left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\left(Z_{0i.}u_{0i} - \frac{n_1}{n_0}\frac{Z_1'Z_1}{n_1}\left(\frac{Z_0'Z_0}{n_0}\right)^{-1} Z_{0i.}v_{0i}\beta \right)\right) \\
&+ \frac{\sqrt{n_1}}{\sqrt{n}}\left(\frac{1}{\sqrt{n_1}}\sum_{i=n_0+1}^{n} Z_{1i.}(u_{1i} + v_{1i}\beta) \right)
\end{aligned} \tag{10}$$

By 2SLS 4, both terms are asymptotically independent, and so the CLT implies

$$\frac{\sqrt{n_0}}{\sqrt{n}}\left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\left(Z_{0i.}u_{0i} - \frac{n_1}{n_0}\frac{Z_1'Z_1}{n_1}\left(\frac{Z_0'Z_0}{n_0}\right)^{-1} Z_{0i.}v_{0i}\beta \right)\right) \xrightarrow{d} N(0_L, \Omega_0)$$

and

$$\frac{\sqrt{n_1}}{\sqrt{n}}\left(\frac{1}{\sqrt{n_1}}\sum_{i=n_0+1}^{n} Z_{1i.}(u_{1i} + v_{1i}\beta) \right) \xrightarrow{d} N(0_L, \Omega_1).$$

The asymptotic variances Ω_0 and Ω_1 are given by

$$\begin{aligned}\Omega_0 =&(1-p)E[Z_{0i.}Z_{0i.}'u_{0i}^2] - 2pE[u_{0i}v_{0i}Z_{0i.}Z_{0i.}']Q_{Z_0Z_0}^{-1}Q_{Z_1Z_1}\beta \\ &+ \frac{p^2}{1-p}Q_{Z_1Z_1}Q_{Z_0Z_0}^{-1}E[v_{0i}^2Z_{0i.}Z_{0i.}']Q_{Z_0Z_0}^{-1}Q_{Z_1Z_1}\beta^2\end{aligned}$$

and

$$\Omega_1 = p\left(E[Z_{1i.}Z_{1i.}'u_{1i}^2] + 2E[u_{1i}v_{1i}Z_{1i.}Z_{1i.}']\beta + E[Z_{1i.}Z_{1i.}'v_{1i}^2]\beta^2\right).$$

Combing both of these gives the limiting distribution of $\frac{Z'\widetilde{u}}{\sqrt{n}}$:

$$\begin{aligned}&\frac{Z'\widetilde{u}}{\sqrt{n}} \xrightarrow{p} N(0_L, \Omega) \\ &where\ \Omega =E[Z_{i.}Z_{i.}'u_i^2] - 2pE[u_{0i}v_{0i}Z_{0i.}Z_{0i.}']Q_{Z_0Z_0}^{-1}Q_{Z_1Z_1}\beta \\ &\qquad + \frac{p^2}{1-p}Q_{Z_1Z_1}Q_{Z_0Z_0}^{-1}E[v_{0i}^2Z_{0i.}Z_{0i.}']Q_{Z_0Z_0}^{-1}Q_{Z_1Z_1}\beta^2 \\ &\qquad + 2pE[u_{1i}v_{1i}Z_{1i.}Z_{1i.}']\beta + pE[Z_{1i.}Z_{1i.}'v_{1i}^2]\beta^2,\end{aligned} \tag{11}$$

where we used the fact that $(1-p)E[Z_{0i.}Z_{0i.}'u_{0i}^2] + pE[Z_{1i.}Z_{1i.}'u_{1i}^2] = E[Z_{i.}Z_{i.}'u_i^2]$. Using Slutsky's Lemma, we can then prove Proposition 1:

$$\begin{aligned}&\sqrt{n}(\widehat{\beta}_{RI} - \beta) \xrightarrow{d} \left(Q_{xZ}Q_{ZZ}^{-1}Q_{Zx}\right)^{-1}Q_{xZ}Q_{ZZ}^{-1}N(0_L, \Omega) = N(0, V_{\widehat{\beta}_{RI}}) \\ &\qquad where\ V_{\widehat{\beta}_{RI}} = \left(Q_{xZ}Q_{ZZ}^{-1}Q_{Zx}\right)^{-1}Q_{xZ}Q_{ZZ}^{-1}\,\Omega\,Q_{ZZ}^{-1}Q_{Zx}\left(Q_{xZ}Q_{ZZ}^{-1}Q_{Zx}\right)^{-1}.\end{aligned} \tag{12}$$

A.2 Proof of Proposition 2

We begin from

$$n\widehat{V}_{\widehat{\beta}_{RI}} = \left(\frac{\widetilde{x}'P_Z\widetilde{x}}{n}\right)^{-1}\frac{\widetilde{x}'Z}{n}\left(\frac{Z'Z}{n}\right)^{-1}\frac{\widehat{W}_{RI}}{n}\left(\frac{Z'Z}{n}\right)^{-1}\frac{Z'\widetilde{x}}{n}\left(\frac{\widetilde{x}'P_Z\widetilde{x}}{n}\right)^{-1}. \tag{13}$$

From 2SLS 2 and 2SLS 3, we have $\widetilde{x}'P_Z\widetilde{x}/n \xrightarrow{p} Q_{xZ}Q_{ZZ}^{-1}Q_{Zx}$, $\widetilde{x}'Z/n \xrightarrow{p} Q_{xZ}$ and $Z'Z/n \xrightarrow{p} Q_{ZZ}$. The same holds for the complete and incomplete parts of the sample due to 2SLS 6, e.g.

$\frac{1}{n_0}\sum_{i=1}^{n_0} Z_{i.}Z_{i.}' = Z_0'Z_0/n_0 \xrightarrow{p} Q_{Z_0Z_0}$. Hence, it remains to show that $\widehat{W}_{RI}/n = \widehat{\Omega} \xrightarrow{p} \Omega$ as $n \xrightarrow{p} \infty$.

$$
\begin{aligned}
\widehat{\Omega} = &\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\widetilde{u}}_i^2 Z_{i.}Z_{i.}'\right) - 2\frac{n_1}{n}\left(\frac{1}{n_0}\sum_{i=1}^{n_0}\widehat{\widetilde{u}}_i\widehat{\widetilde{v}}_i Z_{i.}Z_{i.}'\right)\left(\frac{1}{n_0}\sum_{i=1}^{n_0} Z_{i.}Z_{i.}'\right)^{-1}\left(\frac{1}{n_1}\sum_{i=n_0+1}^{n} Z_{i.}Z_{i.}'\right)\widehat{\beta}_{RI} \\
&+ \frac{n_1^2}{nn_0}\left(\frac{1}{n_1}\sum_{i=n_0+1}^{n} Z_{i.}Z_{i.}'\right)\left(\frac{1}{n_0}\sum_{i=1}^{n_0} Z_{i.}Z_{i.}'\right)^{-1}\left(\frac{1}{n_0}\sum_{i=1}^{n_0}\widehat{\widetilde{v}}_i^2 Z_{i.}Z_{i.}'\right)\left(\frac{1}{n_0}\sum_{i=1}^{n_0} Z_{i.}Z_{i.}'\right)^{-1}\left(\frac{1}{n_1}\sum_{i=n_0+1}^{n} Z_{i.}Z_{i.}'\right)\widehat{\beta}_{RI}^2 \\
&- \frac{n_1}{n}\left(\frac{1}{n_0}\right)\left(\frac{1}{n_1}\sum_{i=n_0+1}^{n} Z_{i.}Z_{i.}'Z_{i.}Z_{i.}'\right)\left(\frac{1}{n_0}\sum_{i=1}^{n_0} Z_{i.}Z_{i.}'\right)^{-1}\left(\frac{1}{n_0}\sum_{i=1}^{n_0}\widehat{\widetilde{v}}_i^2 Z_{i.}Z_{i.}'\right)\left(\frac{1}{n_0}\sum_{i=1}^{n_0} Z_{i.}Z_{i.}'\right)^{-1}\widehat{\beta}_{RI}^2
\end{aligned}
\tag{14}
$$

where $\widehat{\widetilde{u}}_i = y_i - \widetilde{x}_i\widehat{\beta}_{RI}$ *and* $\widehat{\widetilde{v}}_i = \widetilde{x}_i - Z_{i.}'\widehat{\pi}_{CC}$.

In the next steps, it is useful to rewrite the composite residuals as $\widehat{\widetilde{u}}_i = \widetilde{u}_i - \widetilde{x}_i(\widehat{\beta}_{RI} - \beta)$ and $\widehat{\widetilde{v}}_i = \widetilde{v}_i - Z_{i.}'(\widehat{\pi}_{CC} - \pi)$. Rewriting the first term of Eq. (14), we get

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\widehat{\widetilde{u}}_i^2 Z_{i.}Z_{i.}' &= \left(\frac{1}{n}\sum_{i=1}^{n}(\widetilde{u}_i - \widetilde{x}_i(\widehat{\beta}_{RI} - \beta))^2 Z_{i.}Z_{i.}'\right) \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{u}_i^2 Z_{i.}Z_{i.}'\right) - 2\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{u}_i\widetilde{x}_i(\widehat{\beta}_{RI} - \beta)Z_{i.}Z_{i.}'\right) + \left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i^2(\widehat{\beta}_{RI} - \beta)^2 Z_{i.}Z_{i.}'\right) \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{u}_i^2 Z_{i.}Z_{i.}'\right) - 2(\widehat{\beta}_{RI} - \beta)\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{u}_i\widetilde{x}_i Z_{i.}Z_{i.}'\right) + (\widehat{\beta}_{RI} - \beta)^2\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i^2 Z_{i.}Z_{i.}'\right)
\end{aligned}
\tag{15}
$$

For the first term of Eq. (15), we have

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\widetilde{u}_i^2 Z_{i.}Z_{i.}' =& \frac{n_0}{n}\left(\frac{1}{n_0}\sum_{i=1}^{n_0} u_i^2 Z_{i.}Z_{i.}'\right) + \frac{n_1}{n}\left(\frac{1}{n_1}\sum_{i=n_0+1}^{n}\widetilde{u}_i^2 Z_{i.}Z_{i.}'\right) \\
=& \frac{n_0}{n}\left(\frac{1}{n_0}\sum_{i=1}^{n_0} u_i^2 Z_{i.}Z_{i.}'\right) + \frac{n_1}{n}\left(\frac{1}{n_1}\sum_{i=n_0+1}^{n} u_i^2 Z_{i.}Z_{i.}'\right) \\
&+ 2\frac{n_1}{n}\left(\frac{1}{n_1}\sum_{i=n_0+1}^{n} v_i u_i Z_{i.}Z_{i.}'\right)\beta + \left(\frac{1}{n_1}\sum_{i=n_0+1}^{n} v_i^2 Z_{i.}Z_{i.}'\right)\beta^2 \\
&- 2\frac{n_1}{n}\left(\frac{1}{n_1}\sum_{i=n_0+1}^{n}\widetilde{u}_i Z_{i.}'(Z_0'Z_0)^{-1}Z_0'v_0 Z_{i.}Z_{i.}'\right)\beta \\
&- \frac{n_1}{n}\left(\frac{1}{n_1}\sum_{i=n_0+1}^{n} Z_{i.}Z_{i.}'(Z_0'Z_0)^{-1}Z_0'v_0 v_0' Z_0(Z_0'Z_0)^{-1}Z_{i.}Z_{i.}'\right)\beta^2 \\
\xrightarrow{p}& E[u_i^2 Z_{i.}Z_{i.}'] + 2pE[u_{1i}v_{1i}Z_{1i.}Z_{1i.}']\beta + pE[v_{1i}^2 Z_{1i.}Z_{1i.}']\beta^2
\end{aligned}
\tag{16}
$$

Using $(Z_0'Z_0)^{-1}Z_0'v_0 = \widehat{\pi}_{CC} - \pi$, the triangle inequality, and Hölder's inequality, we get for the next to last term of Eq. (16)

$$\begin{aligned}\left\|\frac{1}{n_1}\sum_{i=n_0+1}^{n} Z_{i.}Z_{i.}'\widetilde{u}_i Z_{i.}'(\widehat{\pi}_{CC}-\pi)\right\| &\leq \frac{1}{n_1}\sum_{i=n_0+1}^{n}\left\|Z_{i.}Z_{i.}'\widetilde{u}_i Z_{i.}'(\widehat{\pi}_{CC}-\pi)\right\| \\ &\leq \frac{1}{n_1}\sum_{i=n_0+1}^{n}\|Z_{i.}\|^3|\widetilde{u}_i|\|\widehat{\pi}_{CC}-\pi\| = o_p(1),\end{aligned}$$

as $\|\widehat{\pi}_{CC}-\pi\| \xrightarrow{p} 0$ and $E[\|Z_{i.}\|^3|\widetilde{u}_i|] \leq E[\|Z_{i.}\|^4]^{\frac{3}{4}}E[|\widetilde{u}_i|^4]^{\frac{1}{4}} < \infty$ from the finiteness of $E[\|Z_{i.}\|^4]$, $E[y_i^4]$ and $E[\widetilde{x}_i^4]$. Similarly, we have for the last term of Eq. (16)

$$\begin{aligned}\left\|\frac{1}{n_1}\sum_{i=n_0+1}^{n} Z_{i.}Z_{i.}'Z_{i.}Z_{i.}'(\widehat{\pi}_{CC}-\pi)(\widehat{\pi}_{CC}-\pi)'\right\| &\leq \frac{1}{n_1}\sum_{i=n_0+1}^{n}\left\|Z_{i.}Z_{i.}'Z_{i.}Z_{i.}'(\widehat{\pi}_{CC}-\pi)(\widehat{\pi}_{CC}-\pi)'\right\| \\ &\leq \frac{1}{n_1}\sum_{i=n_0+1}^{n}\|Z_{i.}\|^4\|\widehat{\pi}_{CC}-\pi\|^2 = o_p(1).\end{aligned}$$

Now, we show that the last two terms of Eq. (15) converge to zero as $n \xrightarrow{p} \infty$.

$$\begin{aligned}\left\|(\widehat{\beta}_{RI}-\beta)\frac{1}{n}\sum_{i=1}^{n}\widetilde{u}_i\widetilde{x}_i Z_{i.}Z_{i.}'\right\| &\leq |\widehat{\beta}_{RI}-\beta|\frac{1}{n_1}\sum_{i=n_0+1}^{n}\left\|\widetilde{u}_i\widetilde{x}_i Z_{i.}Z_{i.}'\right\| \\ &\leq |\widehat{\beta}_{RI}-\beta|\frac{1}{n_1}\sum_{i=n_0+1}^{n}|\widetilde{u}_i||\widetilde{x}_i|\|Z_{i.}\|^2 = o_p(1),\end{aligned}$$

and

$$\begin{aligned}\left\|(\widehat{\beta}_{RI}-\beta)^2\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i^2 Z_{i.}Z_{i.}'\right\| &\leq |\widehat{\beta}_{RI}-\beta|^2\frac{1}{n_1}\sum_{i=n_0+1}^{n}\left\|\widetilde{x}_i^2 Z_{i.}Z_{i.}'\right\| \\ &\leq |\widehat{\beta}_{RI}-\beta|^2\frac{1}{n_1}\sum_{i=n_0+1}^{n}|\widetilde{x}_i|^2\|Z_{i.}\|^2 = o_p(1),\end{aligned}$$

since $|\widehat{\beta}_{RI}-\beta| \xrightarrow{p} 0$, $E[|\widetilde{u}_i||\widetilde{x}_i|\|Z_{i.}\|^2] \leq E[|\widetilde{u}_i|^4]^{\frac{1}{4}}E[|\widetilde{x}_i|^4]^{\frac{1}{4}}E[\|Z_{i.}\|^4]^{\frac{2}{4}} < \infty$, and $E[|\widetilde{x}_i|^2\|Z_{i.}\|^2] \leq E[|\widetilde{x}_i|^4]^{\frac{1}{2}}E[\|Z_{i.}\|^4]^{\frac{1}{2}} < \infty$.
This proves that

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{\widetilde{u}}_i^2 Z_{i.}Z_{i.}' \xrightarrow{p} E[u_i^2 Z_{i.}Z_{i.}'] + 2pE[u_{1i}v_{1i}Z_{1i.}Z_{1i.}']\beta + pE[v_{1i}^2 Z_{1i.}Z_{1i.}']\beta^2 \tag{17}$$

Now, analysing the remaining terms of Eq. (14), we get

$$\begin{aligned}\frac{1}{n_0}\sum_{i=1}^{n_0}\widehat{\widetilde{u}}_i\widehat{\widetilde{v}}_i Z_{i.}Z_{i.}' =& \frac{1}{n_0}\sum_{i=1}^{n_0}(\widetilde{u}_i - \widetilde{x}_i(\widehat{\beta}_{RI}-\beta))(\widetilde{v}_i - Z_{i.}'(\widehat{\pi}_{CC}-\pi))Z_{i.}Z_{i.}' \\ =& \frac{1}{n_0}\sum_{i=1}^{n_0}\widetilde{u}_i\widetilde{v}_i Z_{i.}Z_{i.}' + \frac{1}{n_0}\sum_{i=1}^{n_0}\widetilde{x}_i(\widehat{\beta}_{RI}-\beta)\widetilde{v}_i Z_{i.}Z_{i.}' - \frac{1}{n_0}\sum_{i=1}^{n_0}\widetilde{u}_i Z_{i.}'(\widehat{\pi}_{CC}-\pi)Z_{i.}Z_{i.}' \\ &+ \frac{1}{n_0}\sum_{i=1}^{n_0}\widetilde{x}_i(\widehat{\beta}_{RI}-\beta)Z_{i.}'(\widehat{\pi}_{CC}-\pi)Z_{i.}Z_{i.}' \\ \xrightarrow{p}& E[u_{0i}v_{0i}Z_{0i.}Z_{0i.}'],\end{aligned} \tag{18}$$

and

$$\begin{aligned}\frac{1}{n_0}\sum_{i=1}^{n_0}\widehat{\widetilde{v}}_i^2 Z_{i.}Z_{i.}' =& \frac{1}{n_0}\sum_{i=1}^{n_0}\widetilde{v}_i - Z_{i.}'(\widehat{\pi}_{CC}-\pi))^2 Z_{i.}Z_{i.}' \\ =& \frac{1}{n_0}\sum_{i=1}^{n_0}\widetilde{v}_i^2 Z_{i.}Z_{i.}' - 2\frac{1}{n_0}\sum_{i=1}^{n_0}(\widetilde{v}_i Z_{i.}'(\widehat{\pi}_{CC}-\pi)Z_{i.}Z_{i.}' + \frac{1}{n_0}\sum_{i=1}^{n_0}(\widehat{\pi}_{CC}-\pi)' Z_{i.}Z_{i.}'(\widehat{\pi}_{CC}-\pi)Z_{i.}Z_{i.}' \\ \xrightarrow{p}& E[v_{0i}^2 Z_{0i.}Z_{0i.}'].\end{aligned} \tag{19}$$

The proofs are similar to the results above and omitted here.
Finally, we can show that

$$\begin{aligned}\widehat{\Omega} =& \left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\widetilde{u}}_i^2 Z_{i.}Z_{i.}'\right) - 2\frac{n_1}{n}\left(\frac{1}{n_0}\sum_{i=1}^{n_0}\widehat{\widetilde{u}}_i\widehat{\widetilde{v}}_i Z_{i.}Z_{i.}'\right)\left(\frac{1}{n_0}\sum_{i=1}^{n_0}Z_{i.}Z_{i.}'\right)^{-1}\left(\frac{1}{n_1}\sum_{i=n_0+1}^{n}Z_{i.}Z_{i.}'\right)\widehat{\beta}_{RI} \\ &+ \frac{n_1^2}{nn_0}\left(\frac{1}{n_1}\sum_{i=n_0+1}^{n}Z_{i.}Z_{i.}'\right)\left(\frac{1}{n_0}\sum_{i=1}^{n_0}Z_{i.}Z_{i.}'\right)^{-1}\left(\frac{1}{n_0}\sum_{i=1}^{n_0}\widehat{\widetilde{v}}_i^2 Z_{i.}Z_{i.}'\right)\left(\frac{1}{n_0}\sum_{i=1}^{n_0}Z_{i.}Z_{i.}'\right)^{-1}\left(\frac{1}{n_1}\sum_{i=n_0+1}^{n}Z_{i.}Z_{i.}'\right)\widehat{\beta}_{RI}^2 \\ &- \frac{n_1}{n}\left(\frac{1}{n_0}\right)\left(\frac{1}{n_1}\sum_{i=n_0+1}^{n}Z_{i.}Z_{i.}'Z_{i.}Z_{i.}'\right)\left(\frac{1}{n_0}\sum_{i=1}^{n_0}Z_{i.}Z_{i.}'\right)^{-1}\left(\frac{1}{n_0}\sum_{i=1}^{n_0}\widehat{\widetilde{v}}_i^2 Z_{i.}Z_{i.}'\right)\left(\frac{1}{n_0}\sum_{i=1}^{n_0}Z_{i.}Z_{i.}'\right)^{-1}\widehat{\beta}_{RI}^2 \\ \xrightarrow{p}& E[Z_{i.}Z_{i.}'u_i^2] - 2pE[u_{0i}v_{0i}Z_{0i.}Z_{0i.}']Q_{Z_0'Z_0}^{-1}Q_{Z_1Z_1}\beta \\ &+ \frac{p^2}{1-p}Q_{Z_1Z_1}Q_{Z_0'Z_0}^{-1}E[v_{0i}^2 Z_{0i.}Z_{0i.}']Q_{Z_0'Z_0}^{-1}Q_{Z_1Z_1}\beta^2 \\ &+ 2pE[u_{1i}v_{1i}Z_{1i.}Z_{1i.}']\beta + pE[Z_{1i.}Z_{1i.}'v_{1i}^2]\beta^2 = \Omega.\end{aligned} \tag{20}$$

The last term of the estimator converges to zero due to the n_0^{-1} term. Nevertheless, we keep it in the estimator to improve its finite sample performance as this term is essentially part of the first and third terms and hence should be deducted once. The limit is not affected by this choice.

A.3 Proof of Corollary 1

Using the results from Proposition 1 and the simplifications induced by homoskedasticity, e.g. $E[u_i^2 Z_{i.} Z_{i.}'] = Q_{ZZ}\sigma_u^2$, we get

$$\begin{aligned}\Omega^{hom} =& Q_{ZZ}\sigma_u^2 - 2pQ_{Z_1Z_1}\sigma_{uv}^2\beta + 2pQ_{Z_1Z_1}\sigma_{uv}^2\beta + pQ_{Z_1Z_1}\sigma_v^2\beta^2 + p\frac{p}{1-p}Q_{Z_1Z_1}Q_{Z_0Z_0}^{-1}Q_{Z_1Z_1}\sigma_v^2\beta^2 \\ =& Q_{ZZ}\sigma_u^2 + pQ_{Z_1Z_1}\sigma_v^2\beta^2 + p\frac{p}{1-p}Q_{Z_1Z_1}Q_{Z_0Z_0}^{-1}Q_{Z_1Z_1}\sigma_v^2\beta^2 \\ =& Q_{ZZ}\sigma_u^2 - Q_{ZZ}\sigma_v^2\beta^2 + \frac{1}{1-p}Q_{ZZ}Q_{Z_0Z_0}^{-1}Q_{ZZ}\sigma_v^2\beta^2,\end{aligned} \tag{21}$$

where we use the fact that $Q_{ZZ} = pQ_{Z_1Z_1} + (1-p)Q_{Z_0Z_0}$ in the last equality. The asymptotic variance is

$$\begin{aligned}V_{\widehat{\beta}_{RI}}^{hom} =& \left(Q_{xZ}Q_{ZZ}^{-1}Q_{Zx}\right)^{-1}\sigma_u^2 - \left(Q_{xZ}Q_{ZZ}^{-1}Q_{Zx}\right)^{-1}\sigma_v^2\beta^2 \\ &+ \frac{1}{1-p}\left(Q_{xZ}Q_{ZZ}^{-1}Q_{Zx}\right)^{-1}\left(Q_{xZ}Q_{Z_0Z_0}^{-1}Q_{Zx}\right) \\ &\left(Q_{xZ}Q_{ZZ}^{-1}Q_{Zx}\right)^{-1}\sigma_v^2\beta^2.\end{aligned}$$

Under MCAR, we know $Q_{Z_0Z_0} = Q_{ZZ}$, which we can use to derive the asymptotic variance of $\widehat{\beta}_{RI}$ under homoskedasticity and MCAR.