Correcting the standard errors of two-stage residual inclusion estimators for Mendelian randomization studies

Tom M. Palmer[1] Michael V. Holmes[2,3] Brendan J. Keating[3,4]
Nuala A. Sheehan[5]

December 21, 2016

Corresponding author contact details:
Dr Tom Palmer, BSc, MSc, PhD
Telephone: +44 (0)1524 594019
email: t.palmer1@lancaster.ac.uk

Running head: Correcting standard errors for TSRI estimators

Abbreviations: BMI: body-mass index; BS: bootstrap; GMM: generalised method of moments; LSMM: logistic structural mean model; SBP: systolic blood pressure; SE: standard error; TSLS: two-stage least squares; TSPS: two-stage predictor substitution; TSRI: two-stage residual inclusion;

[1]Department of Mathematics and Statistics, Lancaster University, Lancaster, UK.
[2]Clinical Trial Service Unit & Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK.
[3]Department of Surgery, University of Pennsylvania, PA 19104, US.
[4]Children's Hospital of Philadelphia, Philadelphia, PA 19104, US.
[5]Departments of Health Sciences and Genetics, University of Leicester, Leicester, UK.

Abstract

Mendelian randomization studies use genotypes as instrumental variables to test for and estimate the causal effects of modifiable risk factors on outcomes. Two-stage residual inclusion (TSRI) estimators have been used when researchers are willing to make parametric assumptions. However, researchers are currently reporting uncorrected or heteroskedasticity robust standard errors (SEs) for these estimates.

We compare several different forms of the SE for linear and logistic TSRI estimates in simulations and in real data examples. Amongst others we consider SEs modified from the approach of Newey (1987), Terza (2016), and bootstrapping.

In our simulations Newey, Terza, bootstrap, and corrected two-stage least squares (in the linear case) standard errors gave the best results in terms of coverage and type I error. In the real data examples the Newey SEs were 0.5% and 2% larger than the unadjusted standard errors for the linear and logistic TSRI estimators respectively.

We show that TSRI estimators with modified SEs have correct type I error under the null. Researchers should report TSRI estimates with modified SEs instead of reporting unadjusted or heteroskedasticity robust SEs.

Keywords: Causal inference, instrumental variables, Mendelian randomization, two-stage predictor substitution estimators, two-stage residual inclusion estimators.

Introduction

Mendelian randomization studies aim to use genotypes as instrumental variables to test and estimate the causal effect of modifiable exposures on disease related outcomes.[1–4] A variety of instrumental variable estimators have been described and evaluated for use with data in a single study.[5–12] A class of semiparametric estimators known as structural mean models have been found to be most robust to distributional assumptions for binary outcomes but can have problems with identification.[7,13–16] Therefore, researchers may wish to fit models which make more distributional assumptions.

One frequently used instrumental variable estimator is two-stage least squares (TSLS). This is a series of two linear models and is most commonly applied when both the exposure and outcome variables are continuous. The first stage is a linear regression of the exposure on the instrumental variables. The second stage is a linear regression of the outcome on the predicted values of the exposure from the first stage. TSLS is consistent for the causal effect when all relationships are linear and there are no interactions between the instrument and unmeasured confounders and between the exposure and unmeasured confounders. Palmer et al. (2008) investigated two instrumental variable estimators of the causal odds ratio for a binary outcome the "standard" and "adjusted" logistic instrumental variable estimators.[17] The standard logistic instrumental variable estimator replaced the linear regression in the second stage of TSLS with a logistic regression. Such estimators have been referred to as "two-stage predictor substitution" (TSPS) estimators which are written as follows,[18]

$$\text{Stage 1:} \quad X = \alpha_0 + \alpha_1 Z + \varepsilon_1, \quad \varepsilon_1 \sim N(0, \sigma_1^2) \tag{1}$$

$$\text{Stage 2:} \quad h(E[Y]) = \beta_0 + \beta_1 \widehat{X} \tag{2}$$

where X represents the exposure variable, Y the outcome variable, Z the instrumental variable, $h()$ the link function for the appropriate generalised linear model,[19] and ε_1 the stage 1 residuals with variance σ_1^2.

The adjusted logistic instrumental variable estimator included the first stage residuals alongside the predicted values of the exposure in the second stage logistic regression.[17] In the econometrics literature it is more common to fit the second stage of such estimators using the original values of the exposure.[18,9] When the residuals are included as an additive covariate these estimators have been referred to as "two-stage residual inclusion" (TSRI) estimators.[18,20,21] If a function of the residuals is included in the second stage model these estimators have been referred to as control function estimators.[22] Therefore, the second stage of TSRI estimators considered in this paper can be written as follows,

$$\text{Stage 2:} \quad h(E[Y]) = \beta_0 + \beta_1 X + \beta_2 \widehat{\varepsilon}_1. \tag{3}$$

In this paper we use 'linear/logistic TSRI estimator' to refer to the estimator using linear/logistic regression at the second stage (with a linear first stage).

A recent review of Mendelian randomization studies showed that TSRI estimators are commonly used but are typically being reported with unadjusted or heteroskedasticity robust standard errors.[23–34]. One indication that this may not be appropriate is that when TSLS is estimated by fitting the two stages sequentially the standard errors of the second stage parameter estimates are not correct (Web Appendix 1, Web Figures 1–3).[35] Interestingly, for the linear TSRI estimator the standard error of the coefficient on the first stage residuals is correct.[36] For a binary outcome Newey (1987) developed a correction to the standard errors of the second stage intercept and causal effect of the Probit TSRI estimator.[37] More recently Terza (2016) has suggested an alternative correction.[38] The aim of this paper is to investigate these corrections adapted to the linear and logistic TSRI estimators.

This paper proceeds by describing the Probit TSRI estimator and Newey's correction for its standard errors. We then perform two simulation studies using binary and continuous outcomes to investigate the performance of the corrected standard errors. We then apply these corrections to a real data example investigating the causal effect of body-mass index (BMI) on systolic blood pressure (SBP) and on a binary diabetes status indicator.

Methods

Background to TSRI estimators

Two reviews of TSRI estimators and their application have been given.[18,22] The rationale for TSRI estimators is that the first-stage residuals capture some of the variability in the confounders. Therefore, the first stage residuals can be used to correct for confounding between the exposure and the outcome, known as endogeneity in econometrics.[39–43] It is well known that the linear TSRI estimator produces an estimate of the causal effect equivalent to that from TSLS.[36,44] Hausman (1978) showed that the test of the coefficient of the first stage residuals is a test for the presence of unmeasured confounding.[45–47] That it is necessary to correct the standard errors of the second stage estimate of the causal effect of TSRI estimators has been referred to as the problem of using "generated regressors" in the second stage model.[36,48,49]

For binary outcomes the use of Probit TSRI estimator has been discussed.[36,50–52] There are several estimation methods available including maximum likelihood and sequential two-stage methods. For two-stage estimation a correction to the second stage standard errors was proposed by Newey (1987) which is implemented in the `ivprobit` and `ivtobit` Stata (College Station, Texas) commands.[37,53]

It is also important to distinguish between different causal effects. We refer to a conditional causal effect as the value of the causal effect conditioning on the unmeasured confounding and to a marginal effect as the causal effect averaged over some proportion of the unmeasured confounding. The maximum likelihood Probit TSRI estimator estimates the conditional effect, whereas the two-stage Probit and logistic TSRI estimators estimate marginal effects.[53,17,12,18,21]

Probit TSRI estimator and Newey standard errors

Two-stage estimation of the Probit TSRI estimator follows Equations 1 and 3, where the inverse Normal cumulative distribution function is used as the link function. If there are measured confounders, as with TSLS, these can be included as covariates in both stages of estimation. Letting $\widehat{\beta} = \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_0 \end{bmatrix}$ denote the vector of estimates of the causal effect and intercept yielded by the Probit TSRI estimator, and defining the matrix $\widehat{D}$ as $\begin{bmatrix} \widehat{\alpha}_1 & 0 \\ \widehat{\alpha}_0 & 1 \end{bmatrix}$,[37,53]

$$\widehat{\beta} = (\widehat{D}'\widehat{\Omega}^{-1}\widehat{D})^{-1}\widehat{D}'\widehat{\Omega}^{-1}\widehat{\gamma}. \tag{4}$$

The variance of the Probit TSRI estimator is as follows, where $\widehat{\gamma}$, $\widehat{\Omega}$ and its components are defined below,[37]

$$\text{var}(\widehat{\beta}) = (\widehat{D}'\widehat{\Omega}^{-1}\widehat{D})^{-1} \tag{5}$$

$$\text{where } \widehat{\Omega} = J_1^{-1} + \Sigma_2. \tag{6}$$

To obtain $\widehat{D}$, $\widehat{\gamma}$, and $\widehat{\Omega}$ we use the following algorithm as described by Newey (1987).[37]

1. Perform the first stage linear regression of X on Z to compile $\widehat{D}$ and $\widehat{\varepsilon}_1$.
2. Perform a Probit regression of Y on Z and $\widehat{\varepsilon}_1$, from which;
   (i) $\widehat{\gamma}$ is the coefficients of Z and the estimated intercept.
   (ii) J_1^{-1} is the variance-covariance matrix of these coefficients.
   (iii) denote $\widehat{\lambda}$ as the coefficient on $\widehat{\varepsilon}_1$.
3. Fit the second stage of the Probit TSRI estimator by a Probit regression of Y on X and $\widehat{\varepsilon}_1$.
   (i) The coefficient on X is $\widehat{\beta}_1$, the estimate of the causal effect of interest.
4. Generate a new variable equal to $X(\widehat{\lambda} - \widehat{\beta}_1)$.
   (i) Perform a linear regression of this new variable on Z (also including a constant).

(ii) The covariance matrix from this model is the estimate of the second term in the expression for $\widehat{\Omega}$, i.e. Σ_2.

(iii) Add this covariance matrix to J_1^{-1} giving $\widehat{\Omega}$.

5. Calculate $\widehat{\beta}$ and var($\widehat{\beta}$). The standard errors of $\widehat{\beta}$ are simply the square root of the diagonal of var($\widehat{\beta}$).

The rationale for this approach is that we obtain a standard error for our TSRI estimate which incorporates both the variability explained in the estimate by the instrumental variable Z and the predicted first stage residuals $\widehat{\varepsilon}_1$.

To apply these standard errors to other TSRI estimators we propose to replace the Probit regressions in steps 2 and 3 with the second stage models used by the specific TSRI estimator. Example Stata and R code is given in Web Appendix 2.[54,55]

Terza (2016) details an alternative algorithm for obtaining the standard error of TSRI estimators and provides example Stata code.[38] We provide equivalent R code in Web Appendix 2. Terza (2016) uses heteroskedasticity robust standard errors in both stages of the algorithm, which we refer to as Terza SE 1. We additionally investigate using non-robust standard errors , which we refer to as Terza SE 2.[38] By following the code in Web Appendix 2 we can see that the Terza corrected variance covariance matrix is the unadjusted TSRI covariance matrix plus some function of the first stage covariance matrix.

We also investigate two types of non-parametric bootstrap standard errors. The first only bootstraps the second stage, which we refer to as BS 1, whereas the second bootstraps both the first and second stages, which we refer to as BS 2. BS 2 is implented in the `ivprobit` and `ivtobit` Stata commands. And for our binary outcome models we additionally investigate the Probit TSRI estimator, whose estimates we convert to the odds ratio scale by dividing the estimate on the linear predictor scale by 0.6071 and taking the exponential.[10] These Probit estimates use Newey standard errors.

For all estimators we calculate asymptotic normal 95% confidence interval (CI) limits as:

estimate $\pm$ 1.96 $\times$ standard error.

Simulations

Logistic model simulations

Data were simulated using the basic model proposed in Palmer et al. (2008) but modifying the parameter values.[17] Specifically the data generation model was as follows where index i represents an observation and $\text{logit}(p_i) = \log(p_i/(1-p_i))$.

$$
\begin{aligned}
g_i &\sim \text{Binomial}(2,\ 0.3) \\
u_i &\sim N(0,1) \text{ – representing the unmeasured confounding,} \\
x_i &\sim \alpha_0 + \alpha_1 g_i + \alpha_2 u_i + \varepsilon_{1i},\ \varepsilon_{1i} \sim N(0,1) \\
\text{logit}(p_i) &= \beta_0 + \beta_1 x_i + \beta_2 u_i \\
y_i &\sim \text{Binomial}(1,\ p_i) \\
\alpha_0 = 0,\ \alpha_1 = 1,\ &\alpha_2 = \{0,2,4,6,8\}, \beta_0 = \log(0.05/0.95), \beta_1 = 1, \beta_2 = [0,3]
\end{aligned} \tag{7}
$$

Data were simulated for sample sizes of 1 000 and 5 000 and each scenario of values of α_2 and β_2, representing the effects of the unobserved confounding, was repeated 500 times. A number of different estimators were fitted to the data; the direct logistic regression of Y on X, the logistic TSPS, the logistic TSRI with unadjusted, robust, Newey, Terza 1 and 2, TSPS, and BS 1 and 2 standard errors. We also investigated the logistic structural mean model (LSMM) estimated via the generalized method of moments (GMM)[56] and the rescaled Probit estimator with Newey standard errors.

In these simulations, with sample size 1 000 for the first stage model, the average F statistics were 422, 85, 25, 12, and 7 and the average R^2 statistics were 0.30, 0.08, 0.02, 0.01, and 0.007 when α_2 was equal to 0, 2, 4, 6, and 8 respectively. With a sample size of 5 000 the average F statistics increased to 2104, 421, 125, 57, and 33 and the average R^2

statistics were the approximately the same.

Type I error was assessed by generating the data with β_1 set to 0, i.e. which corresponds to the null hypothesis of no causal effect, and counting the percentage of simulations for which the particular estimator gave a p-value less than 0.05. Coverage was defined as a 95% confidence interval including the value of either the conditional or marginal value of β_1. Marginal values of β_1 for the estimators were obtained using the adjustments detailed in the Appendix of Palmer et al.[17] (and Web Appendix 3, Web Figure 4). Simulations were performed in Stata (version 14.1).[54]

Linear model simulations

For a continuous outcome the simulations were modified as follows.

$$y_i \sim \beta_0 + \beta_1 x_i + \beta_2 u_i + \varepsilon_{2i},\ \varepsilon_{2i} \sim N(0,1)$$
$$\alpha_0 = 0,\ \alpha_1 = 1,\ \alpha_2 = \{0,2,4,6,8\},\ \beta_0 = 0,\ \beta_1 = 1,\ \beta_2 = [0,3] \quad (8)$$

For a linear second stage model the conditional and marginal parameter values are the same. Type I error was assessed by setting β_1 to 0. A number of linear estimators were fitted to the data; the direct linear regression of Y on X, TSLS with adjusted and unadjusted (i.e. TSPS) standard errors, the linear TSRI estimator with unadjusted, robust, Newey, Terza 1 and 2, TSLS, and BS 1 and 2 standard errors.

Results

Logistic model simulations

Figure 1 and Web Figure 5 show that with respect to the conditional parameter ($\beta_1 = 1$) all estimators have low coverage at some point in the simulations. This is mainly because of the bias in the parameter estimates. The conditional coverage of several estimators

that we didn't expect to perform well because their standard errors do not account for the uncertainty in both stages of estimation (TSRI with unadjusted, robust, and BS 1 standard errors) was around the 95% level for some larger values of α_2. This occurred because their standard errors increased in proportion with their bias. The conditional coverage of TSRI using the Newey, Terza 1 and 2, and BS 2 standard errors was the closest to 95% for the greatest proportion of simulated scenarios.

Figure 2 and Web Figure 6 show the coverage with respect to the marginal parameter values estimated by the TSRI estimator (the true marginal values are given in Figure 2 of Palmer et al. [2008] and Web Appendix 3).[17] The logistic TSPS estimator with unadjusted and robust SEs had coverage values well below the target value of 95%. The coverage of the logistic TSRI estimator with unadjusted, robust, and BS 1 standard errors was lower than the expected 95%. The coverage of the logistic TSRI estimator with TSPS standard errors was also too low and decreased as the confounding increased. However, the coverage of the logistic TSRI estimator using Newey, Terza 1 and 2,and BS 2 standard errors, and the coverage of LSMM were approximately correct with values around 95%.

Figure 3 and Web Figure 7 show that the type I error of the logistic TSRI estimator with unadjusted, robust, and BS 1 standard errors was too high with values greater than the nominal level of 5%. Type I error was also too high for the logistic TSRI estimator with TSPS standard errors when there was confounding. . For the LSMM estimates the type I error was approximately correct with values around 5%. The logistic TSPS estimator also had approximately correct type I error with unadjusted and robust standard errors with values around 5%. This is because under the null there isn't substantial bias in the logistic TSPS estimates. The logistic TSRI estimator using Newey, Terza 1 and 2, and BS 2 standard errors had approximately correct type I error with values around 5%.

Similar trends can be seen in the results for the simulations using a sample size of 5 000 in Web Figures 8, 9, and 10.

Web Figures 2–3 show that the correction to the logistic TSRI standard error has largest

effect when the absolute value of the correlation between the confounders of the exposure and outcome is greater than about 0.5, or more generally when the effect of the confounder is stronger. The effect of the correction is also more pronounced when the outcome has a higher prevalence (up to 50% beyond which the effect decreases).

Linear model simulations

The results in Figure 4 and Web Figure 11 show that the direct regression of Y and X has poor coverage when there is confounding. This is because of the bias in the point estimate. The TSRI estimator with unadjusted standard errors had poor coverage because the standard error does not account the uncertainty from the first stage estimation. TSRI using BS 1 and robust standard errors also showed poor coverage for the same reason. Usually we want the standard errors for the TSRI estimate to be larger than the unadjusted standard error, but robust standard errors are often smaller. All other estimators demonstrated coverage values around 95%. The coverage values for TSRI using the two Terza standard errors were slightly above 95%. The coverage of TSRI using Newey and TSLS standard errors fell below 95% as the amount of confounding increased. TSRI using BS 2 standard errors had the coverage values consistently closest to 95% over the range of the simulated scenarios.

The type I error results in Figure 5 and Web Figure 12 essentially show the same pattern as for the coverage results. The type I error of TSRI using unadjusted and BS 1 standard errors is inflated, whereas it is approximately correct for the other TSRI standard errors. Again the type I error of the TSLS and Newey standard errors is inflated as the confounding increases. The type I error of the two Newey standard errors is slightly below 5% and the results for BS 2 are the closest to 5% over the range of the simulations.

Similar trends can also be seen in Web Figures 13, 14.

Web Figure 1 shows that the correction to the TSRI standard errors has the largest effect when the absolute value of the correlation between the confounders of the exposure and outcome residuals is greater than about 0.5, or more generally when the confounding is

stronger.

Example: causal effect of BMI on SBP and diabetes

Data were taken on 17 057 participants from 6 prospective cohorts of European ancestry that had been genotyped with the Human CVD BeadArray (Illumina), also termed the "IBC" or "CardioChip" array.[57] The 6 cohorts are Atherosclerosis Risk in Communities (ARIC),[58] the Cardiovascular Health Study (CHS),[59] Coronary Artery Risk Development in Young Adults (CARDIA),[60] the Framingham Heart Study (FHS),[61] Multinational Etoricoxib and Diclofenac Arthritis Long-term (MEDAL),[62] and the Multi-Ethnic Study of Atherosclerosis (MESA).[63]

Individuals had complete data on variables for body mass index (BMI), systolic blood pressure (SBP), and diabetes. An externally weighted allele score was constructed out of the genetic variants for BMI. Details of the genetic variants and the construction of the allele scores have been previously reported.[64] In the first example we estimate the causal effect of BMI on SBP using linear IV estimators. In the second example we estimate the causal odds ratio for diabetes for a unit increase in BMI using binary outcome IV estimators . Analysis was performed using Stata (version 13.1).[54]

The prevalence of the diabetes outcome was 13.7%. Table 1 shows the estimated causal odds ratios for diabetes for a one unit increase in BMI. The direct estimate of the odds ratio was 1.14 (95% CI 1.13, 1.15). In the first stage of TSPS and TSRI estimation the instrument gave a first stage F-statistic of 119, greater than the usual cut-off for a weak instrument of 10, but a low R^2 of 0.7%. The logistic TSPS estimate was larger at 1.32 (95% CI 1.19, 1.48) and also excluded a null effect. The logistic TSRI gave the same point estimate of the causal odds ratio. For the TSRI estimator the unadjusted standard error was 0.058 whereas the Newey standard error was 2% larger at 0.059. The two Terza standard errors were 0.057 and 0.059 respectively. For logistic TSRI the BS 1 standard error was the same as the robust standard error, whereas the BS 2 standard error at 0.061

was larger than the Newey and Terza 2 standard errors. For the logistic TSRI with the Newey standard error the z-statistic was 4.71 whereas the Probit TSRI gave a slightly larger z-statistic of 4.74. The logistic SMM gave a larger point estimate of the causal odds ratio of 1.39 (95% CI 1.19, 1.59) and also a larger standard error as shown by the smaller z-statistic and wider CI. We conclude that the observational estimate of the causal odds ratio has been attenuated by unmeasured confounding and that these data support a causal effect of BMI on the risk of diabetes.

Table 2 shows the estimates of the effect of a 1 unit increase in BMI on SBP. The direct estimate of this association was 0.76mmHg (95% CI 0.70, 0.82). Using the same first stage as the logistic TSPS and TSRI estimators, TSLS gave an estimate for a 1 unit increase of BMI of 0.36mmHg (95% CI -0.37, 1.10) with a standard error of 0.374. The linear TSRI gave the same point estimate with a smaller unadjusted standard error of 0.372. The Newey standard error of 0.374 was equal to the TSLS standard error, and the two Terza standard errors were slightly smaller at 0.370 and 0.372. In this example the BS 2 standard error was the largest at 0.384. The Newey correction increased the standard error by 0.5%. We conclude that the observational association is likely to be partly explained by unmeasured confounding and that the data do not support a causal effect of BMI on SBP.

In this example, the standard errors which don't take into account the uncertainty from both stages of estimation (unadjusted, robust, and BS 1) are only slightly smaller than those that do (TSLS, Newey, Terza 1 and 2, BS 2, LSMM, and Probit) because of the combination of low first stage R^2 and large sample size.

Discussion

In this paper, we have adapted corrections to the standard errors of TSRI estimators, developed by Newey (1987) and Terza (2016), to the linear and logistic TSRI estimators.[37,38] The results of our simulations show that Newey, Terza, BS 2, and corrected

TSLS (for the linear case) standard errors have the best properties in terms of coverage and type I error.

The methods were illustrated in real data examples investigating the effect of BMI on SBP and diabetes risk respectively. In the examples the Newey standard errors were 0.5% and 2% larger than the unadjusted standard errors for the linear and logistic TSRI estimators respectively. In the supplementary material we show that the corrections to the TSRI standard errors have most effect when the unmeasured confounding is greater and when the outcome prevalence is higher (up to 50%, beyond which the effect decreases). In the binary outcome example the Probit TSRI estimator gave a slightly larger z-statistic than the logistic TSRI estimator. The standard error of the logistic TSRI estimator could be scaled to give the same z-statistic. We do not prefer this approach because using the scaled standard error in Equation 4 would not give the same value as sequential two-step estimation.

Further work could investigate the application of Newey and Terza standard errors to TSRI estimators using other generalised linear models at the second stage. For example, Terza et al. (2008) used a parametric Weibull model and an ordered logistic regression model in the second stage.[18] And Tchetgen Tchetgen et al. (2015) discussed TSRI estimators for survival models using an Aalen additive hazard model at the second stage.[65]. Our work has applicability beyond Mendelian randomization studies because TSRI estimators have been used in other areas, for example, using randomized treatment status in a clinical trial as an instrumental variable to correct for non-compliance and in health economics.[66,18,22]

Newey's correction to the standard errors of the two-step Probit TSRI estimator relates to Murphy-Topel standard errors in econometrics which can be used for TSPS estimates.[67] Murphy-Topel standard errors have been implemented in Stata .[68–71] It has been argued that researchers may want to fit the logistic TSPS estimator because it is consistent for the effect averaged over the population,[72,73] whereas it is less clear what effect is identified by the TSRI estimator.[74] Also TSPS estimators have correct type I error under the

null.[17,72,75] However, since we have shown that using Newey, Terza, and BS 2 standard errors for TSRI estimators also gives correct type I error under the null we argue that TSRI estimators are attractive to researchers using non-collapsible models at the second stage. There is also scope to use TSRI estimates, with corrected standard errors, as part of the algorithms in the recently proposed MR-Egger and median estimators, which are robust to different proportions of invalid instruments.[76,77]

In conclusion, we recommend that researchers fitting TSRI estimators should not report unadjusted or heteroskedasticity robust standard errors but should report standard errors using the Newey or Terza corrections or from bootstrapping including both stages of estimation.

Acknowledgments

Author affiliations: Children's Hospital of Philadelphia, Philadelphia, PA 19104, US (Brendan J. Keating); Clinical Trial Service Unit & Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK (Michael V. Holmes); Departments of Health Sciences and Genetics, University of Leicester, Leicester, UK (Nuala A. Sheehan); Department of Mathematics and Statistics, Lancaster University, Lancaster, UK (Tom M. Palmer); Department of Surgery, University of Pennsylvania, PA 19104, US (Michael V. Holmes, Brendan J. Keating).

The authors thank Prof. Joe Terza (Indiana University Purdue University, Indianapolis) for pre-publication access to his Stata Journal article.

Conflicts of interest: none declared.

References

[1] G. Davey Smith and S. Ebrahim. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease. *International Journal of Epidemiology*, 32(1):1–22, 2003.

[2] V. Didelez and N. Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330, 2007.

[3] S. Burgess, N. J. Timpson, S. Ebrahim, and G. Davey Smith. Mendelian randomization: where are we now and where are we going? *International Journal of Epidemiology*, 44(2):379–388, 2015.

[4] S. Burgess, A. S. Butterworth, and J. R. Thompson. Beyond Mendelian randomization: How to interpret evidence of shared genetic predictors. *Journal of Clinical Epidemiology*, 69:208–216, 2016.

[5] S. Vansteelandt and E. Goetghebeur. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society: Series B*, 65(4):817–835, 2003.

[6] K. M. Johnston, P. Gustafson, A. R. Levy, and P. Grootendorst. Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Statistics in Medicine*, 27(9):1539–1556, 2008.

[7] V. Didelez, S. Meng, and N. A. Sheehan. Assumptions of IV methods for observational epidemiology. *Statistical Science*, 25(1):22–40, 2010.

[8] J. Bowden and S. Vansteelandt. Mendelian randomisation analysis of case-control data using structural mean models. *Statistics in Medicine*, 30(6):678–694, 2011.

[9] T. M. Palmer, J. A. C. Sterne, R. M. Harbord, D. A. Lawlor, N. A. Sheehan, S. Meng, R. Granell, G. D. Smith, and V. Didelez. Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses. *American Journal of Epidemiology*, 173(12):1392–1403, 2011.

[10] S. Vansteelandt, J. Bowden, M. Babanezhad, and E. Goetghebeur. On instrumental variables estimation of causal odds ratios. *Statistical Science*, 26(3):403–422, 2011.

[11] S. Burgess and S. G. Thompson. Improving bias and coverage in instrumental variable analysis with weak instruments for continuous and binary outcomes. *Statistics in Medicine*, 31(15):1582–1600, 2012.

[12] R. M. Harbord, V. Didelez, T. M. Palmer, S. Meng, J. A. Sterne, and N. A. Sheehan. Severity of bias of a simple estimator of the causal odds ratio in Mendelian randomization studies. *Statistics in Medicine*, 32(7):1246–1258, 2013.

[13] J. M. Robins. *Health Services Research Methodology: A focus on AIDS*, chapter The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. US Public Health Service, Washington DC, US, 1989.

[14] J. M. Robins. Correcting for non-compliance in randomized trials using struc-

tural nested mean models. *Communications in Statistics: Theory and Methods*, 23(8):2379–2412, 1994.

[15] P. S. Clarke and F. Windmeijer. Identification of causal effects on binary outcomes using structural mean models. *Biostatistics*, 11(4):756–770, 2010.

[16] S. Burgess, R. Granell, T. M. Palmer, J. A. C. Sterne, and V. Didelez. Lack of identification in semiparametric instrumental variable models with binary outcomes. *American Journal of Epidemiology*, 180(1):111–119, 2014.

[17] T. M. Palmer, J. R. Thompson, M. D. Tobin, N. A. Sheehan, and P. R. Burton. Adjusting for bias and unmeasured confounding in the analysis of Mendelian randomization studies with binary responses. *International Journal of Epidemiology*, 37(5):1161–1168, 2008.

[18] J. V. Terza, A. Basu, and P. J. Rathouz. Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics*, 27(3):531–543, 2008.

[19] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall, second edition, 1989.

[20] A. J. O'Malley, R. G. Frank, and S.-L. T. Normand. Estimating cost-offsets of new medications: Use of new antipsychotics and mental health costs for schizophrenia. *Statistics in Medicine*, 30(16):1971–1988, 2011.

[21] B. Cai, D. S. Small, and T. R. T. Have. Two-stage instrumental variable methods for estimating the causal odds ratio: Analysis of bias. *Statistics in Medicine*, 30(15):1809–1824, 2011.

[22] M. M. Garrido, P. Deb, J. F. Burgess, and J. D. Penrod. Choosing models for health care cost analyses: Issues of nonlinearity and endogeneity. *Health Services Research*, 47(6):2377–2397, 2012.

[23] A. G. Boef, O. M. Dekkers, and S. le Cessie. Mendelian randomization studies: A review of the approaches used and the quality of reporting. *International Journal of Epidemiology*, 44(2):496–511, 2015.

[24] M. Benn, A. Tybjærg-Hansen, S. Stender, R. Frikke-Schmidt, and B. G. Nordestgaard. Low-density lipoprotein cholesterol and the risk of cancer: A Mendelian randomization study. *Journal of the National Cancer Institute*, 103(6):508–519, 2011.

[25] S. Collin, C. Metcalfe, T. Palmer, H. Refsum, S. Lewis, G. Davey, D. Donovan1, F. Hamdy, A. Smith, and R. Martin. The causal roles of vitamin B12 and transcobalamin in prostate cancer: can Mendelian randomization analysis provide definitive answers? *International Journal of Molecular Epidemiology and Genetics*, 2(4):316–327, 2011.

[26] N. M. G. De Silva, R. M. Freathy, T. M. Palmer, L. A. Donnelly, J. Luan, T. Gaunt, C. Langenberg, M. N. Weedon, B. Shields, B. A. Knight, K. J. Ward, M. S. Sandhu, R. M. Harbord, M. I. McCarthy, G. D. Smith, S. Ebrahim, A. T. Hattersley, N. Wareham, D. A. Lawlor, A. D. Morris, C. N. Palmer, and T. M. Frayling. Mendelian randomization studies do not support a role for raised circulating triglyceride levels

influencing type 2 diabetes, glucose levels, or insulin resistance. *Diabetes*, 60(3):1008–1018, 2011.

[27] D. A. Lawlor, R. M. Harbord, A. Tybjærg-Hansen, T. M. Palmer, J. Zacho, M. Benn, N. J. Timpson, G. Davey Smith, and B. G. Nordestgaard. Using genetic loci to understand the relationship between adiposity and psychological distress: A Mendelian randomization study in the Copenhagen General Population Study of 53221 adults. *Journal of Internal Medicine*, 269(5):525–537, 2011.

[28] M. Islam, T. Jafar, A. Wood, N. De Silva, M. Caulfield, N. Chaturvedi, and T. Frayling. Multiple genetic variants explain measurable variance in type 2 diabetes-related traits in pakistanis. *Diabetologia*, 55(8):2193–2204, 2012.

[29] E. Theodoratou, T. Palmer, L. Zgaga, S. M. Farrington, P. McKeigue, F. V. N. Din, A. Tenesa, G. Davey Smith, M. G. Dunlop, and H. Campbell. Instrumental variable estimation of the causal effect of plasma 25-hydroxy-vitamin D on colorectal cancer risk: A Mendelian randomization analysis. *PLoS ONE*, 7(6):e37662, 2012.

[30] D. A. Lawlor, B. G. Nordestgaard, M. Benn, L. Zuccolo, A. Tybjaerg-Hansen, and G. Davey Smith. Exploring causal associations between alcohol and coronary heart disease risk factors: Findings from a Mendelian randomization study in the Copenhagen General Population Study. *European Heart Journal*, 34(32):2519–2528, 2013.

[31] R. Haring, A. Teumer, U. Vlker, M. Drr, M. Nauck, R. Biffar, H. Vlzke, S. E. Baumeister, and H. Wallaschofski. Mendelian randomization suggests non-causal associations of testosterone with cardiometabolic risk factors and mortality. *Andrology*, 1(1):17–23, 2013.

[32] A. P. Thrift, N. J. Shaheen, M. D. Gammon, L. Bernstein, B. J. Reid, L. Onstad, H. A. Risch, G. Liu, N. C. Bird, A. H. Wu, D. A. Corley, Y. Romero, S. J. Chanock, W.-H. Chow, A. G. Casson, D. M. Levine, R. Zhang, W. E. Ek, S. MacGregor, W. Ye, L. J. Hardie, T. L. Vaughan, and D. C. Whiteman. Obesity and risk of esophageal adenocarcinoma and Barretts esophagus: A Mendelian randomization study. *Journal of the National Cancer Institute*, 106(11), 2014.

[33] M. V. Holmes, F. W. Asselbergs, T. M. Palmer, F. Drenos, M. B. Lanktree, C. P. Nelson, C. E. Dale, S. Padmanabhan, C. Finan, D. I. Swerdlow, V. Tragante, E. P. van Iperen, S. Sivapalaratnam, S. Shah, C. C. Elbers, T. Shah, J. Engmann, C. Giambartolomei, J. White, D. Zabaneh, R. Sofat, S. McLachlan, P. A. Doevendans, A. J. Balmforth, A. S. Hall, K. E. North, B. Almoguera, R. C. Hoogeveen, M. Cushman, M. Fornage, S. R. Patel, S. Redline, D. S. Siscovick, M. Y. Tsai, K. J. Karczewski, M. H. Hofker, W. M. Verschuren, M. L. Bots, Y. T. van der Schouw, O. Melander, A. F. Dominiczak, R. Morris, Y. Ben-Shlomo, J. Price, M. Kumari, J. Baumert, A. Peters, B. Thorand, W. Koenig, T. R. Gaunt, S. E. Humphries, R. Clarke, H. Watkins, M. Farrall, J. G. Wilson, S. S. Rich, P. I. de Bakker, L. A. Lange, G. Davey Smith, A. P. Reiner, P. J. Talmud, M. Kivimäki, D. A. Lawlor, F. Dudbridge, N. J. Samani, B. J. Keating, A. D. Hingorani, and J. P. Casas. Mendelian randomization of blood lipids for coronary heart disease. *European Heart Journal*, 36(9):539–550, 2015.

[34] Z. Ye, P. C. Haycock, D. Gurdasani, C. Pomilla, S. M. Boekholdt, S. Tsimikas, K.-T. Khaw, N. J. Wareham, M. S. Sandhu, and N. G. Forouhi. The association between

circulating lipoprotein(a) and type 2 diabetes: Is it causal? *Diabetes*, 63(1):332–342, 2014.

[35] R. Davidson and J. G. Mackinnon. *Estimation and inference in econometrics*. Oxford University Press, New York, US, 1993.

[36] J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2002.

[37] W. K. Newey. Efficient estimation of limited dependent variable models with endogenous explanatory variables. *Journal of Econometrics*, 36(3):231–250, 1987.

[38] J. V. Terza. Simpler standard errors for two-stage optimization estimators. *Stata Journal*, 16(2):368–385, 2016.

[39] J. Garen. The returns to schooling: A selectivity bias approach with a continuous choice variable. *Econometrica*, 52(5):1199–1218, 1984.

[40] J. Heckman and R. Robb. Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*, 30(1-2):239–67, 1985.

[41] J. M. Wooldridge. On two stage least squares estimation of the average treatment effect in a random coefficient model. *Economics Letters*, 56(2):129–133, 1997.

[42] W. K. Newey, J. L. Powell, and F. Vella. Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, 67(3):565–603, 1999.

[43] R. W. Blundell and J. L. Powell. *Advances in Economics and Econometrics: Theory and Applications. 8th World Congress of the Econometric Society*, chapter Endogeneity in nonparametric and semiparametric regression models, pages 312–357. Cambridge University Press, Cambridge, UK, 2003.

[44] P. J. Dhrymes. *Econometrics: Statistical Foundations and Applications*. Harper and Row, New York, NY, 1970.

[45] J. Durbin. Errors in variables. *Review of the International Statistical Institute*, 22(1):23–32, 1954.

[46] D.-M. Wu. Alternative tests of independence between stochastic regressors and disturbances: Finite sample results. *Econometrica*, 42(3):529–546, 1974.

[47] J. A. Hausman. Specification tests in econometrics. *Econometrica*, 46(6):1251–1271, 1978.

[48] A. Pagan. Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, 25(1):221–247, 1984.

[49] A. C. Cameron and P. K. Trivedi. *Microeconometrics: Methods and Applications*. Cambridge University Press, New York, 2005.

[50] R. J. Smith and R. W. Blundell. An exogeneity test for a simultaneous equation tobit model with an application to labor supply. *Econometrica*, 54(3):679–685, 1986.

[51] D. Rivers and Q. H. Vuong. Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics*, 39(3):347–366, 1988.

[52] R. W. Blundell and R. J. Smith. Estimation in a class of simultaneous equation limited dependent variable models. *Review of Economics and Statistics*, 56(1):37–57, 1989.

[53] Stata Corp LP,. *StataBase Reference Manual Release 13*, chapter ivprobit – Probit model with continuous endogenous regressors, pages 910–922. Stata Press, College Station, Texas, 2013.

[54] Stata Corp. *Stata Statistical Software*. College Station, Texas, 2015. Version 14.1.

[55] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. Version 3.2.1.

[56] P. S. Clarke, T. M. Palmer, and F. Windmeijer. Estimating structural mean models with multiple instrumental variables using the generalised method of moments. *Statistical Science*, 30(1):96–117, 2015.

[57] B. J. Keating, S. Tischfield, S. S. Murray, T. Bhangale, T. S. Price, J. T. Glessner, L. Galver, J. C. Barrett, S. F. A. Grant, D. N. Farlow, H. R. Chandrupatla, M. Hansen, S. Ajmal, G. J. Papanicolaou, Y. Guo, M. Li, S. DerOhannessian, P. I. W. de Bakker, S. D. Bailey, A. Montpetit, A. C. Edmondson, K. Taylor, X. Gai, S. S. Wang, M. Fornage, T. Shaikh, L. Groop, M. Boehnke, A. S. Hall, A. T. Hattersley, E. Frackelton, N. Patterson, C. W. K. Chiang, C. E. Kim, R. R. Fabsitz, W. Ouwehand, A. L. Price, P. Munroe, M. Caulfield, T. Drake, E. Boerwinkle, D. Reich, A. S. Whitehead, T. P. Cappola, N. J. Samani, A. J. Lusis, E. Schadt, J. G. Wilson, W. Koenig, M. I. McCarthy, S. Kathiresan, S. B. Gabriel, H. Hakonarson, S. S. Anand, M. Reilly, J. C. Engert, D. A. Nickerson, D. J. Rader, J. N. Hirschhorn, and G. A. FitzGerald. Concept, design and implementation of a cardiovascular gene-centric 50 K SNP array for large-scale genomic association studies. *PLoS ONE*, 3(10):e3583, 2008.

[58] The ARIC Investigators,. The Atherosclerosis risk in communities (ARIC) study: Design and objectives. *American Journal of Epidemiology*, 129(4):687–702, 1989.

[59] L. P. Fried, N. O. Borhani, P. Enright, C. D. Furberg, J. M. Gardin, R. A. Kronmal, L. H. Kuller, T. A. Manolio, M. B. Mittelmark, A. Newman, D. H. O'Leary, B. Psaty, P. Rautaharju, R. P. Tracy, and P. G. Weiler. The cardiovascular health study: Design and rationale. *Annals of Epidemiology*, 1(3):263–276, 1991.

[60] G. D. Friedman, G. R. Cutter, R. P. Donahue, G. H. Hughes, S. B. Hulley, D. R. J. Jr., K. Liu, and P. J. Savage. CARDIA: Study design, recruitment, and some characteristics of the examined subjects. *Journal of Clinical Epidemiology*, 41(11):1105 – 1116, 1988.

[61] M. Feinleib, W. B. Kannel, R. J. Garrison, P. M. McNamara, and W. P. Castelli. The Framingham Offspring Study: Design and preliminary data. *Preventive Medicine*, 4(4):518–525, 1975.

[62] C. P. Cannon, S. P. Curtis, G. A. FitzGerald, H. Krum, A. Kaur, J. A. Bolognese, A. S. Reicin, C. Bombardier, M. E. Weinblatt, D. van der Heijde, E. Erdmann, and L. Laine. Cardiovascular outcomes with etoricoxib and diclofenac in patients with osteoarthritis and rheumatoid arthritis in the Multinational Etoricoxib and

Diclofenac Arthritis Long-term (MEDAL) programme: A randomised comparison. *The Lancet*, 368(9549):1771–1781, 2006.

[63] D. E. Bild, D. A. Bluemke, G. L. Burke, R. Detrano, A. V. Diez Roux, A. R. Folsom, P. Greenland, D. R. JacobsJr., R. Kronmal, K. Liu, J. C. Nelson, D. OLeary, M. F. Saad, S. Shea, M. Szklo, and R. P. Tracy. Multi-ethnic study of atherosclerosis: Objectives and design. *American Journal of Epidemiology*, 156(9):871–881, 2002.

[64] M. Holmes, L. Lange, T. Palmer, M. Lanktree, K. North, B. Almoguera, S. Buxbaum, H. Chandrupatla, C. Elbers, Y. Guo, R. Hoogeveen, J. Li, Y. Li, D. Swerdlow, M. Cushman, T. Price, S. Curtis, M. Fornage, H. Hakonarson, S. Patel, S. Redline, D. Siscovick, M. Tsai, J. Wilson, Y. vanderSchouw, G. FitzGerald, A. D. Hingorani, J. Casas, P. deBakker, S. Rich, E. Schadt, F. Asselbergs, A. Reiner, and B. Keating. Causal effects of body mass index on cardiometabolic traits and events: A Mendelian randomization analysis. *The American Journal of Human Genetics*, 94(2):198–208, 2014.

[65] E. J. Tchetgen Tchetgen, S. Walter, S. Vansteelandt, T. Martinussen, and M. Glymour. Instrumental variable estimation in a survival context. *Epidemiology*, 26(3):402–410, 2015.

[66] N. Nagelkerke, V. Fidler, R. Bernsen, and M. Borgdorff. Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine*, 19(14):1849–64, 2000. Erratum Stat Med 2001; 20: 982.

[67] K. M. Murphy and R. H. Topel. Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics*, 3(4):370–379, 1985. Reprinted 2002, 20(1):88–97.

[68] J. W. Hardin. The robust variance estimator for two-stage models. *Stata Journal*, 2(3):253–265, 2002.

[69] A. R. Hole. Calculating Murphy-Topel variance estimates in Stata: A simplified procedure. *The Stata Journal*, 6(4):521–529, 2006.

[70] J. W. Hardin and R. J. Carroll. Variance estimation for the instrumental variables approach to measurement error in generalized linear models. *Stata Journal*, 3(4):342–350, 2003.

[71] J. W. Hardin, H. Schmiediche, and R. J. Carroll. Instrumental variables, bootstrapping, and generalized linear models. *Stata Journal*, 3(4):351–360, 2003.

[72] S. Burgess and CRP CHD Genetics Collaboration. Identifying the odds ratio estimated by a two-stage instrumental variable analysis with a logistic regression model. *Statistics in Medicine*, 32(27):4726–4747, 2013.

[73] S. Burgess and S. G. Thompson. *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation*. Chapman and Hall/CRC, London, UK, 2015.

[74] S. Burgess, D. S. Small, and S. G. Thompson. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research*. [Published online 17 August, 2015], doi: 10.1177/0962280215597579.

[75] S. Burgess. Consistency and collapsibility: are they crucial for instrumental variable analysis with a survival outcome in Mendelian randomization? *Epidemiology*, 26(3):411–413, 2015.

[76] J. Bowden, G. Davey Smith, and S. Burgess. Mendelian randomization with invalid instruments: Effect estimation and bias detection through egger regression. *International Journal of Epidemiology*, 44(2):512–525, 2015.

[77] J. Bowden, G. Davey Smith, P. C. Haycock, and S. Burgess. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40(4):304–314, 2016.

Tables

Table 1: Estimates of the causal odds ratios for diabetes for a one unit increase in body mass index across 6 cohorts ARIC, CHS, CARDIA, FHS, MEDAL, and MESA (All N=17 057).

Estimator	SE (log OR scale)	z	OR	95% CI
Direct logistic	0.004	29.6	1.14	1.13, 1.15
Logistic TSPS (Stage 1: F=119, R^2=0.007)	0.056	4.96	1.32	1.19, 1.48
Logistic TSRI (unadjusted SE)	0.058	4.79	1.32	1.18, 1.48
Logistic TSRI (robust SE)	0.057	4.86	1.32	1.18, 1.47
Logistic TSRI (TSPS unadjusted SE)	0.056	4.96	1.32	1.18, 1.47
Logistic TSRI (BS 1)	0.057	4.80	1.32	1.18, 1.48
Logistic TSRI (BS 2)	0.061	4.50	1.32	1.17, 1.49
Logistic TSRI (Newey SE)	0.059	4.71	1.32	1.17, 1.48
Logistic TSRI (Terza SE 1)	0.057	4.83	1.32	1.18, 1.47
Logistic TSRI (Terza SE 2)	0.059	4.77	1.32	1.18, 1.48
Logistic SMM	0.101	3.26	1.39	1.19, 1.59
Probit TSRI (on OR scale)	0.090	4.74	1.28	1.15, 1.42

SEs given on log odds ratio scale. Bootstrapping using 500 replications (BS: bootstrap, CI: confidence interval, IV: instrumental variable, OR: odds ratio, SE: standard error, SMM: structural mean model, TSPS: two-stage predictor substitution, TSRI: two-stage residual inclusion).

Table 2: Estimates of the causal effect of a one unit increase in body mass index on systolic blood pressure (mmHg) across 6 cohorts ARIC, CHS, CARDIA, FHS, MEDAL, and MESA (All N=17 057).

Estimator	SE	Estimate	95% CI
Direct linear	0.031	0.76	0.70, 0.82
TSLS (Stage 1: F=119, R^2=0.007)	0.374	0.36	-0.37, 1.10
TSPS (unadjusted SE)	0.378	0.36	-0.38, 1.11
Linear TSRI (unadjusted SE)	0.372	0.36	-0.37, 1.09
Linear TSRI (robust SE)	0.370	0.36	-0.36, 1.09
Linear TSRI (TSPS unadjusted SE)	0.378	0.36	-0.38, 1.11
Linear TSRI (BS 1 SE)	0.376	0.36	-0.37, 1.10
Linear TSRI (BS 2 SE)	0.384	0.36	-0.39, 1.12
Linear TSRI (Newey SE)	0.374	0.36	-0.37, 1.10
Linear TSRI (Terza SE 1)	0.370	0.36	-0.36, 1.09
Linear TSRI (Terza SE 2)	0.372	0.36	-0.37, 1.09

Bootstrapping using 500 replications (BS: bootstrap, CI: confidence interval, SE: standard error, TSLS: two-stage least squares, TSRI: two-stage residual inclusion).

Figure legends

Figure 1: Coverage of the logistic TSRI estimators for $N = 1\,000$ with respect to the conditional parameter, $\beta_1 = 1$. The labels in the legend refer to the type of SE. BS 2: bootstrapping both stages; Newey, Terza 1, and Terza 2 SEs explained in main text; TSRI: two-stage residual inclusion. The panels correspond to α_2 being set to the following values A:0, B:2, C:4, D:6, and E:8.

Figure 2: Coverage of the logistic TSRI estimators for $N = 1\,000$ with respect to the marginal parameter. The labels in the legend refer to the type of SE. BS 2: bootstrapping both stages; Newey, Terza 1, and Terza 2 SEs explained in main text; SE: standard error; TSRI: two-stage residual inclusion. The panels correspond to α_2 being set to the following values A:0, B:2, C:4, D:6, and E:8.

Figure 3: Type I error of the logistic TSRI estimators for $N = 1\,000$. The labels in the legend refer to the type of SE. BS 2: bootstrapping both stages; Newey, Terza 1, and Terza 2 SEs explained in main text; SE: standard error; TSRI: two-stage residual inclusion. The panels correspond to α_2 being set to the following values A:0, B:2, C:4, D:6, and E:8.

Figure 4: Coverage of the linear TSRI estimators for $N = 1\,000$. The labels in the legend refer to the type of SE. BS 2: bootstrapping both stages; SE: standard error; TSLS: two-stage least squares; Newey, Terza 1 and 2, SEs explained in main text. The panels correspond to α_2 being set to the following values A:0, B:2, C:4, D:6, and E:8.

Figure 5: Type I error of the linear TSRI estimators for $N = 1\,000$. The labels in the legend refer to the type of SE. BS 2: bootstrapping both stages; SE: standard error; TSLS: two-stage least squares; Newey, Terza 1 and 2, SEs explained in main text. The panels correspond to α_2 being set to the following values A:0, B:2, C:4, D:6, and E:8.

Figures

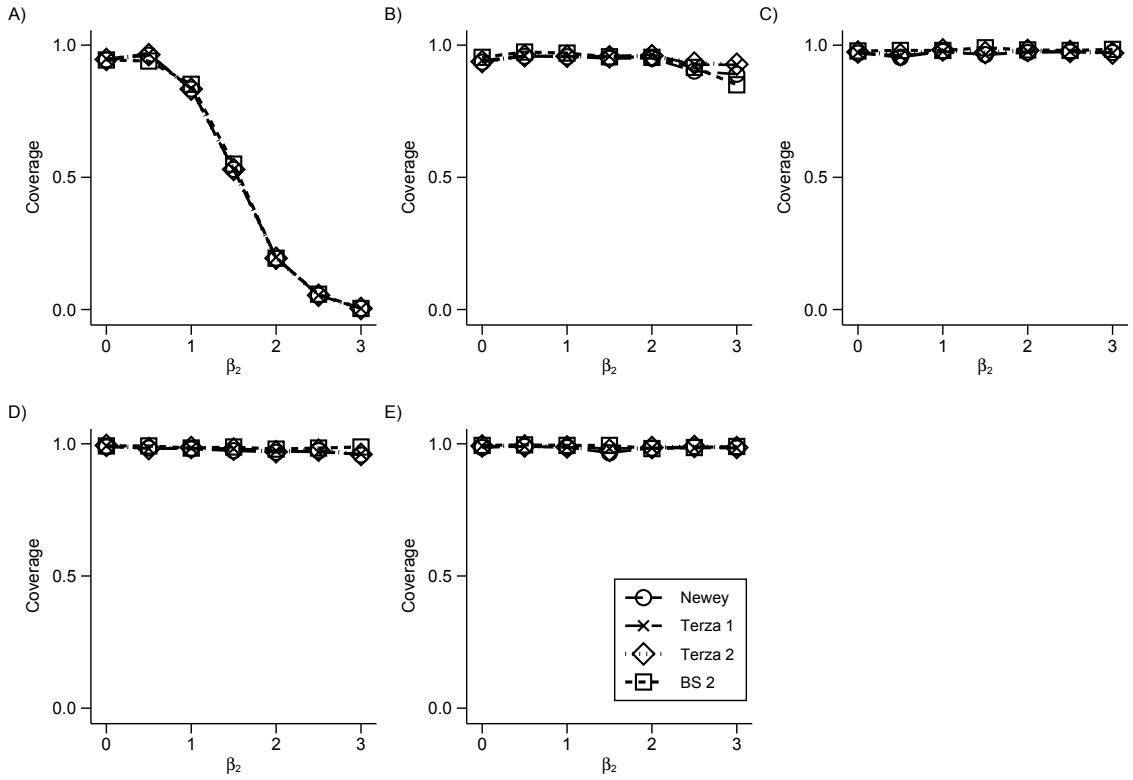


Figure 1

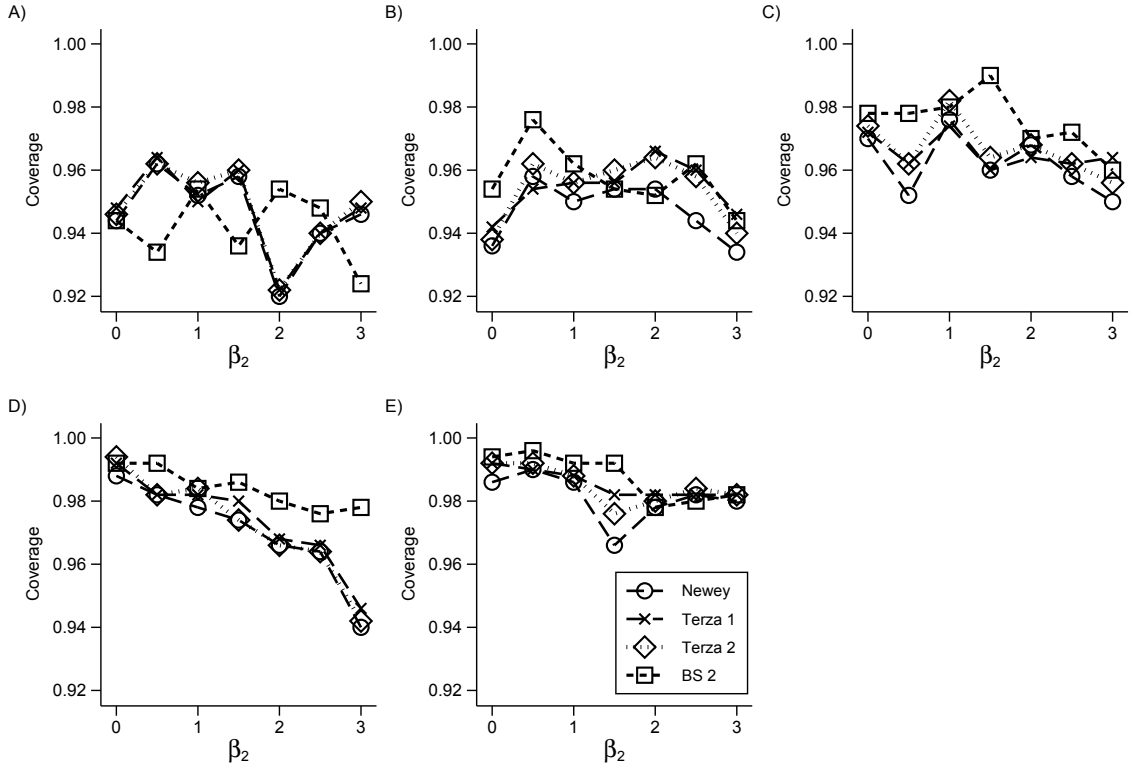

A)
B)
C)
D)
E)
Coverage
1.00
0.98
0.96
0.94
0.92
0
1
2
3
β2
Newey
Terza 1
Terza 2
BS 2

Figure 2

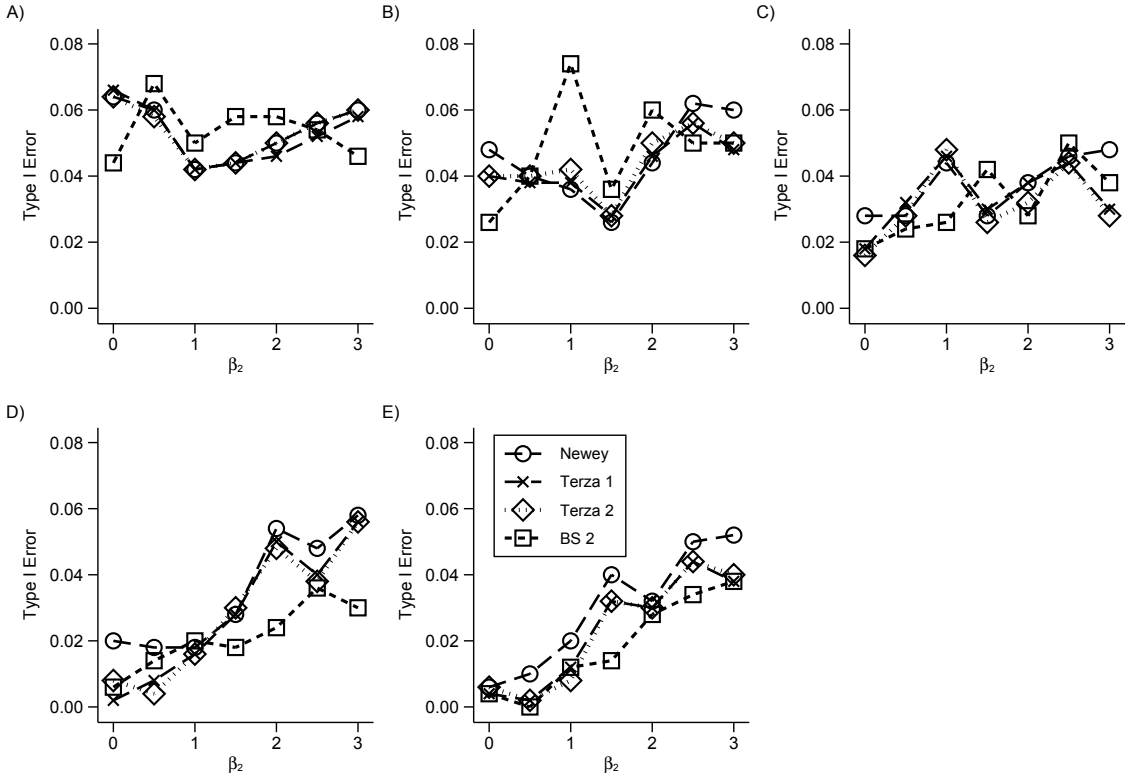

A)
B)
C)
D)
E)
Type I Error
0.00
0.02
0.04
0.06
0.08
0
1
2
3
β₂
Newey
Terza 1
Terza 2
BS 2

Figure 3

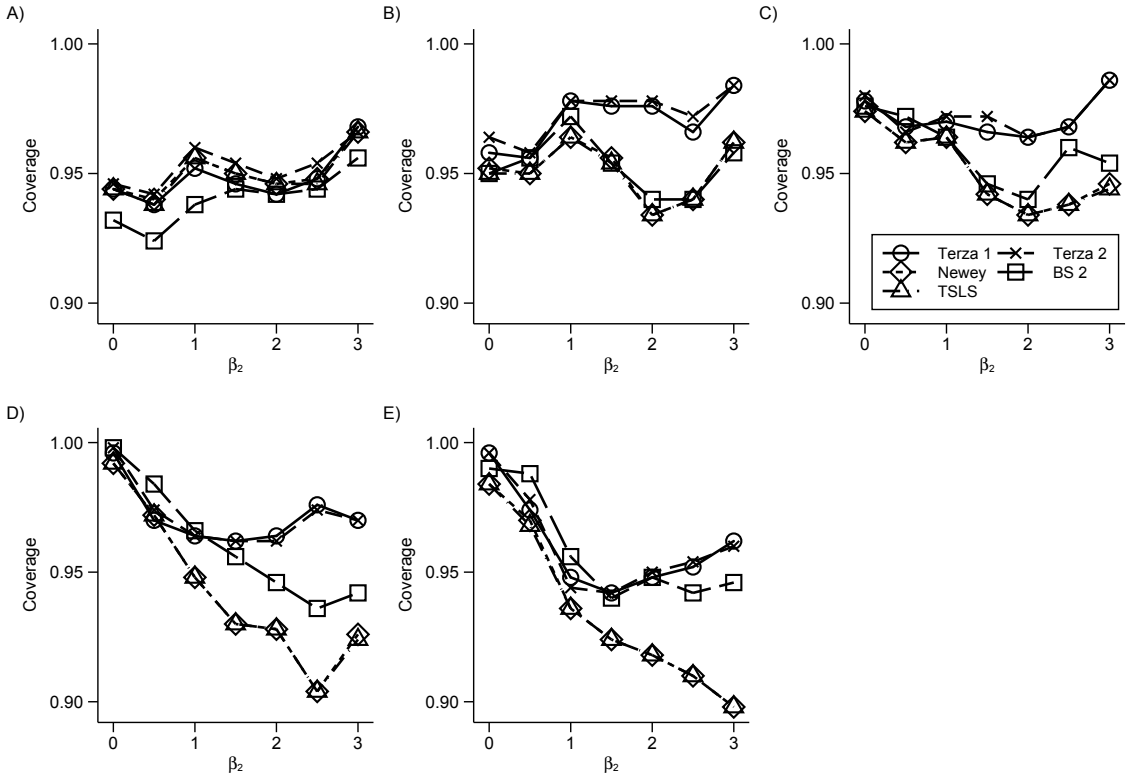

A)
B)
C)
D)
E)
Coverage
1.00
0.95
0.90
0
1
2
3
β₂
Terza 1
Terza 2
Newey
BS 2
TSLS

Figure 4

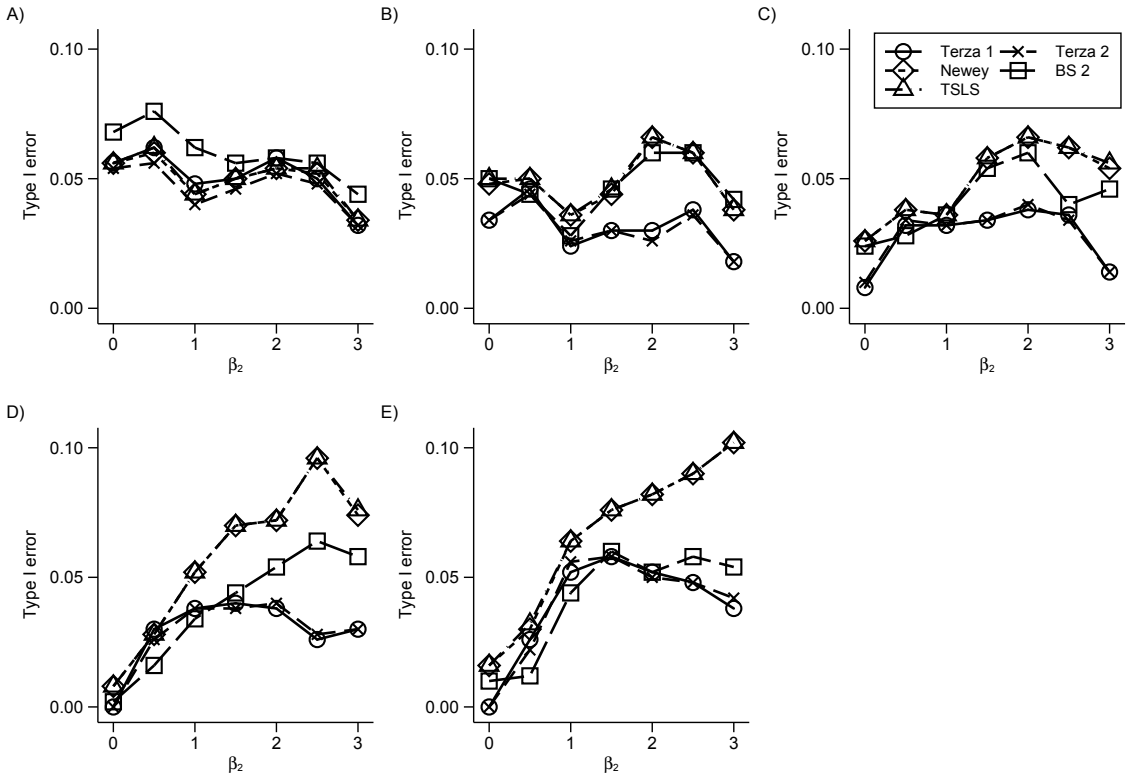

A)
B)
C)
D)
E)
Type I error
0.10
0.05
0.00
0
1
2
3
β₂
Terza 1
Terza 2
Newey
BS 2
TSLS

Figure 5

Supplementary material for Palmer TM, Holmes MV, Keating BJ, and Sheehan NA. Correcting the standard errors of two-stage residual inclusion estimators for Mendelian randomization studies. American Journal of Epidemiology

Web Appendix 1: The difference between unadjusted and corrected standard errors for TSRI estimators

Linear estimators

We consider the case of two-stage least squares. The true underlying model for the data is,

$$\begin{aligned} g_i &\sim \text{Binomial}(2,\ p_g) \\ x_i &= \alpha_0 + \alpha_1 g_i + \varepsilon_{1i}, \\ y_i &= \beta_0 + \beta_1 x_i + \varepsilon_{2i}, \end{aligned} \qquad \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim \text{MVN}\left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right) \tag{A9}$$

However, the models fitted in TSLS estimation are,

$$x_i = \alpha_0 + \alpha_1 g_i + \varepsilon_{3i} \tag{A10}$$

$$y_i = \beta_0 + \beta_1 \widehat{x}_i + \varepsilon_{4i} \tag{A11}$$

After the fitting the second stage manually the variance of our vector of causal effect estimates $\widehat{\beta}$ is given below, where $\widehat{X}$ denotes a matrix made up of the predicted values of X and a column of 1s for the intercept, N the number of observations and k the number of covariates,

$$\text{var}(\widehat{\beta}) = s^2(\widehat{X}'\widehat{X})^{-1} \tag{A12}$$

$$\text{where } s^2 = \frac{\sum_{i=1}^{N}(Y - \widehat{X}\widehat{\beta})^2}{(N-k)}. \tag{A13}$$

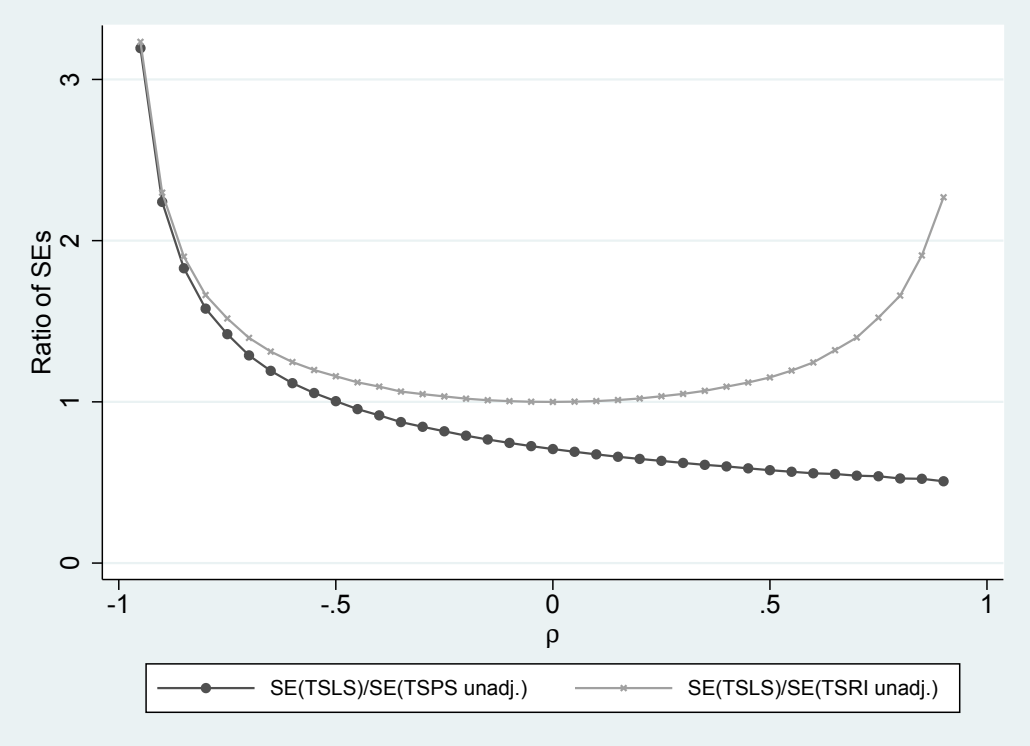
This gives us incorrect standard errors on our causal effects because s^2 is in terms of $\widehat{X}$ whereas our causal model is in terms of X. Hence the corrected variance of $\widehat{\beta}$ is given by,

$$s^2 = \frac{\sum_{i=1}^{N}(Y - X\widehat{\beta})^2}{N}. \tag{A14}$$

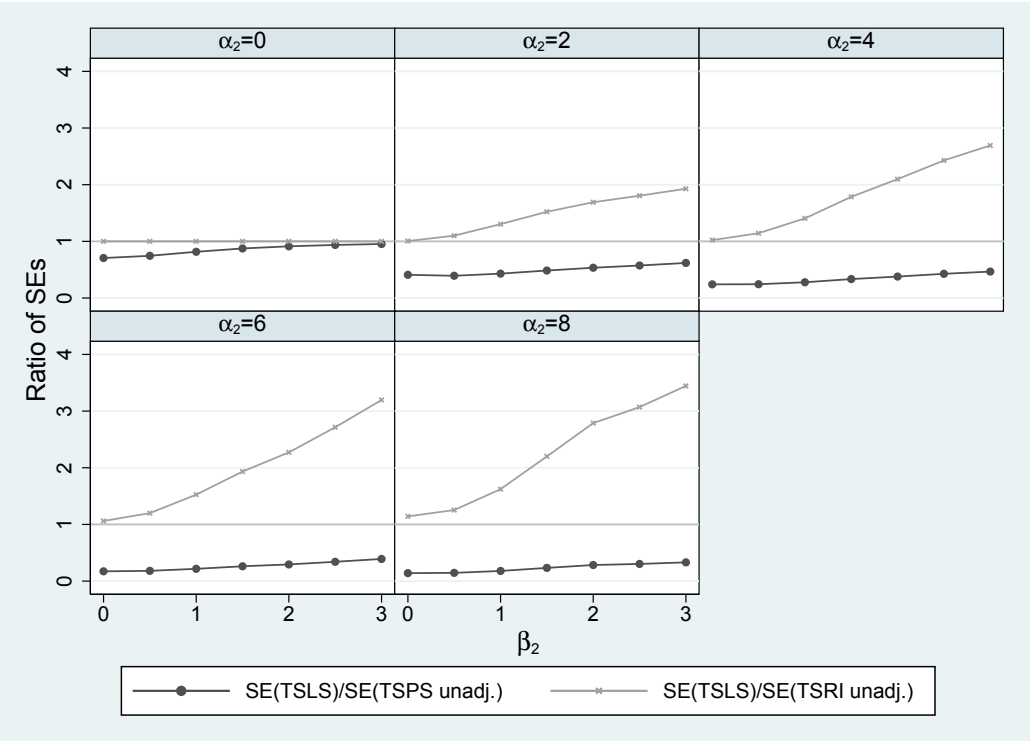
By comparing the numerators of the two terms for s^2 it is apparent that the corrected standard errors will be similar to the uncorrected standard errors when $\widehat{X}$ are close to their observed values X.

We simulated data based on Equation A9 using $p_g = 0.3$, $\alpha_0 = 0$, $\alpha_1 = 1$, $\beta_0 = 0$, and $\beta_1 = 1$. Figure 1a shows that the corrected (TSLS) standard errors are larger than TSPS unadjusted standard errors for $\rho < -0.5$ and that the corrected standards are always larger than TSRI unadjusted standard errors, this second curve being symmetrical about $\rho = 0$ and only starts to reach ratios above 1.1 for $|\rho| > 0.5$.

Figure 1b shows that in the simulations in the main text the corrected standard errors were always less than the TSPS unadjusted standard errors. For $\alpha_2 = 0$, i.e. no unmeasured confounding in the first stage model, the corrected standard errors were the same as the TSRI unadjusted standard errors, for all other values of α_2 the corrected standard errors were up to 3.5 times larger.



(a) Average SEs in simulations with N=1 000 using 50 replications.



(b) Average SEs in the linear simulations with N=1 000.

Web Figure 1: Ratio of TSLS SEs to unadjusted TSPS and TSRI SEs (SE: standard error; TSLS: two-stage least squares; TSPS: two-stage predictor substitution; TSRI: two-stage residual inclusion).

Logistic estimators

For the theoretical example for the logistic estimators we use the following model,

$$\begin{aligned} g_i &\sim \text{Binomial}(2,\ p_g) \\ x_i &= \alpha_0 + \alpha_1 g_i + \varepsilon_{1i}, \\ \log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 x_i + \varepsilon_{2i}, \qquad \begin{pmatrix}\varepsilon_1 \\ \varepsilon_2\end{pmatrix} \sim \text{MVN}\left(0, \begin{bmatrix}1 & \rho \\ \rho & 1\end{bmatrix}\right) \\ y_i &\sim \text{Bernoulli}(p_i). \end{aligned} \tag{A15}$$

For logistic regression the variance of the parameter estimates is given by the following; where X is the design matrix of covariates including a vector of 1s for the intercept, I_N is an N by N identity matrix, $\widehat{p}$ is a vector of predicted probabilities of the outcome from the model, $\circ$ denotes element-wise multiplication, and diag() extracts the diagonal elements of a matrix,

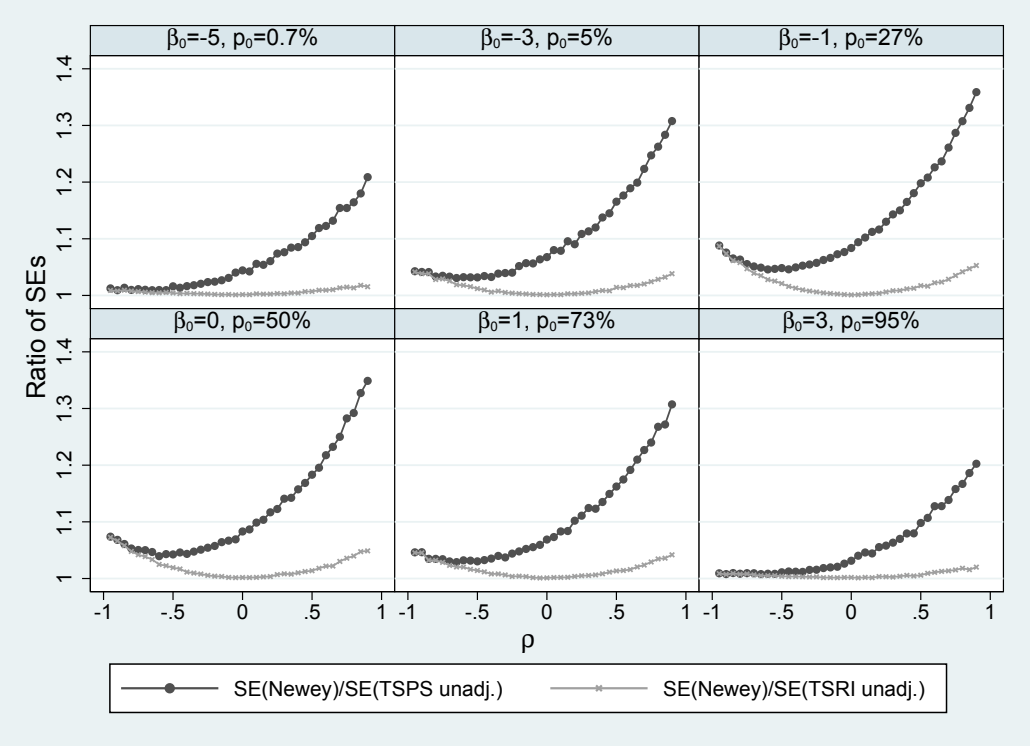
$$\text{var}(\widehat{\beta}) = (X'VX)^{-1} \tag{A16}$$

$$V = I_N \circ \text{diag}(\widehat{p}(1-\widehat{p})'). \tag{A17}$$

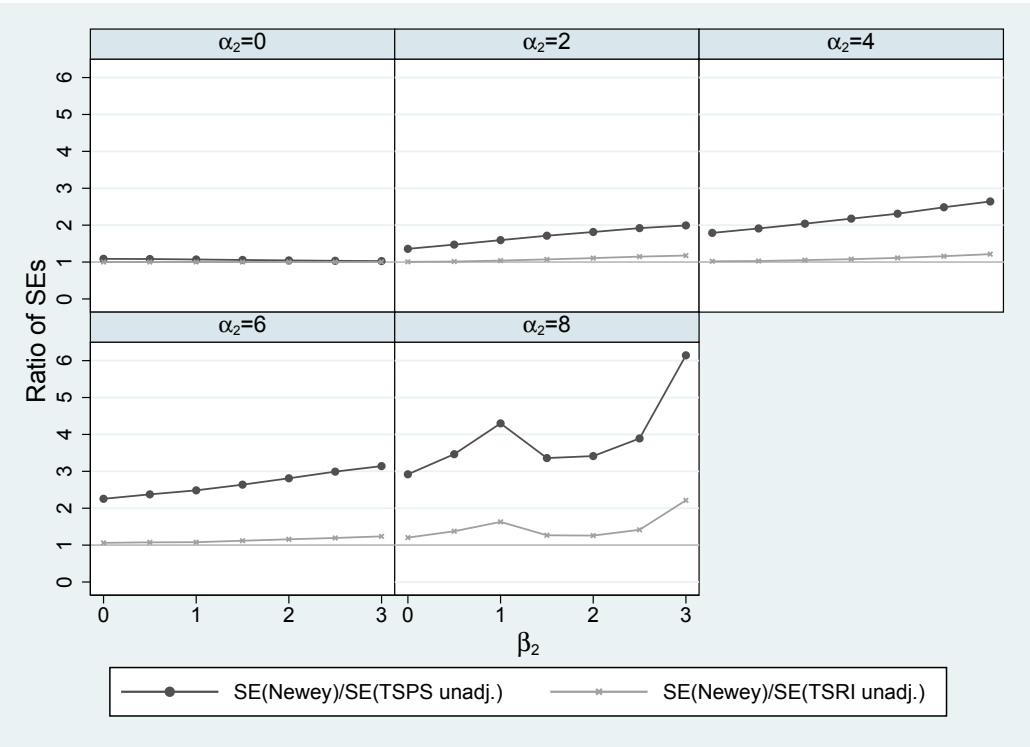
Hence, the variance of the estimates is affected by the values of the covariates in the model and also by the predicted probabilites of the outcome. The variance is maximised for predicted probabilities at 50%. Hence, it might be reasonable to expect the difference between the unadjusted and corrected TSRI standard errors to be greatest when the standard errors are greatest (i.e. for prevalences around 50%) and also when the unmeasured confounding is stronger (since this would mean there is more uncertainty around the predicted values of the first stage residuals).

This can be seen in Figure 2 which shows results for data simulated under the model in Equation A15 setting $p_g = 0.3$, $\alpha_0 = 0$, $\alpha_1 = 1$, and $\beta_1 = 1$. The value of β_0 was set from -5 to 3 to change the prevalence of the outcome from around 0.7% up to around 95%. In general the unadjusted standard errors are closer to the Newey standard errors when the prevalence is further away from 50%. The TSRI unadjusted standard errors are closer to the Newey standard errors than the TSPS unadjusted standard errors. The ratio of the Newey standard errors to the TSRI unadjusted standard errors is approximately symmetric about $\rho = 0$ whereas the ratio of the Newey standard errors to the unadjusted TSPS standard errors is constant or reaches a minimum for $\rho < -0.5$ and then increases as ρ increases.

Figure 3 shows the ratio of the Newey standard errors to the unadjusted TSPS and TSRI standard errors in the simulations in the main text with N=1 000. For $\alpha_2 = 0$ all three standard errors are approximately equal. For the other values of α_2 the Newey standard errors larger than both the unadjusted standard errors. The unadjusted TSRI standard errors are closer to Newey standard errors than the TSPS unadjusted standard errors. Plots using Terza 1 and 2 standard errors were very similar.



Web Figure 2: Ratio of logistic TSRI Newey SEs to unadjusted logistic TSPS and TSRI SEs. These are average SEs in simulations with N=1 000 using 50 replications (SE: standard error; TSLS: two-stage least squares; TSPS: two-stage predictor substitution; TSRI: two-stage residual inclusion).



Web Figure 3: Ratio of logistic TSRI Newey SEs to unadjusted logistic TSPS and TSRI SEs in the logistic simulations with N=1 000 (SE: standard error; TSLS: two-stage least squares; TSPS: two-stage predictor substitution; TSRI: two-stage residual inclusion).

Web Appendix 2: Example Stata and R code implementing Newey and Terza standard errors for two-stage residual inclusion estimators

Newey standard errors: Stata code

In this code we assume that the exposure, outcome, and instrumental variables are named `x`, `y`, and `g` respectively.

The code starts by fitting the first stage model to generate variables containing the predicted values (`xb`) and residuals (`res`). We then create a matrix in Stata's Mata environment called `DPi`.

```
* first stage
regress x g
capture noisily drop xb
predict double xb, xb
capture noisily drop res
predict double res, res

* DPi matrix
mata DPi = I(2)
mata b1 = st_matrix("e(b)")
mata DPi[,1] = b1'
```

Linear TSRI estimator

For the linear TSRI estimator the code proceeds following the steps described for the Probit TSRI estimator in the Methods section replacing the Probit regressions in steps 2 and 3 with linear regressions.

```
* 2. regress to solve (2)
cap noi regress y g res
mata gamma = st_matrix("e(b)")[(1,3)]
mata lambdahat = st_matrix("e(b)")[2]
mata J1inv = st_matrix("e(V)")[(1,3),(1,3)]

* 3. Evaluate linear TSRI estimator
regress y x res
mata cfivb = st_matrix("e(b)")
mata cfivV = st_matrix("e(V)")
mata betahat = st_matrix("e(b)")[1]
putmata x=x, replace
mata y2new = x:*(lambdahat - betahat)
drop y2new
getmata y2new, double replace

* 4. y2*(lambda - beta) is regressed on z - the cov matrix is added to J1inv
regress y2new g
mata S2 = st_matrix("e(V)")
mata Omega = J1inv :+ S2

* evaluating equations 4 and 5 yields psihat and var(psihat)
mata finalv = invsym(DPi' * invsym(Omega) * DPi)
```

```
mata finalv
mata neweyse = sqrt(diagonal(finalv)[1])
mata printf("Newey SE= %9.0g", neweyse)
mata neweylow = cfivb[1] - invnormal(.975)*sqrt(finalv[1,1])
mata neweyupp = cfivb[1] + invnormal(.975)*sqrt(finalv[1,1])
mata printf("95 percent CI using Newey SE: (%9.0g , %9.0g)", neweylow, neweyupp)
mata finalb = finalv * DPi' * invsym(Omega) * gamma'
mata finalb
```

Logistic TSRI

The code for the logistic TSRI is identical to that for the linear TSRI except that the linear regressions in steps 2 and 3 are replaced by logistic regressions as follows.

```
* 2. logit to solve (2)
logit y g res, nolog

* 3. Evaluate logistic TSRI
logit y x res, nolog
```

Poisson TSRI

The code for the Poisson TSRI is identical to that for the linear TSRI except that the linear regressions in steps 2 and 3 are replaced by Poisson regressions as follows.

```
* 2. poisson to solve (2)
poisson y g res, nolog

* 3. Evaluate Poisson TSRI
poisson y x res, nolog
```

Gamma TSRI

The code for the Gamma TSRI is identical to that for the linear TSRI except that the linear regressions in steps 2 and 3 are replaced by Gamma regressions with log links as follows.

```
* 2. gamma regression to solve (2)
glm y g res, fam(gamma) link(log)

* 3. Evaluate Gamma TSRI
glm y x res, fam(gamma) link(log)
```

Newey standard errors: R code

In this code we assume that the exposure, outcome, and instrumental variables are in a data-frame named `data` and are called `x`, `y`, and `g` respectively.

The code starts by attaching the data-frame into the workspace and fitting the first stage model to generate variables containing the predicted values (`xb`) and residuals (`res`). We then create the `DPi` matrix.

```
attach(data)

first <- lm(x ~ g)
xb <- fitted.values(first)
res <- residuals(first)

# DPi matrix
DPi <- diag(2)
DPi[,1] <- rev(as.matrix(coef(first)))
```

Linear TSRI

For the linear TSRI the code proceeds following the steps described for the Probit TSRI in the Methods section replacing the Probit regressions in steps 2 and 3 with linear regressions.

```
# 2. regression to solve (2)
step2 <- lm(y ~ g + res)
gamma <- t(rev(coef(step2)[-3]))
lambdahat <- coef(step2)[3]
J1inv <- vcov(step2)[-3,-3]
J1inv <- J1inv[c(2,1),c(2,1)]

# 3. evaluate linear TSRI
second <- lm(y ~ x + res)
y2new <- x*(lambdahat - coef(second)[2])

# 4. y2new*(lambda - beta) is regressed on z
four <- lm(y2new ~ g)
S2 <- vcov(four)
S2 <- S2[c(2,1),c(2,1)]
Omega <- J1inv + S2

finalv <- solve(t(DPi) %*% solve(Omega) %*% DPi)
finalv
finalb <- finalv %*% t(DPi) %*% solve(Omega) %*% t(gamma)
finalb
```

Logistic TSRI

The code for the logistic TSRI is identical to that for the linear TSRI except that the linear regressions in steps 2 and 3 are replaced by logistic regressions as follows.

```
# 2. logistic regression to solve (2)
step2 <- glm(y ~ g + res, family=binomial(logit))

# 3. evaluate logistic TSRI
second <- glm(y ~ x + res, family=binomial(logit))
```

Poisson TSRI

The code for the Poisson TSRI is identical to that for the linear TSRI except that the linear regressions in steps 2 and 3 are replaced by Poisson regressions as follows.

```
# 2. Poisson regression to solve (2)
step2 <- glm(y ~ g + res, family=poisson(log))

# 3. evaluate Poisson TSRI
second <- glm(y ~ x + res, family=poisson(log))
```

Gamma TSRI

The code for the Gamma TSRI is identical to that for the linear TSRI except that the linear regressions in steps 2 and 3 are replaced by Gamma regressions as follows. Also in R we must replace the zero values in `y` with a small value (e.g. 0.0001) since all values must be greater than zero.

```
# 2. Gamma regression to solve (2)
y[y==0] <- 1E-4
step2 <- glm(y ~ g + res, family=Gamma(log), control=list(maxit=1000))

# 3. evaluate Gamma TSRI
second <- glm(y ~ x + res, family=Gamma(log), control=list(maxit=1000))
```

Terza standard errors: Stata code

Terza (2016) provides Stata code for his method of estimating the standard error for TSRI estimators.[38] When fitting these models we recommend centering the covariates about their means in both stages of estimation.

Terza standard errors: R code

This code uses the linear TSRI estimator (calculating the standard error for the other estimators proceeds similar but changes the second stage model).

```
# first stage
first <- lm(x ~ g)
expWalpha <- fitted.values(first) # xb
res <- residuals(first) # xuhat
alpha <- coef(first)
covalpha <- vcov(first)

# TSRI
second <- lm(y ~ x + res)
expXbeta <- fitted.values(second)
beta <- coef(second)
covbeta <- vcov(second)
bxu <- beta[3]

W <- cbind(1, g)
X <- cbind(1, x, res)
Xbeta <- X %*% beta

# Compute the asymptotic covariance matrix of
# the 2SRI estimate of beta.
```

```
paJ <- -bxu * expXbeta * expWalpha * W
pbJ <- expXbeta * X
Bba <- t(pbJ) %*% paJ
Bbb <- t(pbJ) %*% pbJ
d22 <- solve(Bbb) %*% Bba %*% covalpha %*% t(Bba) %*% solve(Bbb) + covbeta

# Terza standard errors
ses <- sqrt(diag(d22))

# t-statistics
tstats <- beta / ses

# pvalues
pvalues <- 2 * pnorm(-1*abs(tstats))

# estimates with 95% CI limits
cbind(beta, beta - 1.96*ses, beta + 1.96*ses)

# alternative terza standard errors using robust first and second stage SEs
library(sandwich)
covalpha2 <- vcovHC(first, type="HC1")
covbeta2 <- vcovHC(second, type="HC1")
d222 <- solve(Bbb) %*% Bba %*% covalpha2 %*% t(Bba) %*% solve(Bbb) + covbeta2
ses2 <- sqrt(diag(d222))
cbind(beta, beta - 1.96*ses2, beta + 1.96*ses2)
```

Web Appendix 3: Marginal parameter values for the logistic TSPS and TSRI estimators

Using the notation in Equation 7 and as per the appendix of Palmer et al. (2008) we define the marginal parameter value (β_{1m}) estimated by the logistic TSPS and TSRI estimators, and the direct logistic regression below.[17]

$$\beta_{1m} = \beta_1 \frac{1}{\sqrt{1 + c^2 V}}, \quad \text{where } c = \frac{16\sqrt{3}}{15\pi}. \tag{A18}$$

Where σ_1^2 denotes the variance of the residuals in the first stage regression, for the logistic TSPS estimator V is given by,

$$V = (\beta_1\alpha_2 + \beta_2)^2 + \beta_1^2\sigma_1^2. \tag{A19}$$

For the logistic TSRI estimator V is given by,

$$V = (\beta_1\alpha_2 + \beta_2)^2 + \beta_1^2\sigma_1^2 - \frac{(\alpha_2(\beta_1\alpha_2 + \beta_2) + \beta_1\sigma_1^2)^2}{\alpha_2^2 + \sigma_1^2}. \tag{A20}$$

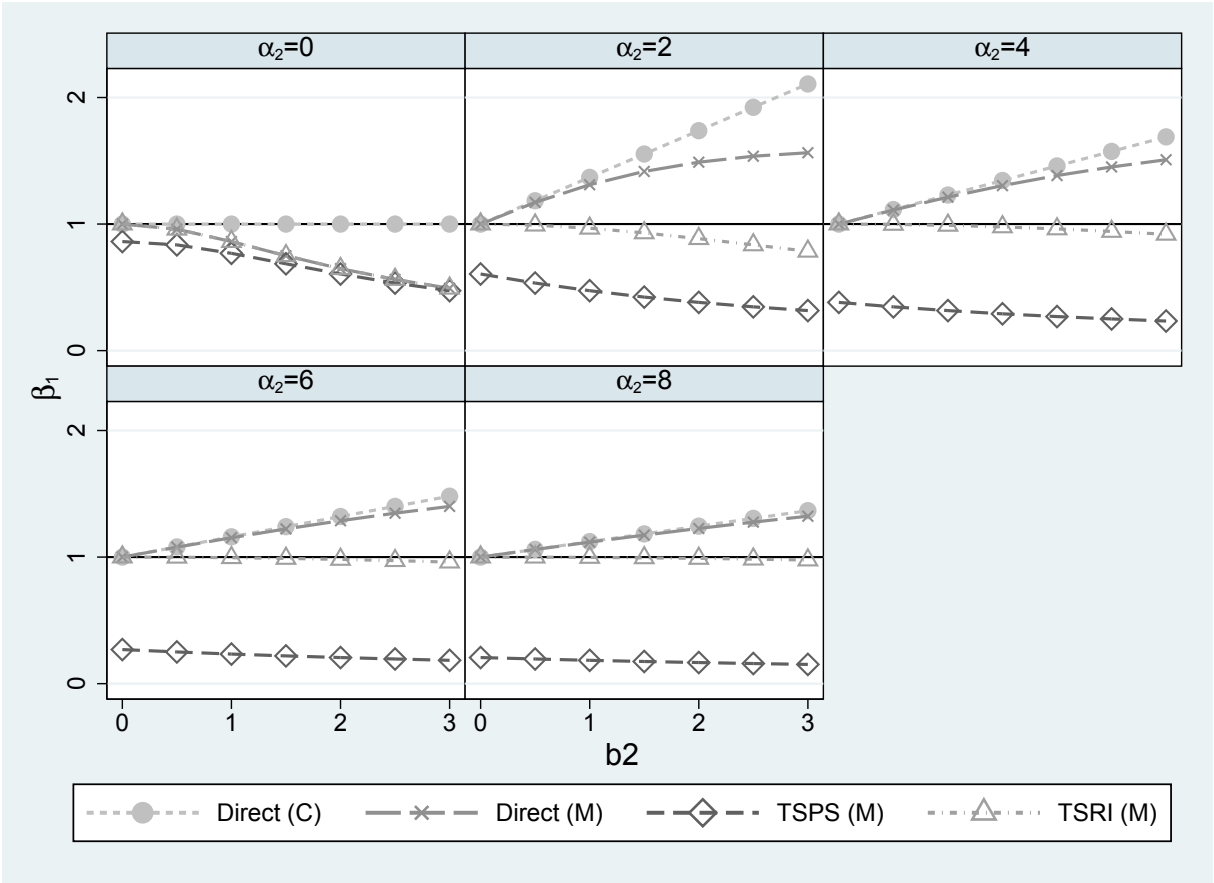
For the direct logistic regression of Y on X we need, where p_g is the minor allele frequency of the genetic variant used as the single instrumental variable and β_{1c} is the value of the conditional effect for the other two estimators (i.e. set as 1 in these simulations),

$$V_g = 2p_g(1 - p_g) \tag{A21}$$

$$\beta_1 = \beta_{1c} + \frac{\alpha_2\beta_2}{\alpha_1^2 V_g + \alpha_2^2 + \sigma_1^2} \tag{A22}$$

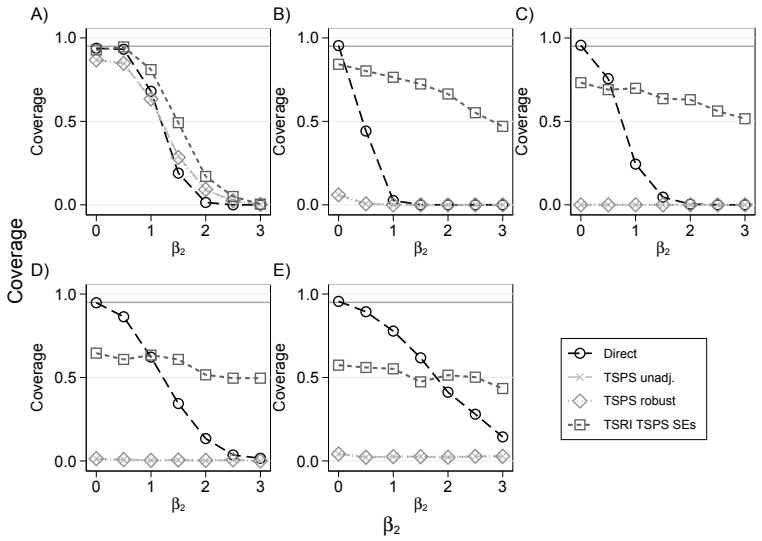
$$V = (\alpha_1\beta_{1c})^2 V_g + (\beta_{1c}\alpha_2 + \beta_2)^2 + \beta_{1c}^2\sigma_1^2 - \frac{(\alpha_1^2\beta_{1c}V_g + \alpha_2(\beta_{1c}\alpha_2 + \beta_2) + \beta_{1c}\sigma_1^2)^2}{\alpha_1^2 V_g + \alpha_2^2 + \sigma_1^2}. \tag{A23}$$

The values of β_{1m} and β_{1c} for the three estimators are shown in Figure 4. The marginal TSRI estimate is much closer to the conditional value of 1 than the marginal TSPS estimate.

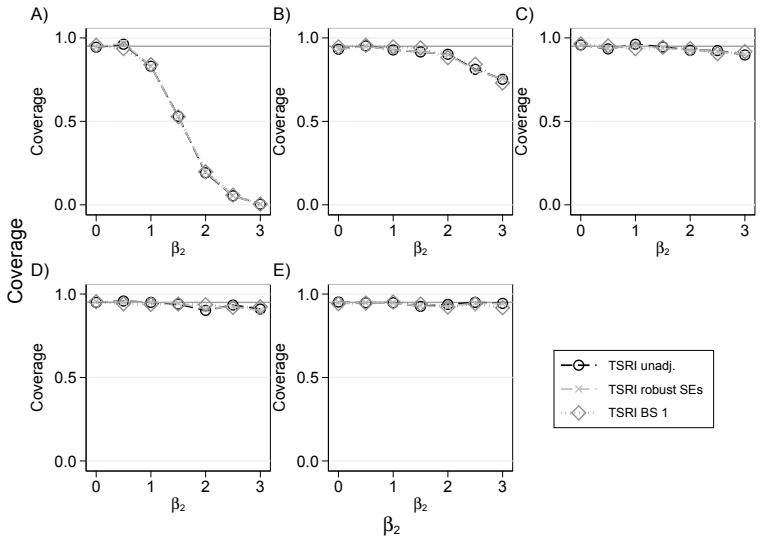


Web Figure 4: Values of the marginal (M) and conditional (C) parameters of the direct logistic regression of Y on X, logistic two-stage predictor substitution (TSPS), and logistic two-stage residual inclusion (TSRI) estimators used in the simulations.

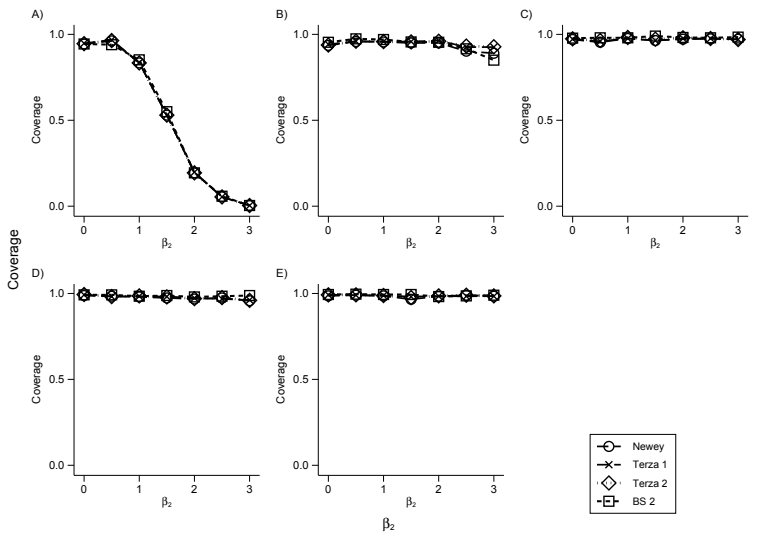
Web Appendix 4: Additional simulation results



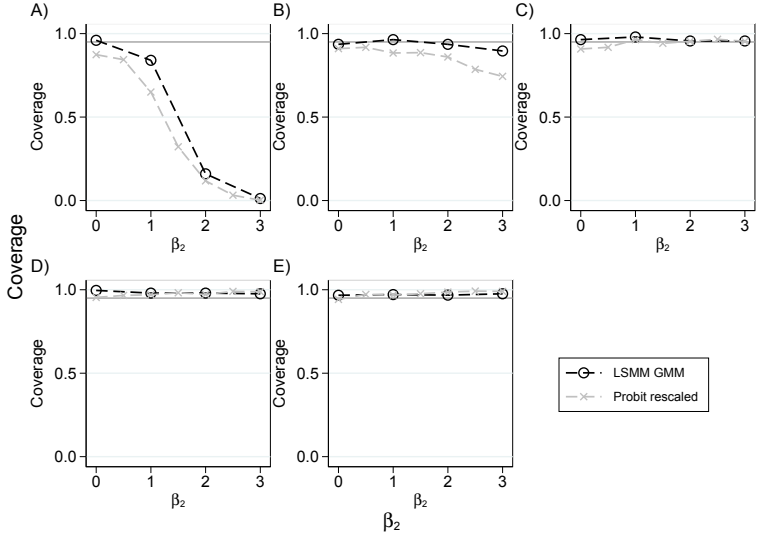
(a)



(b)

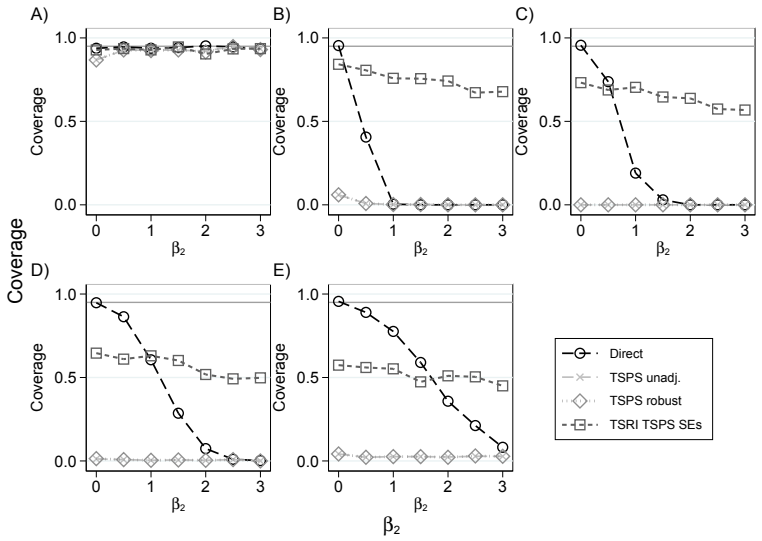


(c)

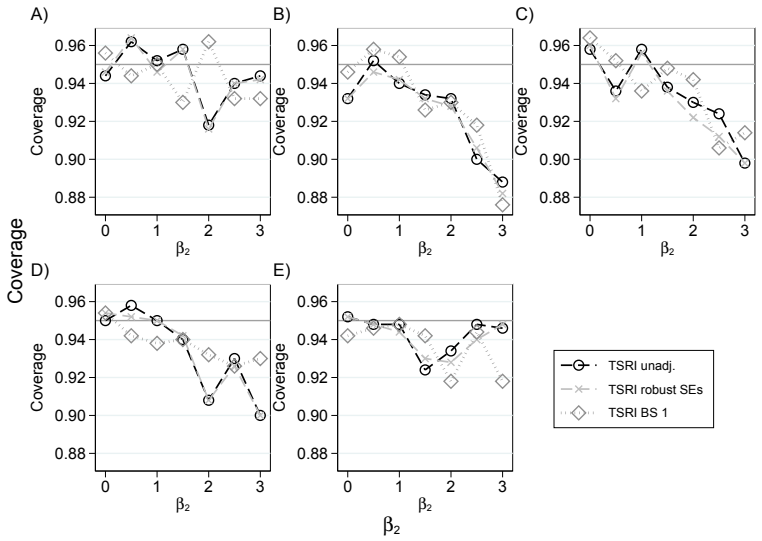


(d)

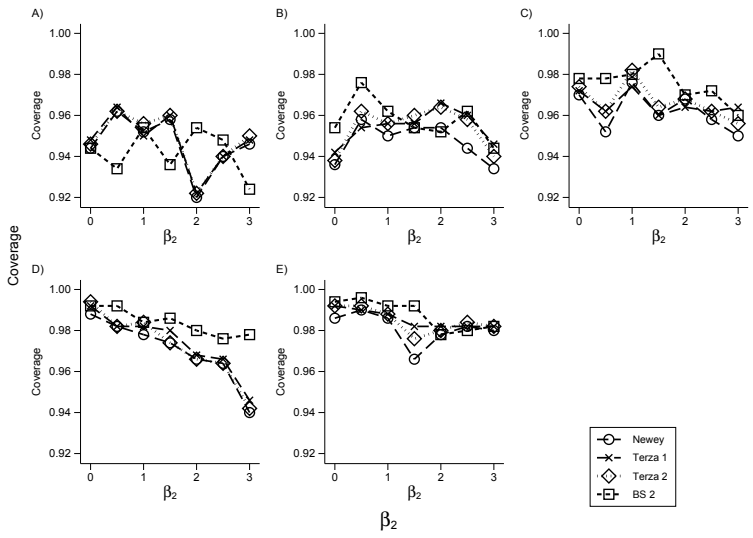
Web Figure 5: Coverage of the logistic estimators for $N = 1\,000$ with respect to the conditional parameter, $\beta_1 = 1$ (BS: bootstrap; LSMM: logistic structural mean model; GMM: generalized method of moments; SE: standard error; TSPS: two-stage predictor substitution; TSRI: two-stage residual inclusion). The panels correspond to α_2 being set to the following values A:0, B:2, C:4, D:6, and E:8.



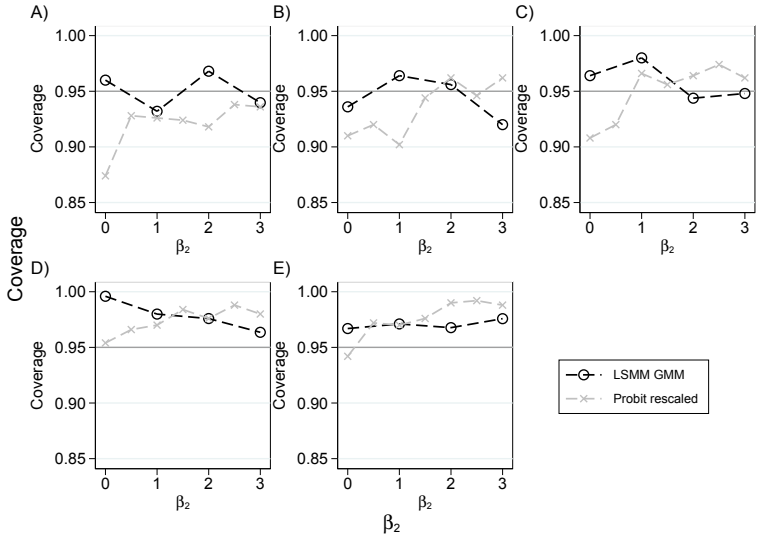
(a)



(b)

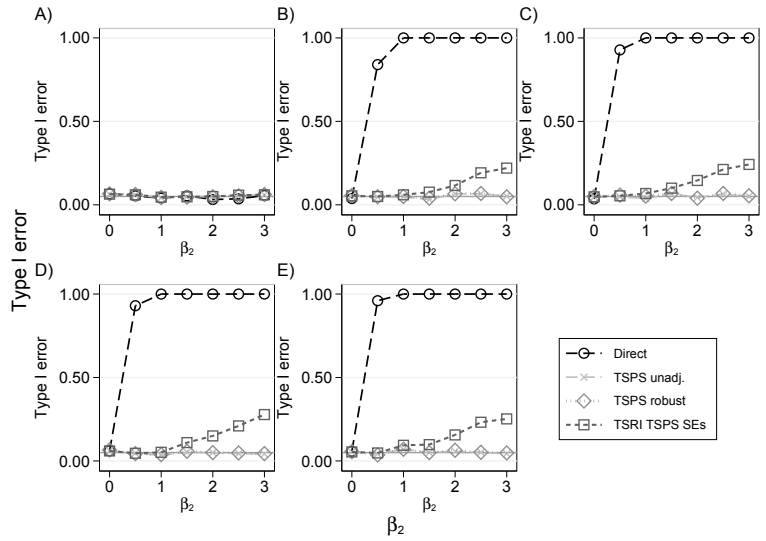


(c)

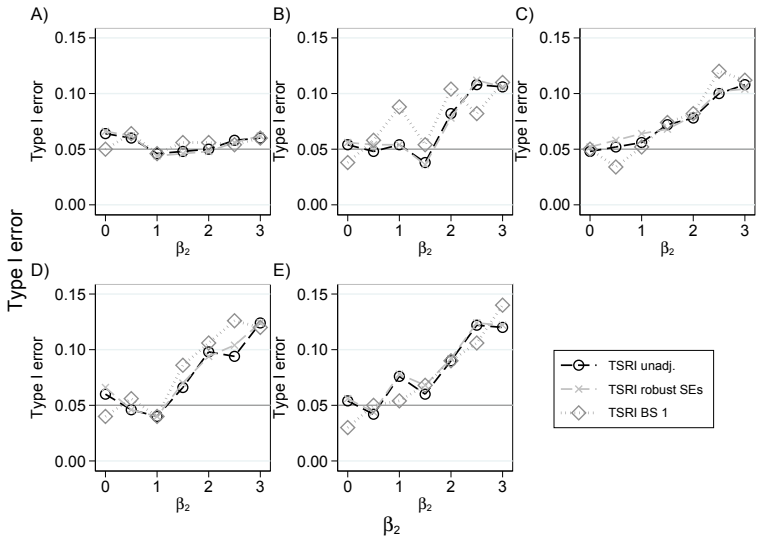


(d)

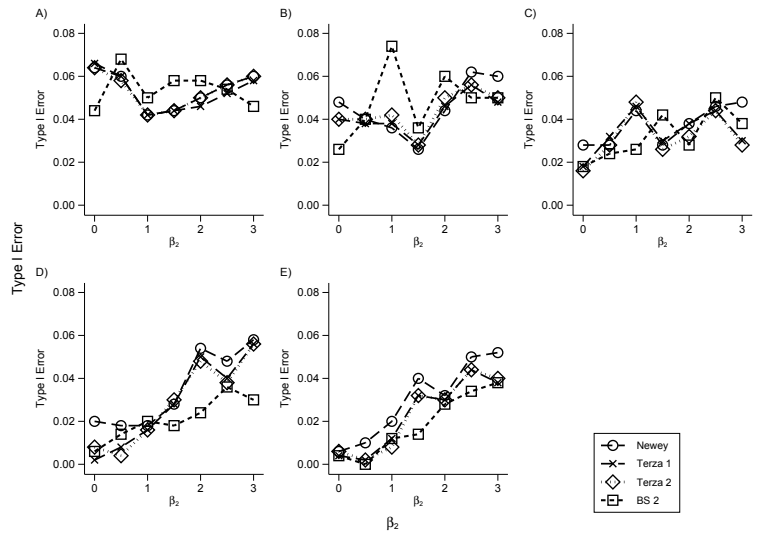
Web Figure 6: Coverage of the logistic estimators for $N = 1\,000$ with respect to the marginal parameter estimated by the TSRI estimator (BS: bootstrap; LSMM: logistic structural mean model; GMM: generalized method of moments; SE: standard error; TSPS: two-stage predictor substitution; TSRI: two-stage residual inclusion). The panels correspond to α_2 being set to the following values A:0, B:2, C:4, D:6, and E:8.



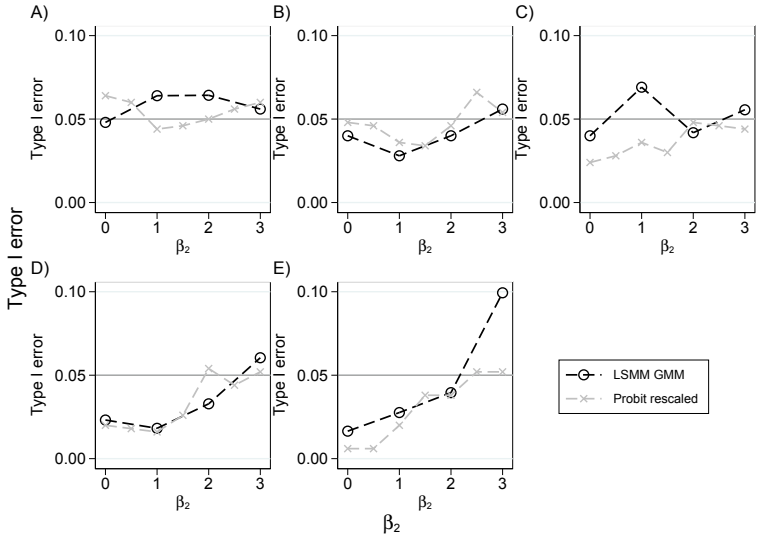
(a)



(b)

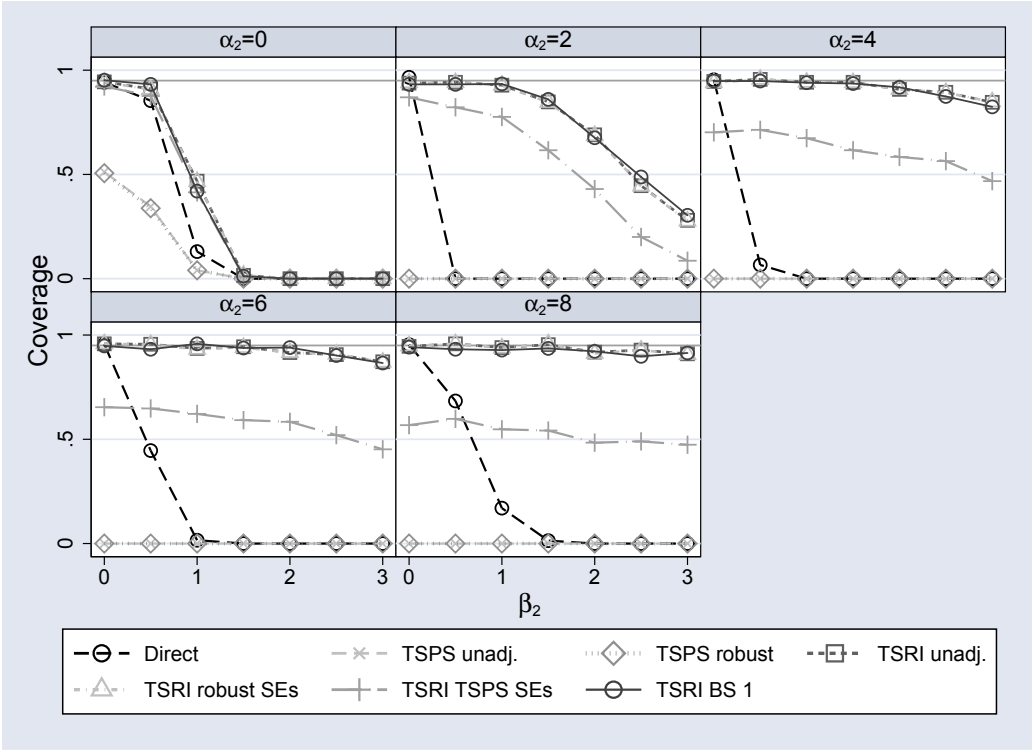


(c)

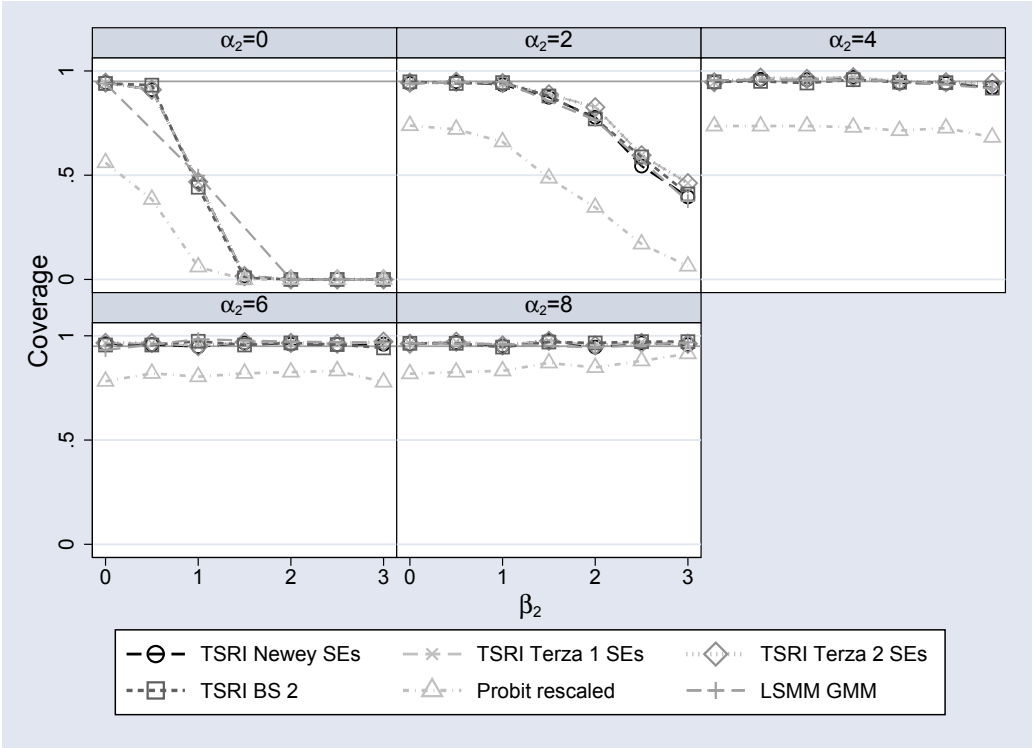


(d)

Web Figure 7: Type I error of the logistic estimators for $N = 1\,000$ (BS: bootstrap; LSMM: logistic structural mean model; GMM: generalized method of moments; SE: standard error; TSPS: two-stage predictor substitution; TSRI: two-stage residual inclusion). The panels correspond to α_2 being set to the following values A:0, B:2, C:4, D:6, and E:8.

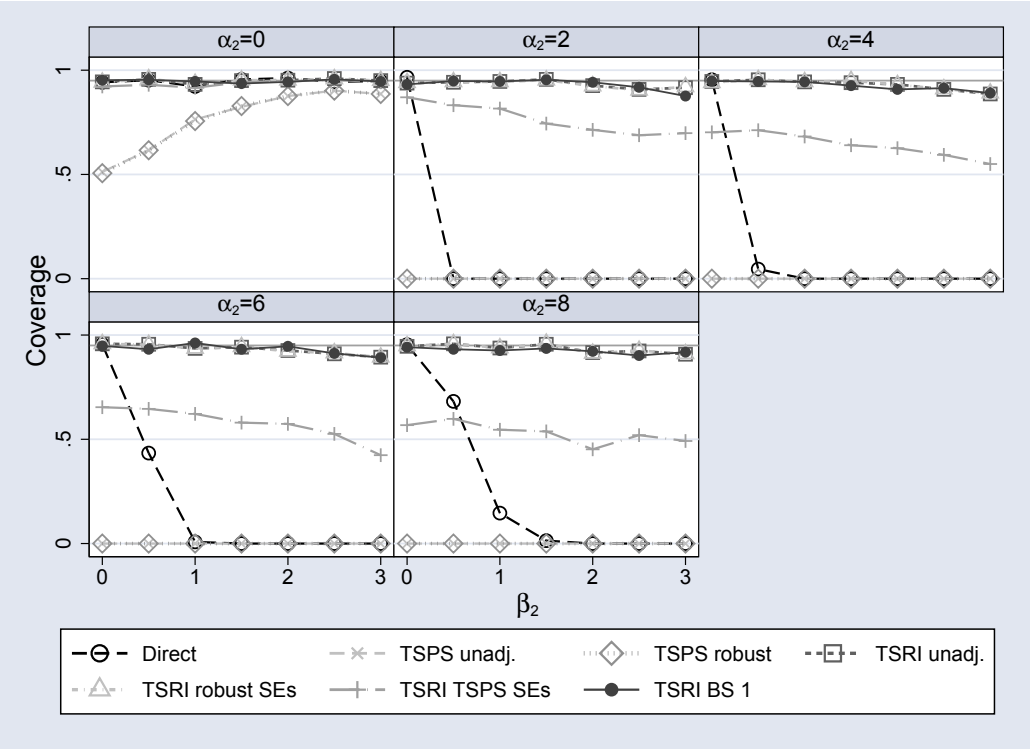


(a)

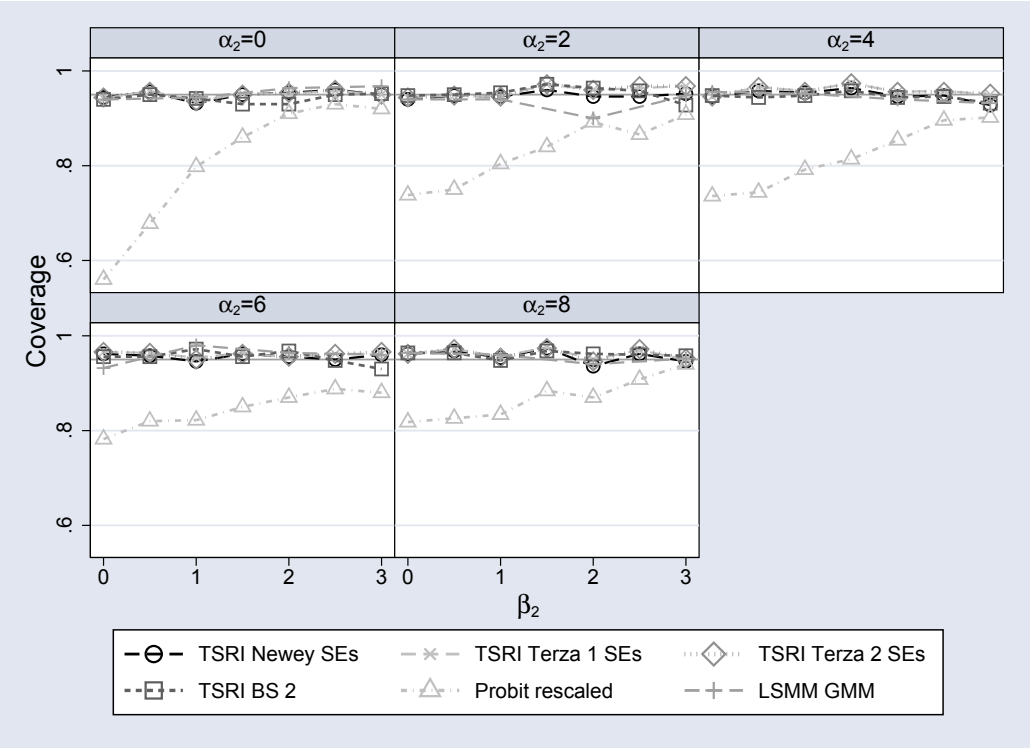


(b)

Web Figure 8: Coverage of the logistic estimators for $N = 5\,000$ with respect to the conditional parameter, $\beta_1 = 1$ (BS: bootstrap; LSMM: logistic structural mean model; GMM: generalized method of moments; SE: standard error; TSPS: two-stage predictor substitution; TSRI: two-stage residual inclusion).

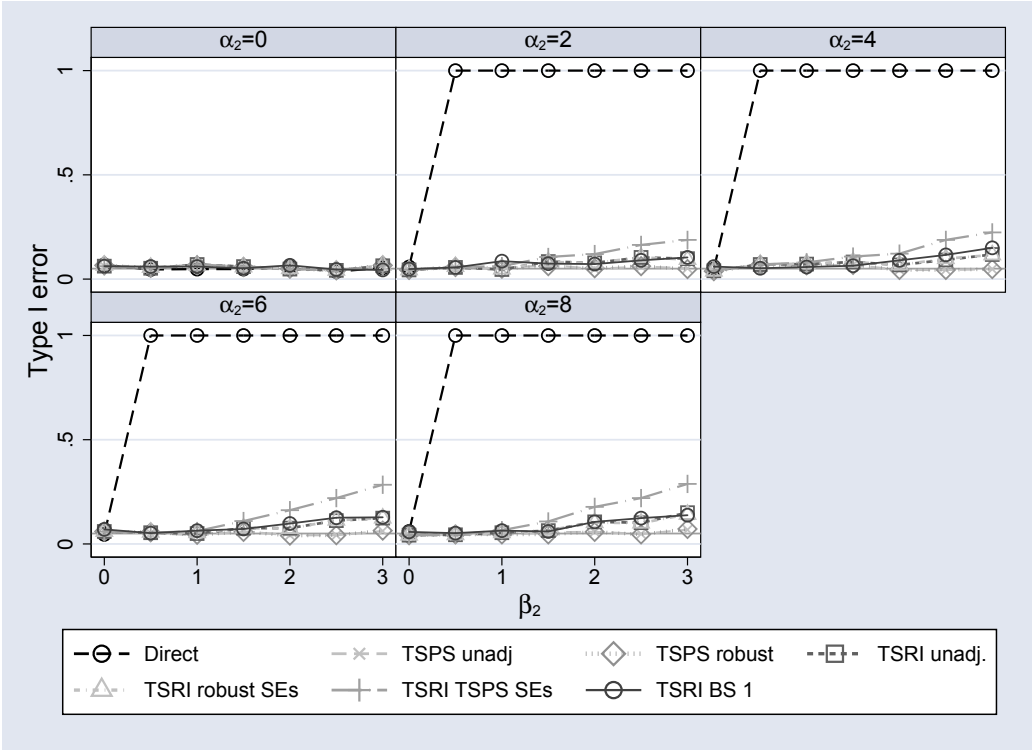


(a)

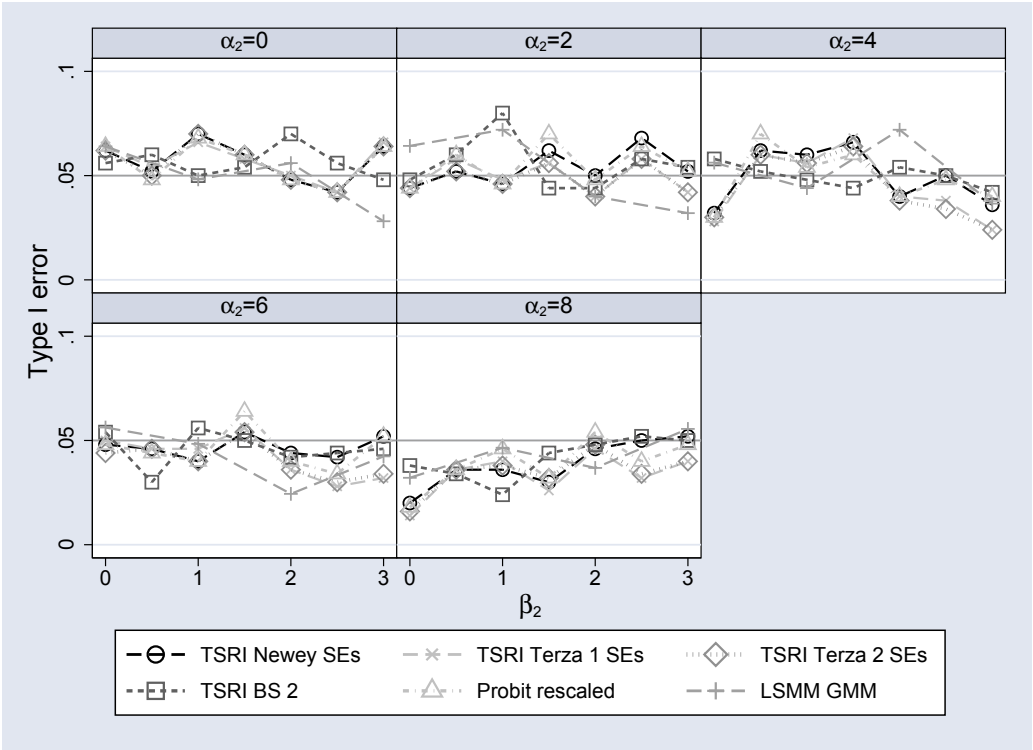


(b)

Web Figure 9: Coverage of the logistic estimators for $N = 5\,000$ with respect to the marginal parameter estimated by the TSRI estimator (BS: bootstrap; LSMM: logistic structural mean model; GMM: generalized method of moments; SE: standard error; TSPS: two-stage predictor substitution; TSRI: two-stage residual inclusion).

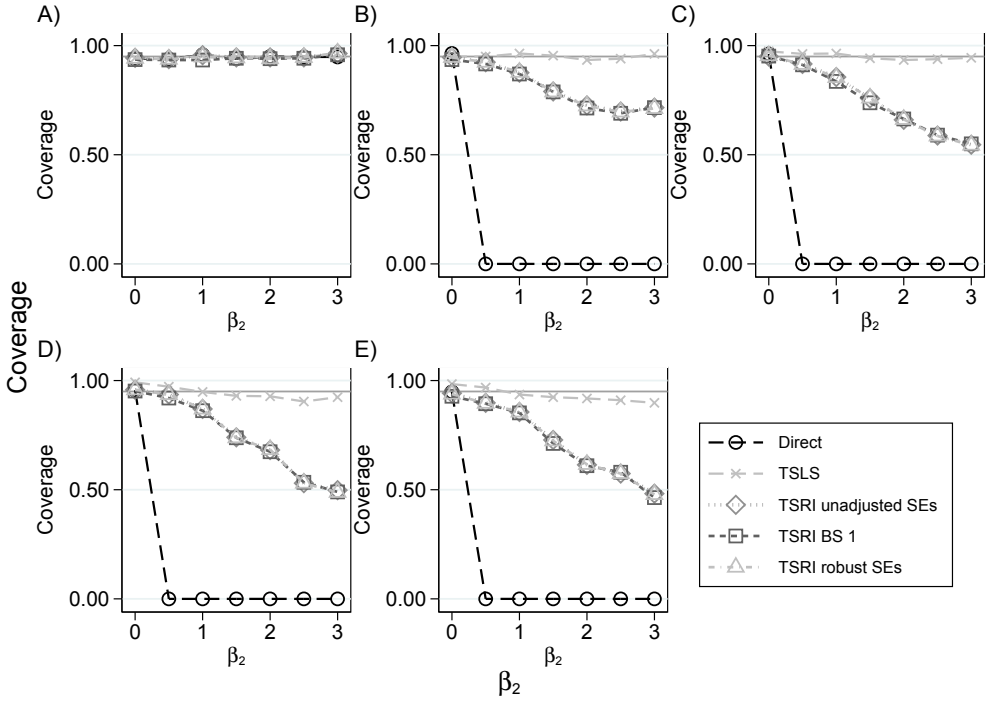


(a)

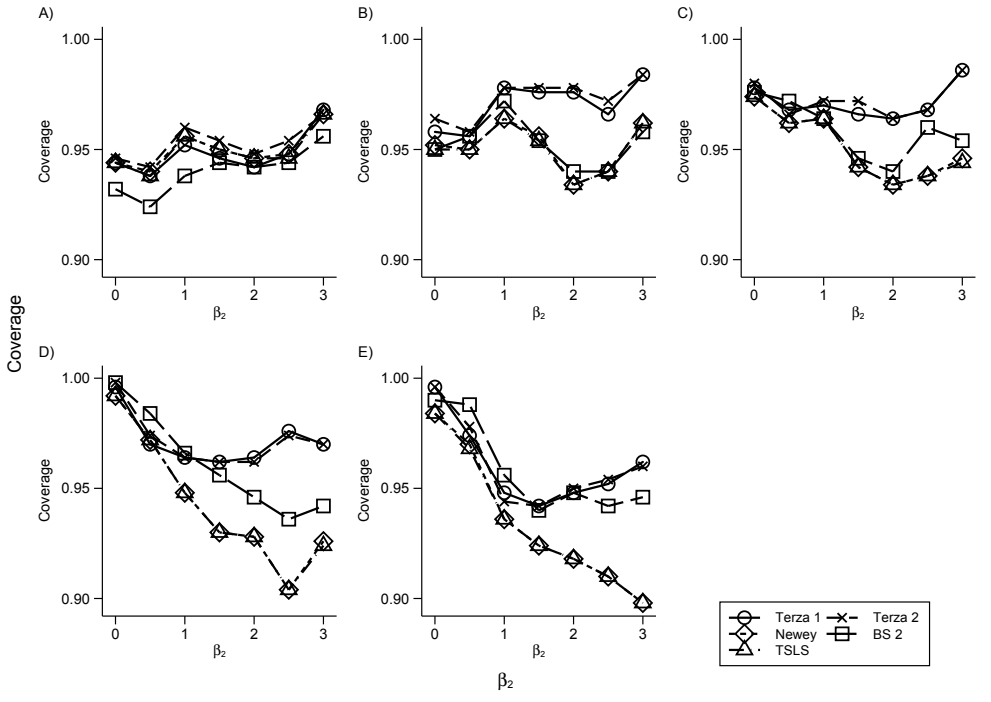


(b)

Web Figure 10: Type I error of the logistic estimators for $N = 5\,000$ (BS: bootstrap; LSMM: logistic structural mean model; GMM: generalized method of moments; SE: standard error; TSPS: two-stage predictor substitution; TSRI: two-stage residual inclusion).

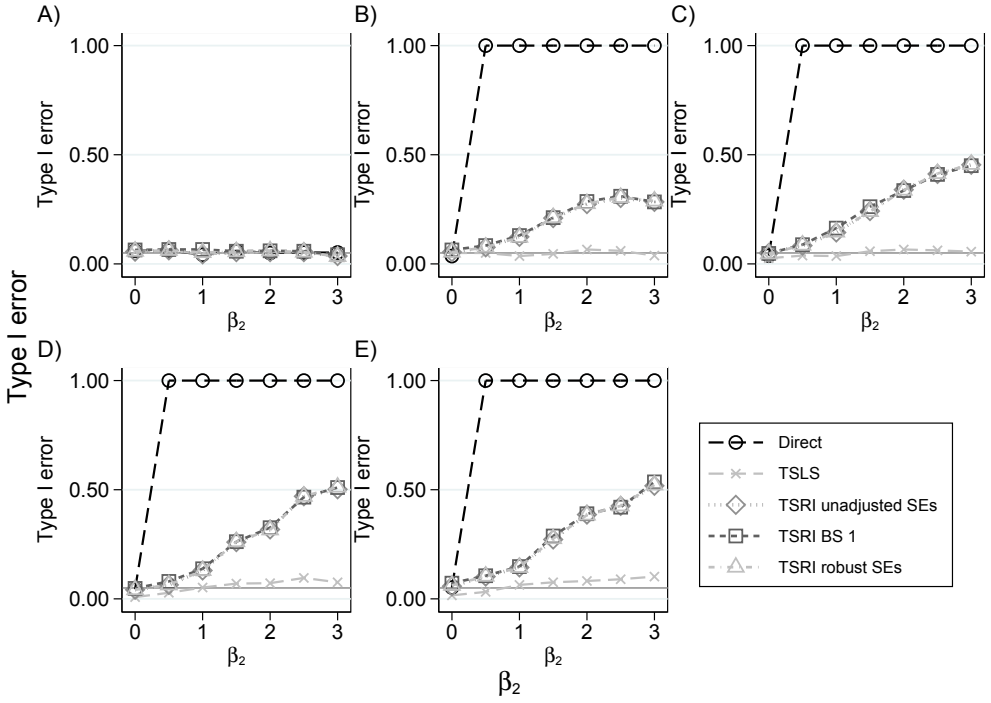


(a)

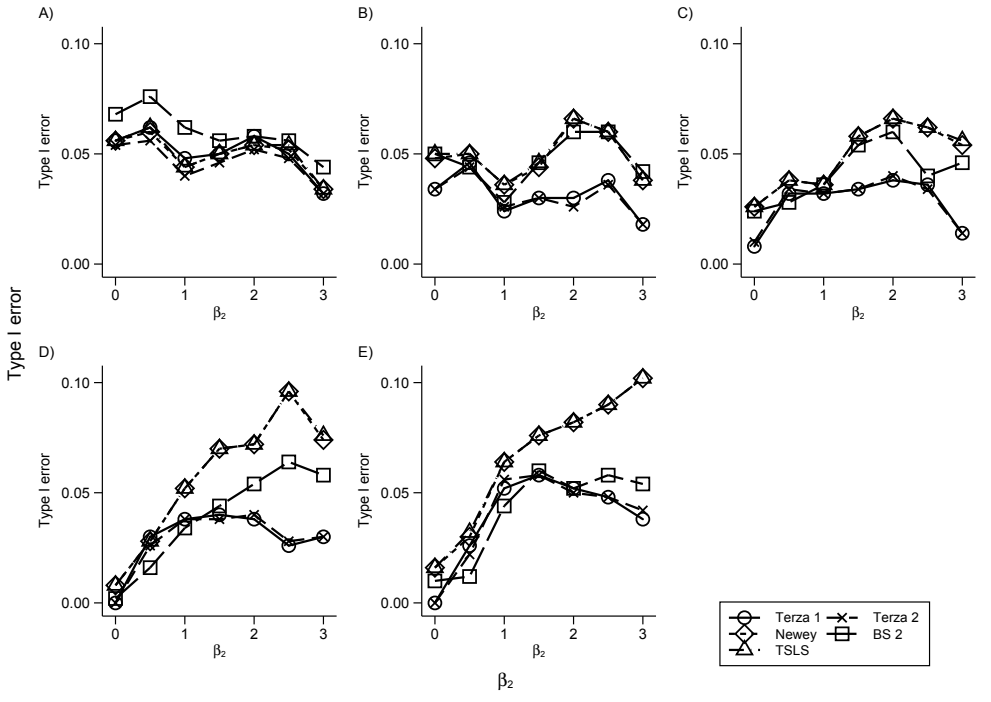


(b)

Web Figure 11: Coverage of the linear estimators for $N = 1\,000$ (BS: bootstrap; SE: standard error; TSLS: two-stage least squares; TSRI: two-stage residual inclusion). The panels correspond to α_2 being set to the following values A:0, B:2, C:4, D:6, and E:8.

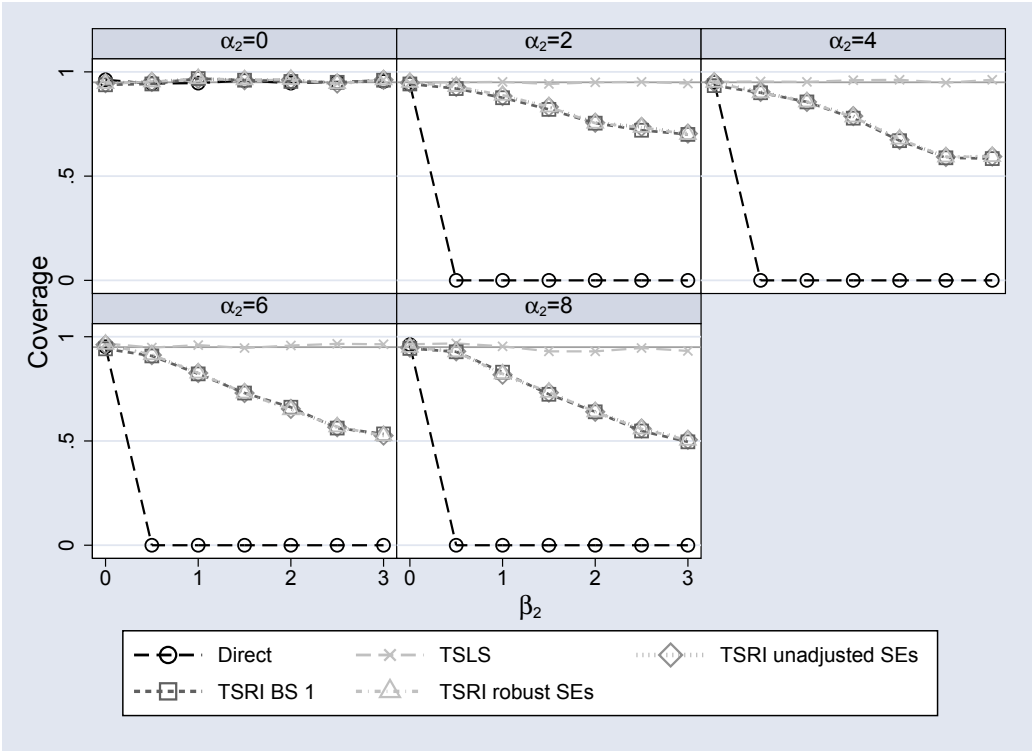


(a)

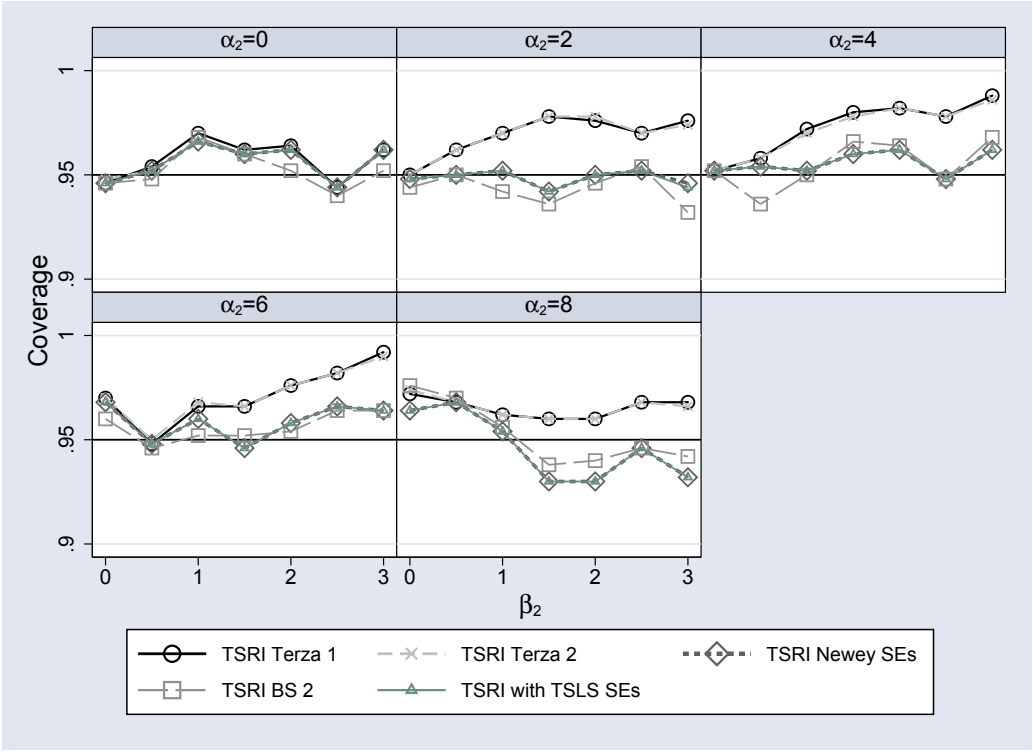


(b)

Web Figure 12: Type I error of the linear estimators for $N = 1\,000$ (BS: bootstrap; SE: standard error; TSLS: two-stage least squares; TSRI: two-stage residual inclusion). The panels correspond to α_2 being set to the following values A:0, B:2, C:4, D:6, and E:8.

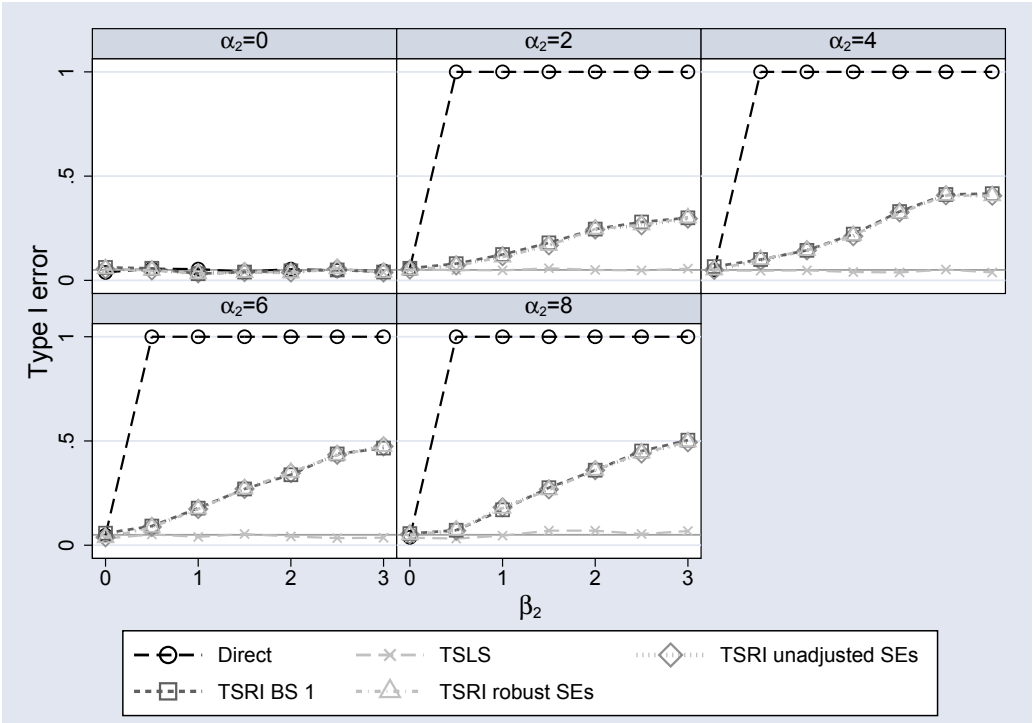


(a)

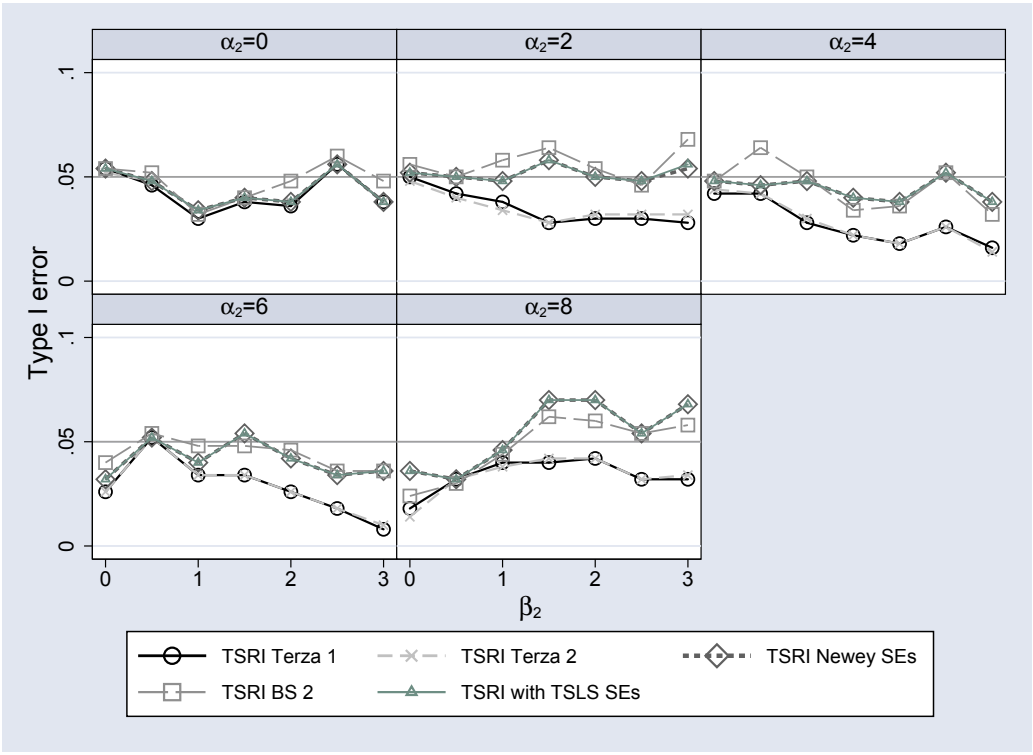


(b)

Web Figure 13: Coverage of the linear estimators for $N = 5\,000$ (BS: bootstrap; SE: standard error; TSLS: two-stage least squares; TSRI: two-stage residual inclusion).



(a)



(b)

Web Figure 14: Type I error of the linear estimators for $N = 5\,000$ (BS: bootstrap; SE: standard error; TSLS: two-stage least squares; TSRI: two-stage residual inclusion).