Corrections to Probit and logistic control function estimator standard errors for marginal parameters

Dr Tom Palmer

June 2015

Mathematics & Statistics | Lancaster University

Outline

- Control function estimators (CFEs)
- Previous work (Palmer et al., 2008)
- TSLS standard errors
- Probit CFE and Newey's standard error correction
- Application to logistic CFE
- Application to linear CFE
- Summary

Control function estimators (CFEs)

- Similar to TSLS
- Additionally include first stage residuals in second stage model
- With X phenotype, Y outcome, Z instrument;

$$X = \alpha_0 + \alpha_1 Z + u$$
$$Y = \beta_0 + \beta_1 X + \beta_2 \widehat{u} \quad \text{or some GLM}$$

- Linear 2$^{\text{nd}}$ stage: $\widehat{\beta}_1 = \widehat{\beta}_{\text{TSLS}}$
- Linear 2$^{\text{nd}}$ stage: test of $\widehat{\beta}_2 = 0$ is an endogeneity test

Previous work (Palmer et al., 2008) I

Published by Oxford University Press on behalf of the International Epidemiological Association
© The Author 2008; all rights reserved. Advance Access publication 7 May 2008

International Journal of Epidemiology 2008;37:1161–1168
doi:10.1093/ije/dyn080

Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses

Tom M Palmer,[1]* John R Thompson,[1] Martin D Tobin,[2] Nuala A Sheehan[2] and Paul R Burton[2]

Accepted 3 April 2008

Background Mendelian randomization uses a carefully selected gene as an instrumental-variable (IV) to test or estimate an association between a phenotype and a disease. Classical IV analysis assumes linear relationships between the variables, but disease status is often binary and modelled by a logistic regression. When the linearity assumption between the variables does not hold the IV estimates will be biased. The extent of this bias in the phenotype-disease log odds ratio of a Mendelian randomization study is investigated.

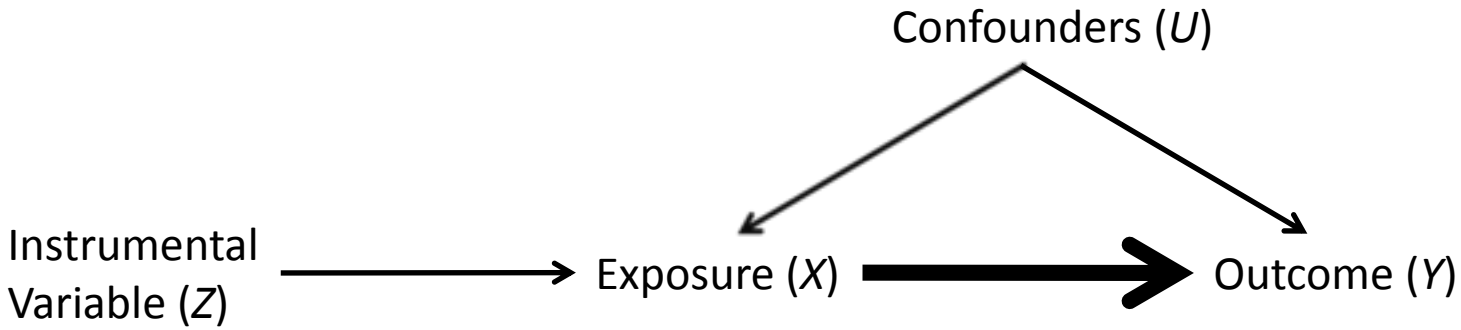
Methods Three estimators termed direct, standard IV and adjusted IV, of the phenotype-disease log odds ratio are compared through a simulation study which incorporates unmeasured confounding. The simulations are verified using formulae relating marginal and conditional estimates given in the Appendix.

Results The simulations show that the direct estimator is biased by unmeasured confounding factors and the standard IV estimator is attenuated towards the null. Under most circumstances the adjusted IV estimator has the smallest bias, although it has inflated type I error when the unmeasured confounders have a large effect.

Conclusions In a Mendelian randomization study with a binary disease outcome the bias associated with estimating the phenotype-disease log odds ratio may be of practical importance and so estimates should be subject to a sensitivity analysis against different amounts of hypothesized confounding.

Keywords Instrumental-variable analysis, Mendelian randomization, bias, unobserved confounding

Previous work (Palmer et al., 2008) II

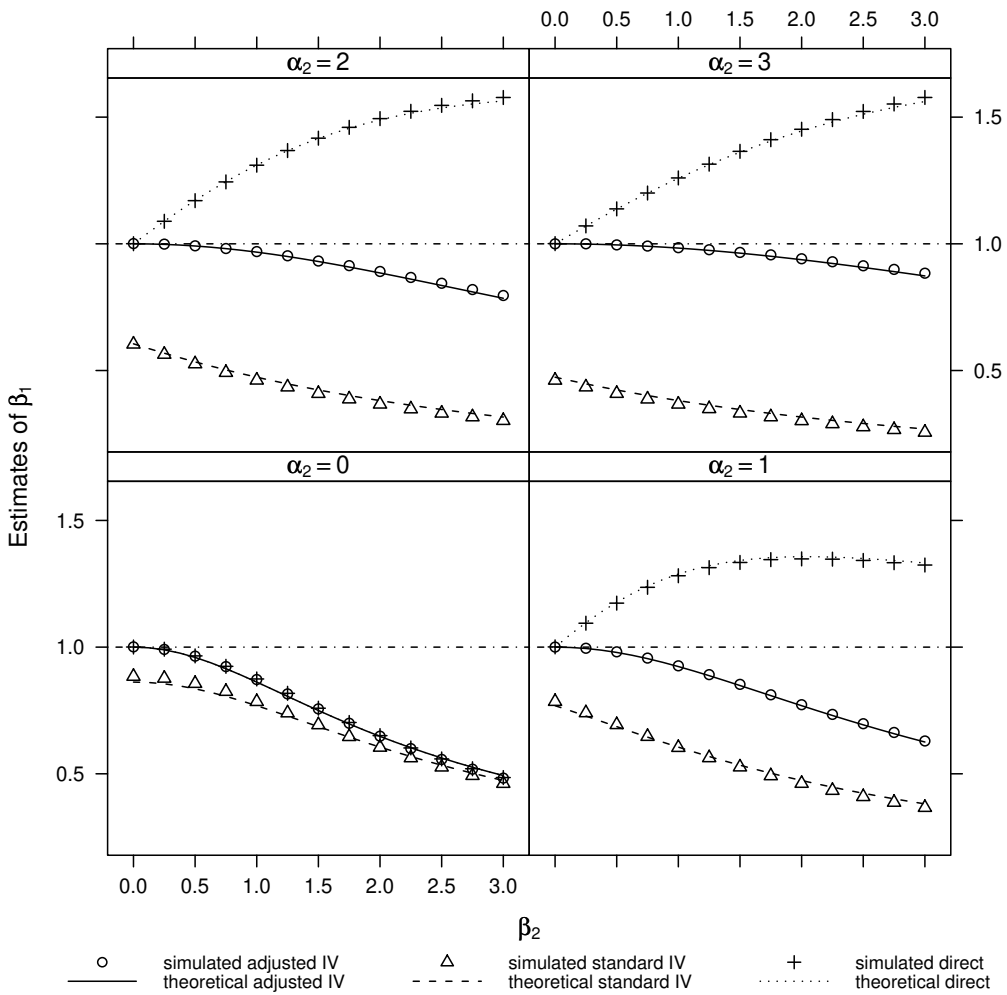


- Binary outcome
- 3 estimators
  1. Direct: `logistic outcome phenotype`
  2. Standard IV: `logistic outcome predicted`
  3. Adjusted IV: `logistic outcome predicted residuals`
- Assessed properties with simulation study

Previous work (Palmer et al., 2008) III: simulation setup

- gen double g = rbinomial(2, 0.3) // hwe, risk allele 30%
- gen double u = rnormal()
- gen double x = rnormal(0 + z + α_2u, 1)
- gen double p = invlogit(log(0.05/0.95) + β_1x + $\beta_2 u$)
- gen byte y = rbinomial(1, p)
- Vary: $\alpha_2 = (0, 3)$, $\beta_2 = (0, 3)$

Previous work (Palmer et al., 2008) III: estimates



- Conditional parameter: set at $\beta_1 = 1$
- Marginal parameters: dashed lines

Previous work (Palmer et al., 2008) IV: estimates

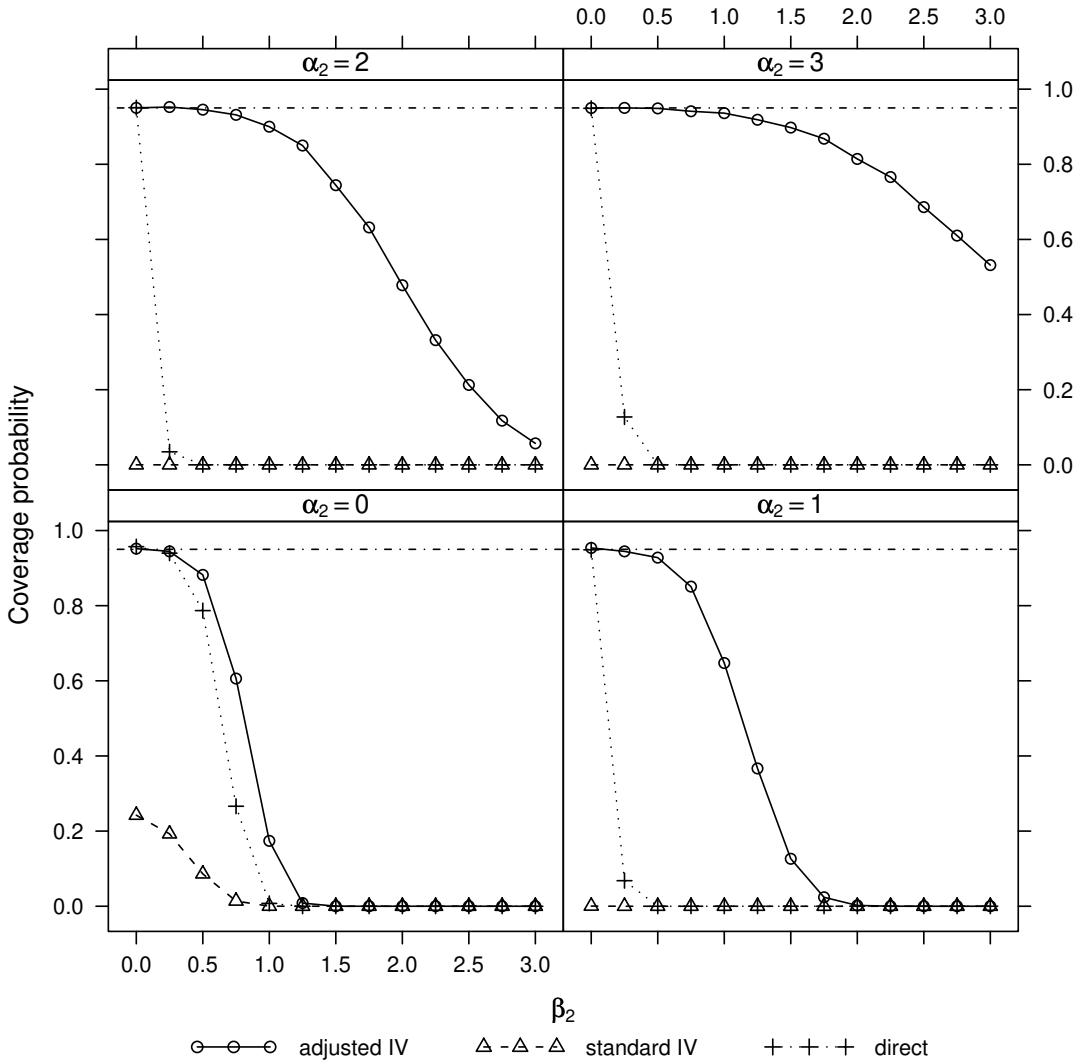
- Adjusted IV (logistic control function estimator) estimates marginal parameter

$$\text{Conditional: } \beta_1$$

$$\text{Marginal: } \beta_1 \times \frac{1}{\sqrt{1 + c^2 V}}$$

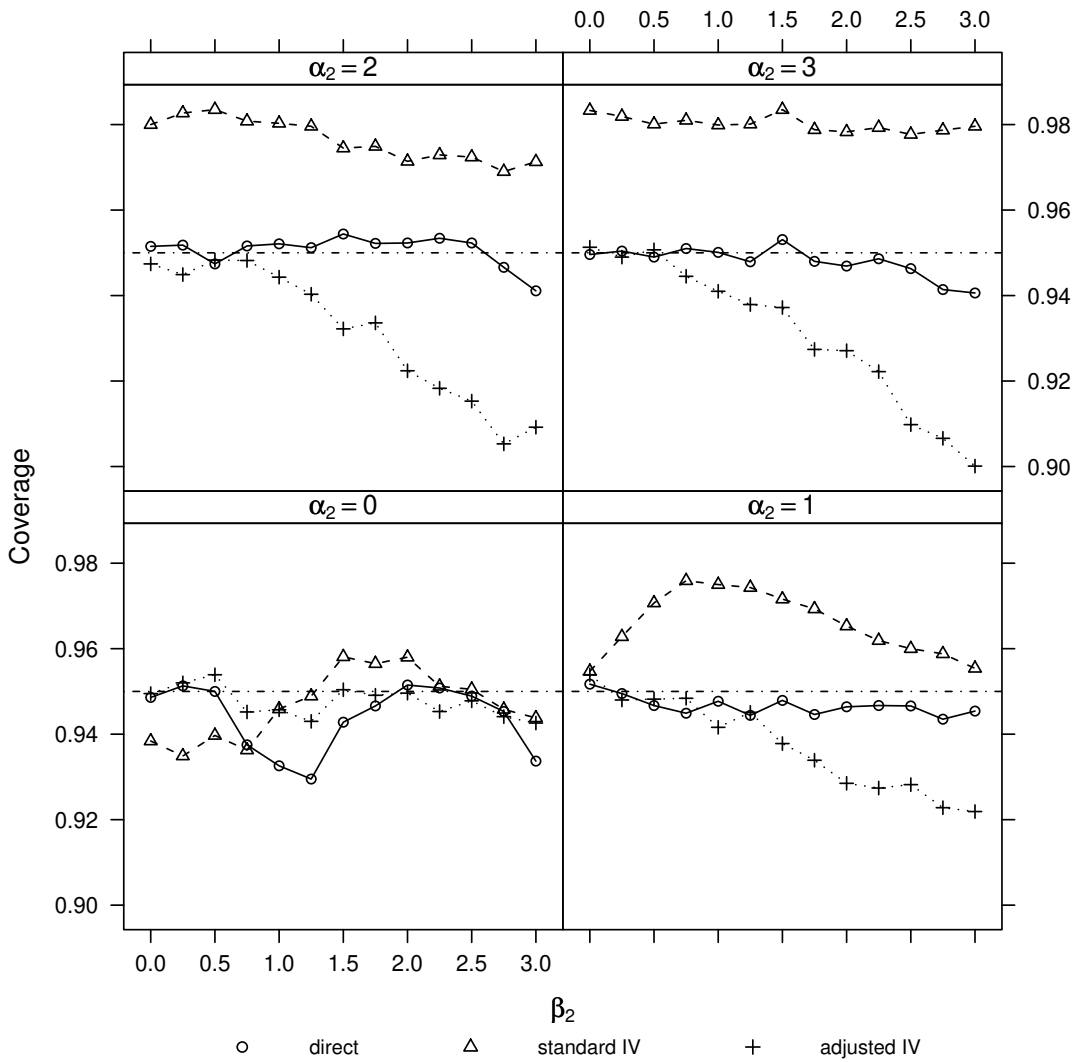
$c = \frac{16\sqrt{3}}{15\pi}$
V: variance of the covariates over which marginal estimates are averaged.

Previous work (Palmer et al., 2008) IV: coverage



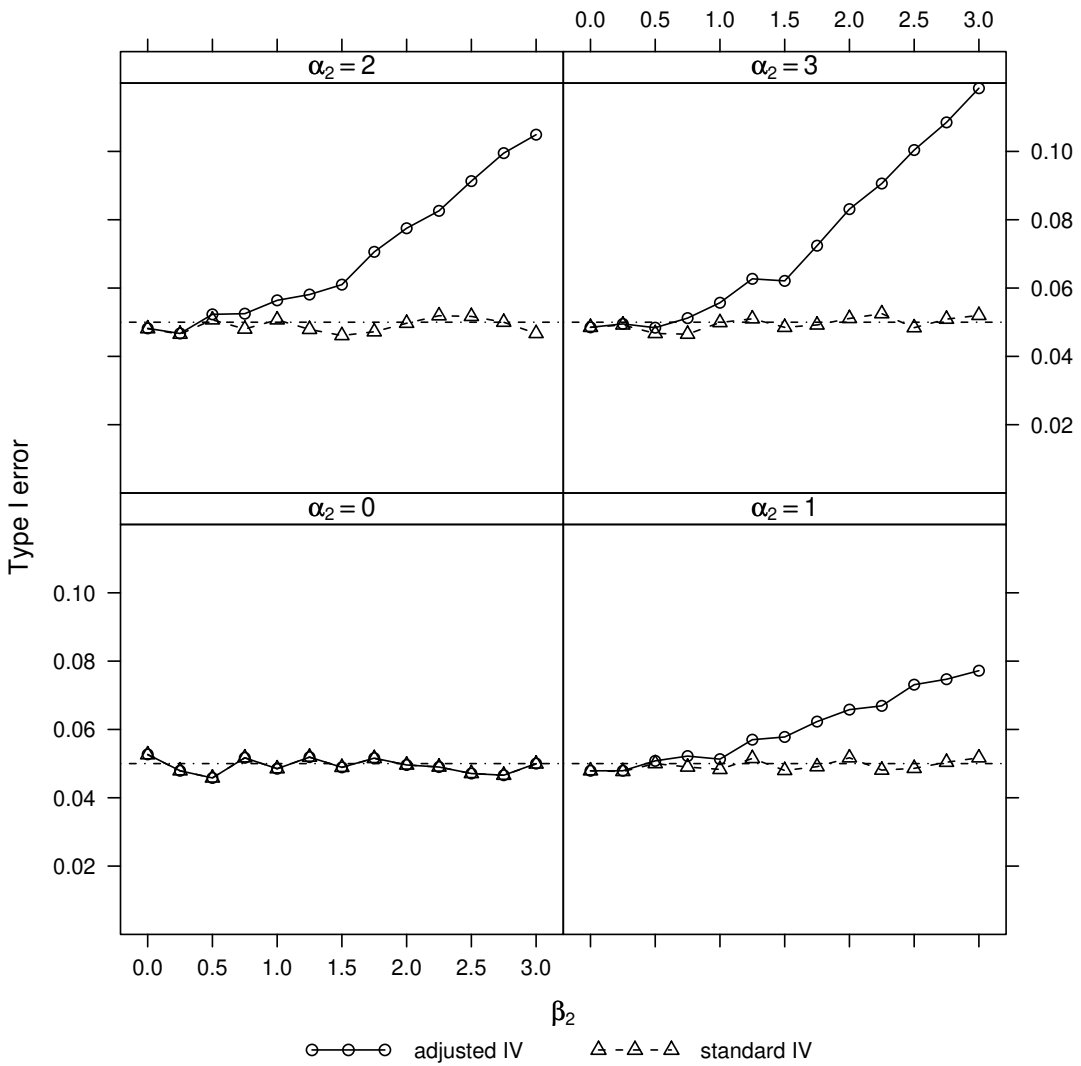
Coverage wrt conditional parameter – very low

Previous work (Palmer et al., 2008) IV: coverage



Coverage wrt marginal parameter (too low; adjusted IV $< 95\%$)

Previous work (Palmer et al., 2008) IV: type I error



How to get correct coverage and type I error?

TSLS standard errors I

- Running TSLS manually we have:
  1. $X = \alpha_0 + \alpha_1 Z + u$
  2. $Y = \beta_0 + \beta_1 \widehat{X} + \epsilon$
- The second stage standard errors are corrected to account for use of $\widehat{X}$ in second stage instead of X (which causal model is formulated in terms of)
- This affects estimate of variance of the residuals
- Define projection matrix, $P_Z = Z(Z'Z)^{-1}Z'$, so $\widehat{X} = P_Z X$, then

$$\widehat{\beta}_{\text{2SLS}} = (X' P_Z X)^{-1} X' P_Z Y$$

TSLS standard errors II

- Uncorrected manual 2nd stage SEs are:

$$\text{var}(\widehat{\beta}_{\text{2nd}}) = s^2(\widehat{X}'\widehat{X})^{-1}$$
$$\text{with } s^2 = (Y - \widehat{X}\widehat{\beta}_{\text{IV}})^2/(n-k).$$

- Corrected 2nd stage SEs are (assuming homoskedasticity):

$$\text{var}(\beta_{\text{2SLS}}) = \sigma^2(X'P_ZX)^{-1}$$
$$\text{with } \sigma^2 = (Y - X\widehat{\beta}_{\text{IV}})^2/n.$$

Newey (1987) standard errors for Probit CFE I

- Binary Y

$$X = Z\Pi + v_i$$
$$Y = X\delta + u_i$$
$$(u_i, v_i) \sim \text{MVN}(0, \Sigma)$$

- Estimation: maximum likelihood and twostep.

Two-step estimation:

1. Regress X on Z and estimate the residuals ($\widehat{v_i}$).
2. Probit regression of Y on X and $\widehat{v_i}$

- Stata implementations:
  - `probitiv` (Gelbach, 1997) – unadjusted SEs
  - `ivprob` (Harkness) – Newey SEs
  - `ivprobit` (Stata) – Newey SEs

Newey (1987) standard errors for Probit CFE II

- Newey, Efficient estimation of limited dependent variable models, Journal of Econometrics, 1987
- 3 key equations:

$$\widehat{\delta} = (\widehat{D}\widehat{\Omega}^{-1}\widehat{D})^{-1}\widehat{D}'\widehat{\Omega}^{-1}\tilde{\alpha} \text{ (Newey Eq. 5.6)}$$

$$\text{var}(\widehat{\delta}) = (\widehat{D}\widehat{\Omega}^{-1}\widehat{D})^{-1}$$

$$\widehat{\Omega} = J_{\alpha\alpha}^{-1} + (\lambda - \beta)'\Sigma_{22}(\lambda - \beta)Q^{-1} \text{ (Newey Eq. 5.4)}$$

- How to obtain the elements of $\widehat{\Omega}$?

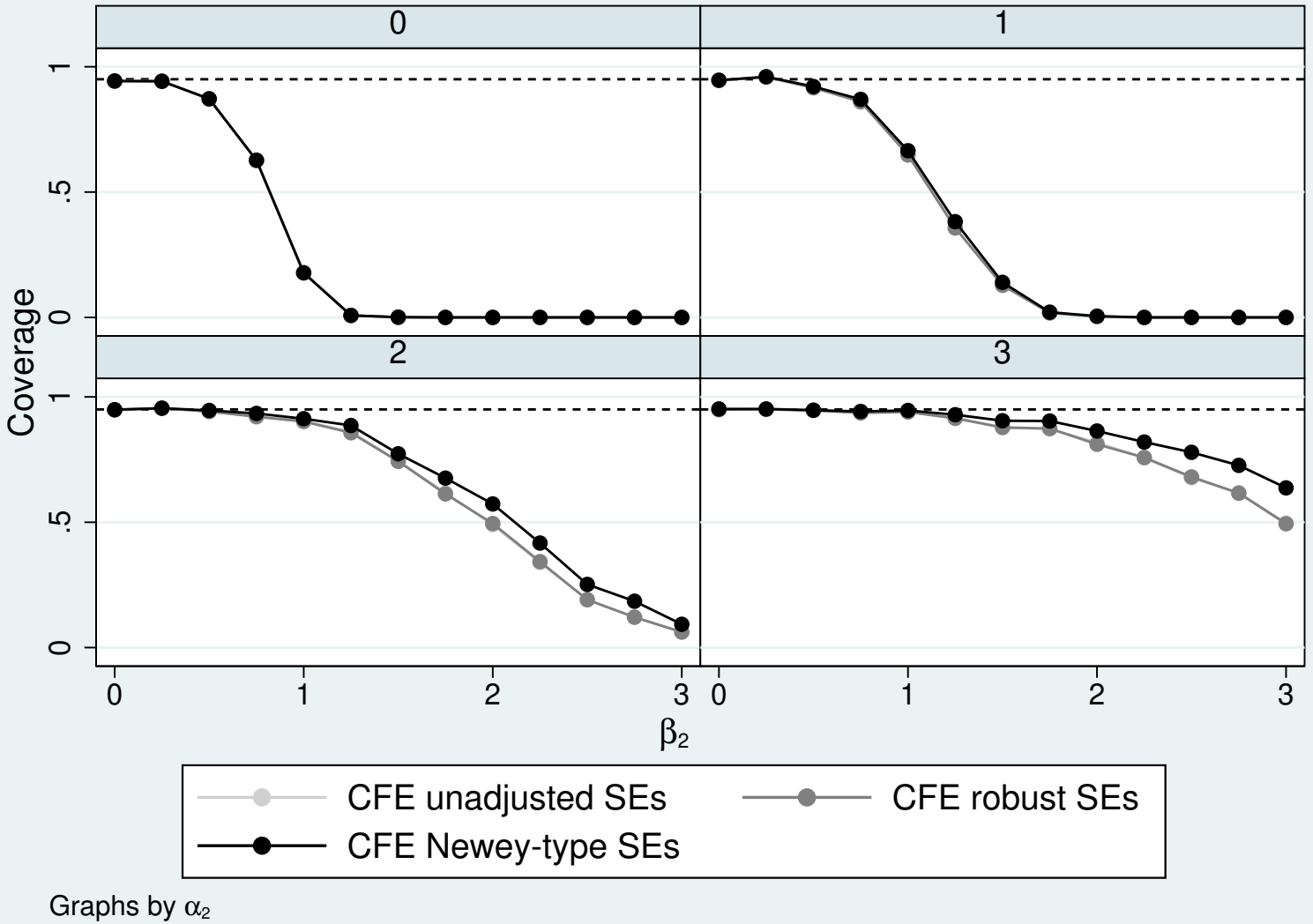
Newey (1987) standard errors for Probit CFE III

Stata [R] Reference Manual page 921 (`ivprobit`: Methods and formulas):

1. Regress X on Z to compile $\widehat{D}$ (I(k) with D[,1] coefs from this)
2. Probit regression of Y on Z and $\widehat{v}_i$
   - $\tilde{a}$:=coefficients on Z and constant
   - $J_{\alpha\alpha}^{-1}$ is var-covar matrix of these coefficients
   - $\widehat{\lambda}$ = coefficient on $\widehat{v}_i$.
3. Probit CFE: Probit regression of Y on X and $\widehat{v}_i$
   - coefficient on X is $\widehat{\beta}$
4. Generate $X(\widehat{\lambda} - \widehat{\beta})$
   - regress $X(\widehat{\lambda} - \widehat{\beta})$ on Z
   - covariance matrix is estimate of term after $+$ in $\widehat{\Omega}$
   - add to $J_{\alpha\alpha}^{-1}$ giving $\widehat{\Omega}$

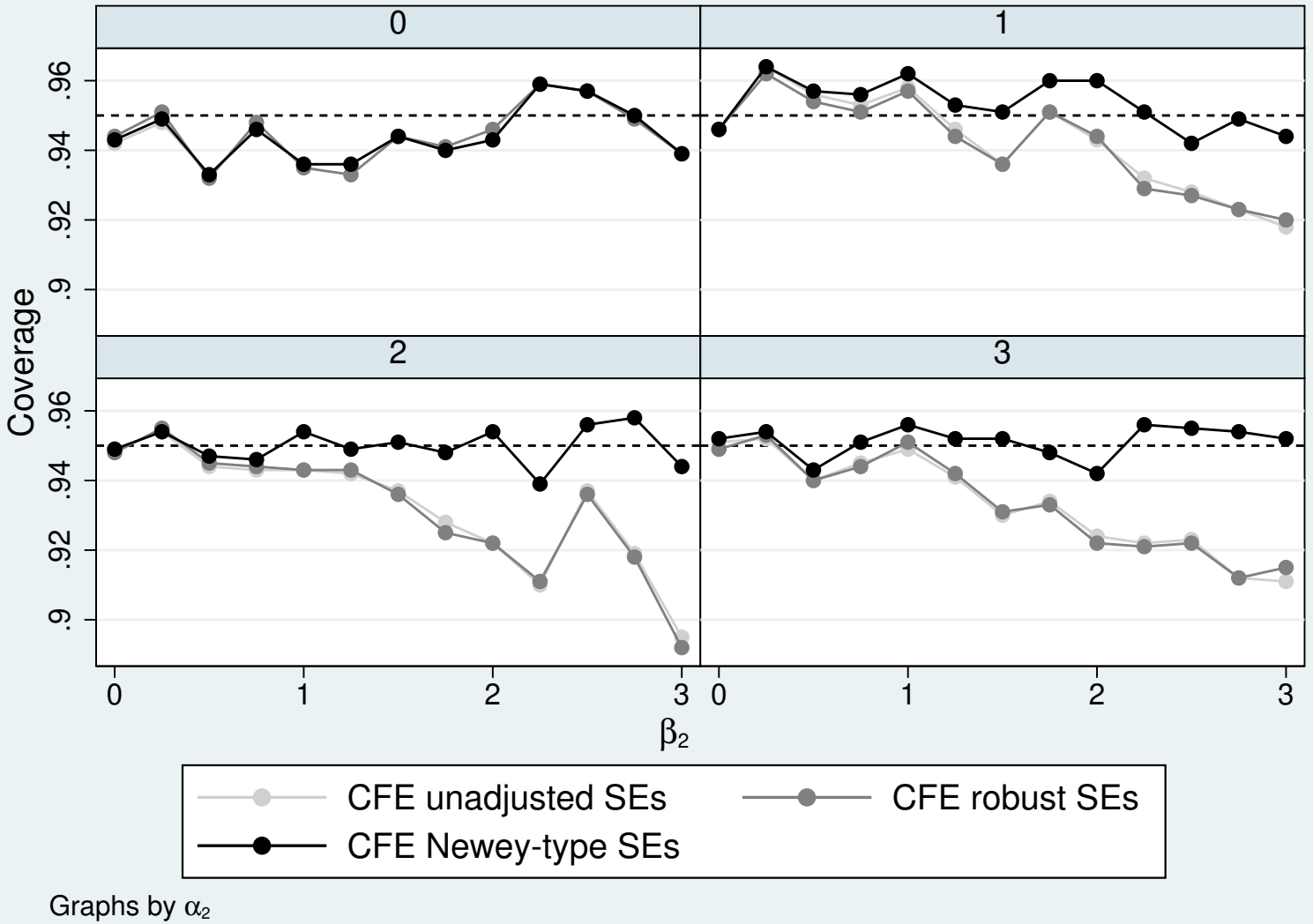
- Approximation for logistic CFE: replace the 2 Probit regressions in (2) and (3) with logistic regressions (and do the same for other GLMs at second stage).

Logistic CFE: coverage simulation results



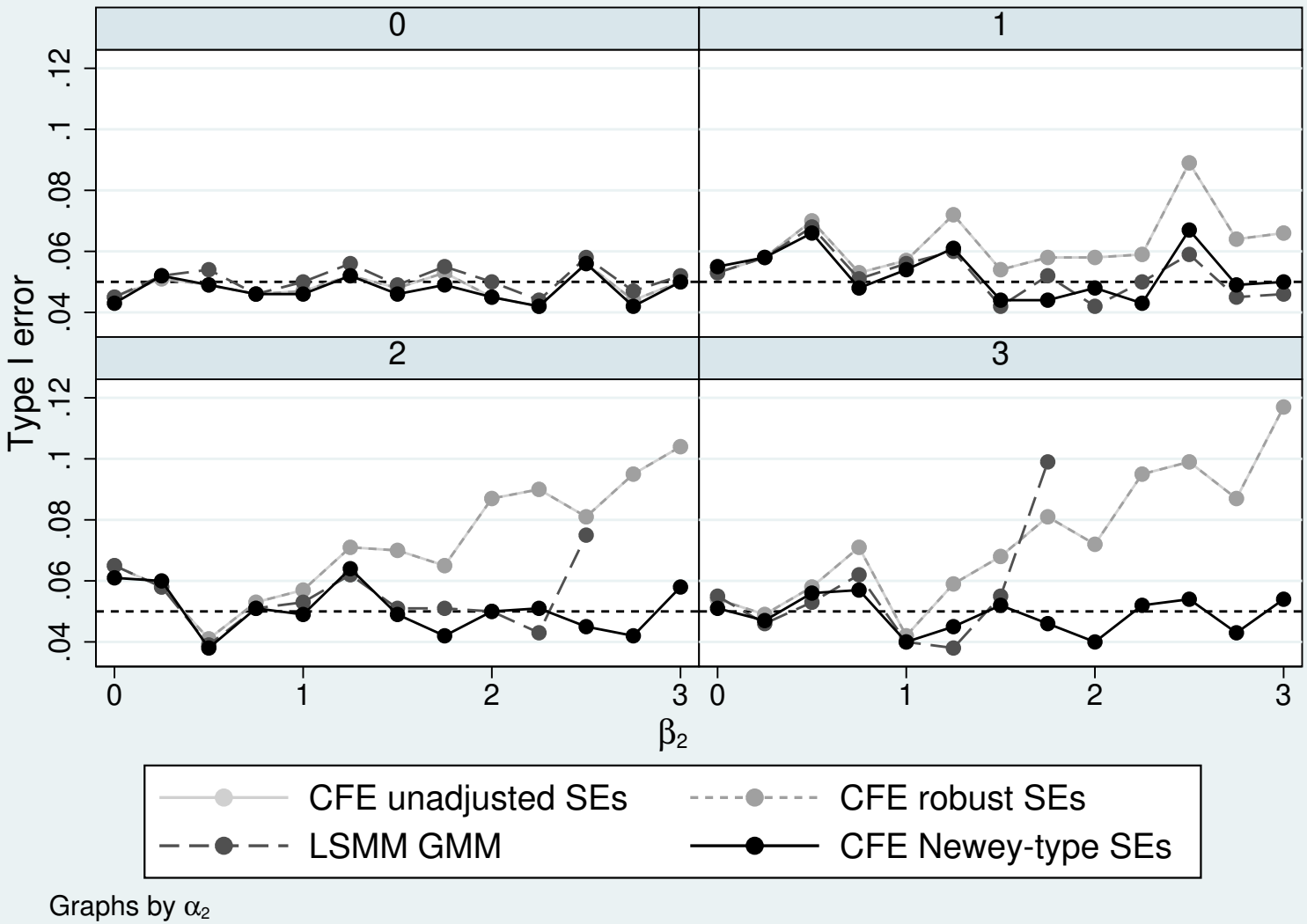
Coverage wrt conditional parameter

Logistic CFE: coverage simulation results

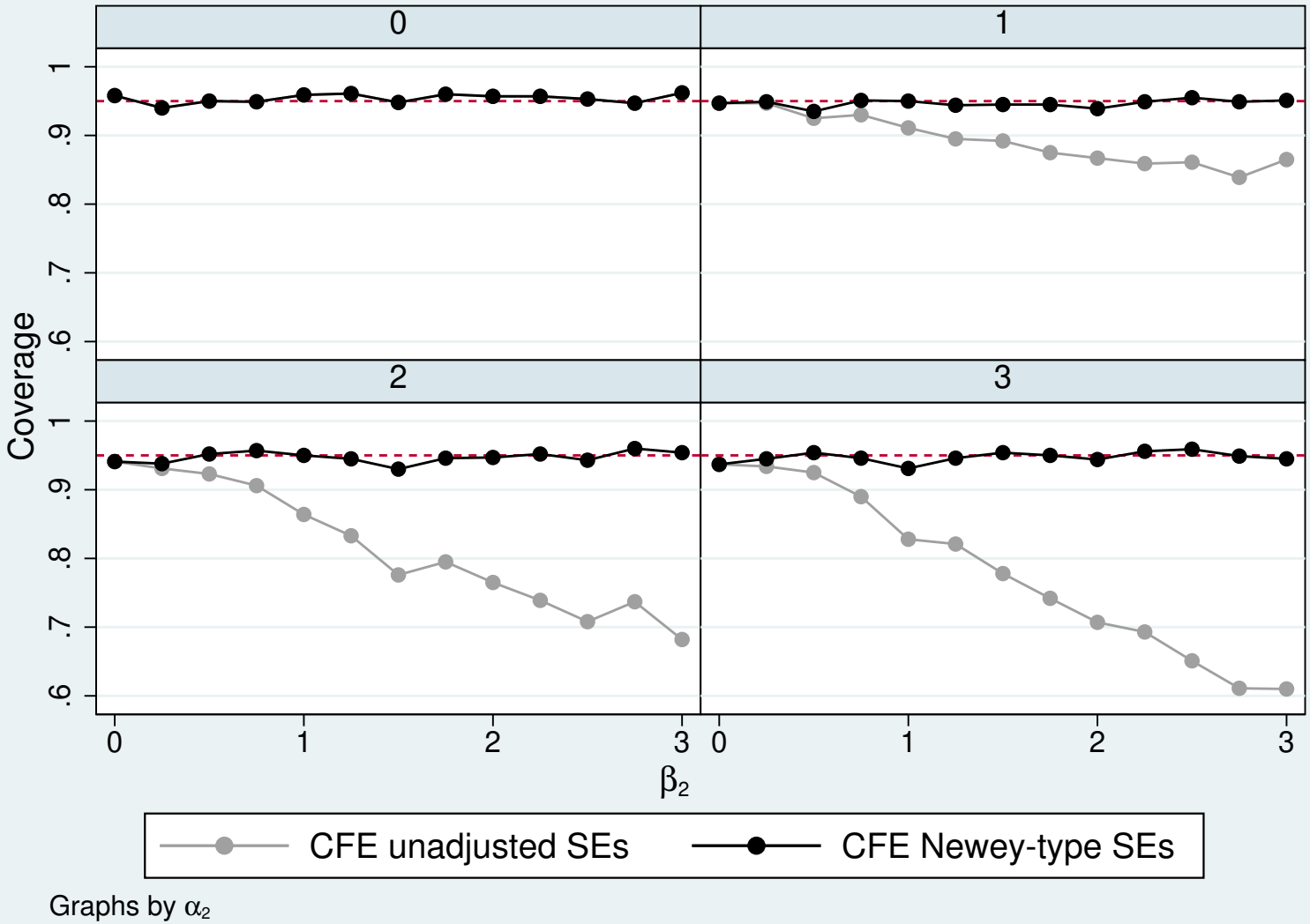


Coverage wrt marginal parameter

Logistic CFE: type I error simulation results

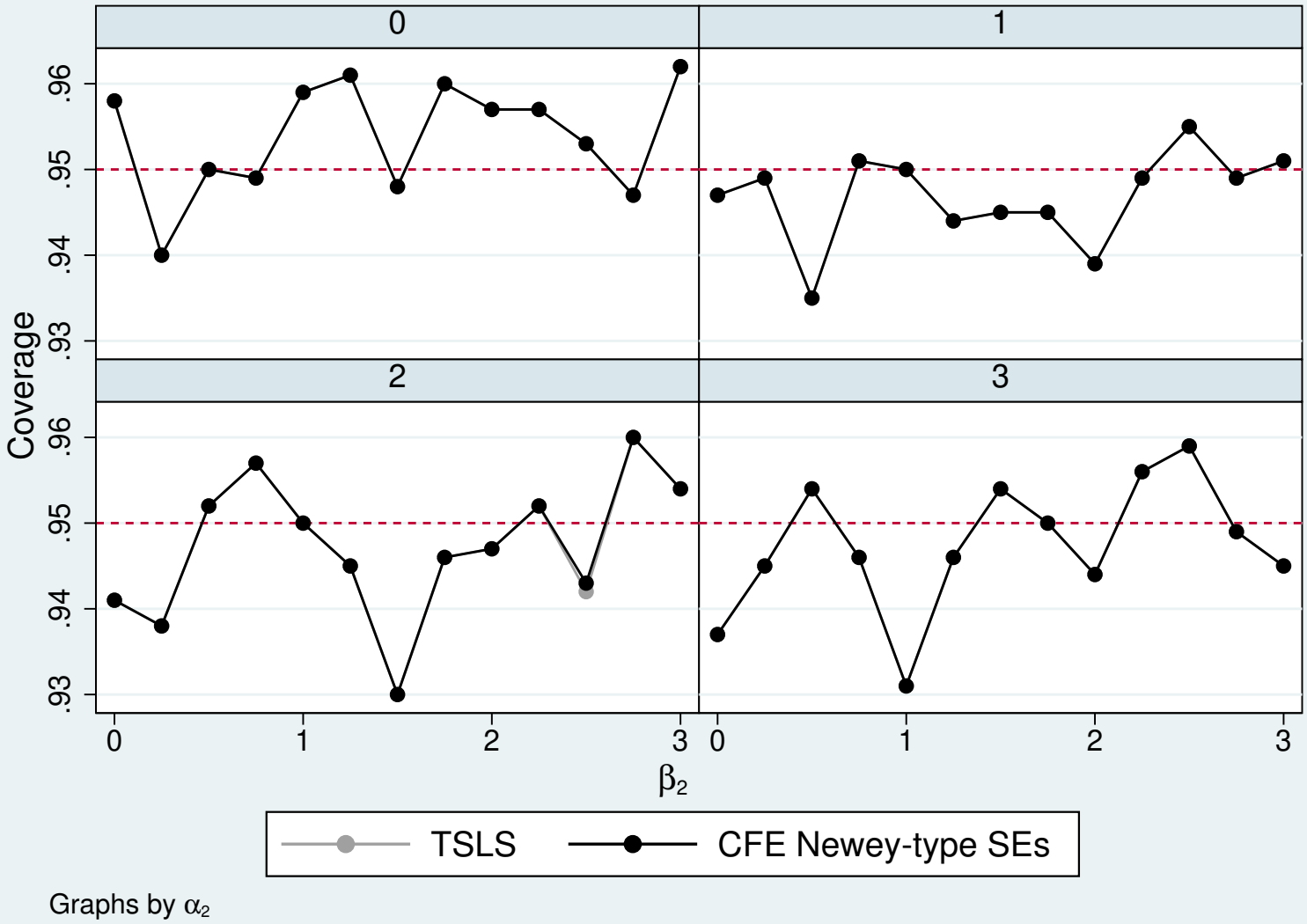


Application to linear CFE



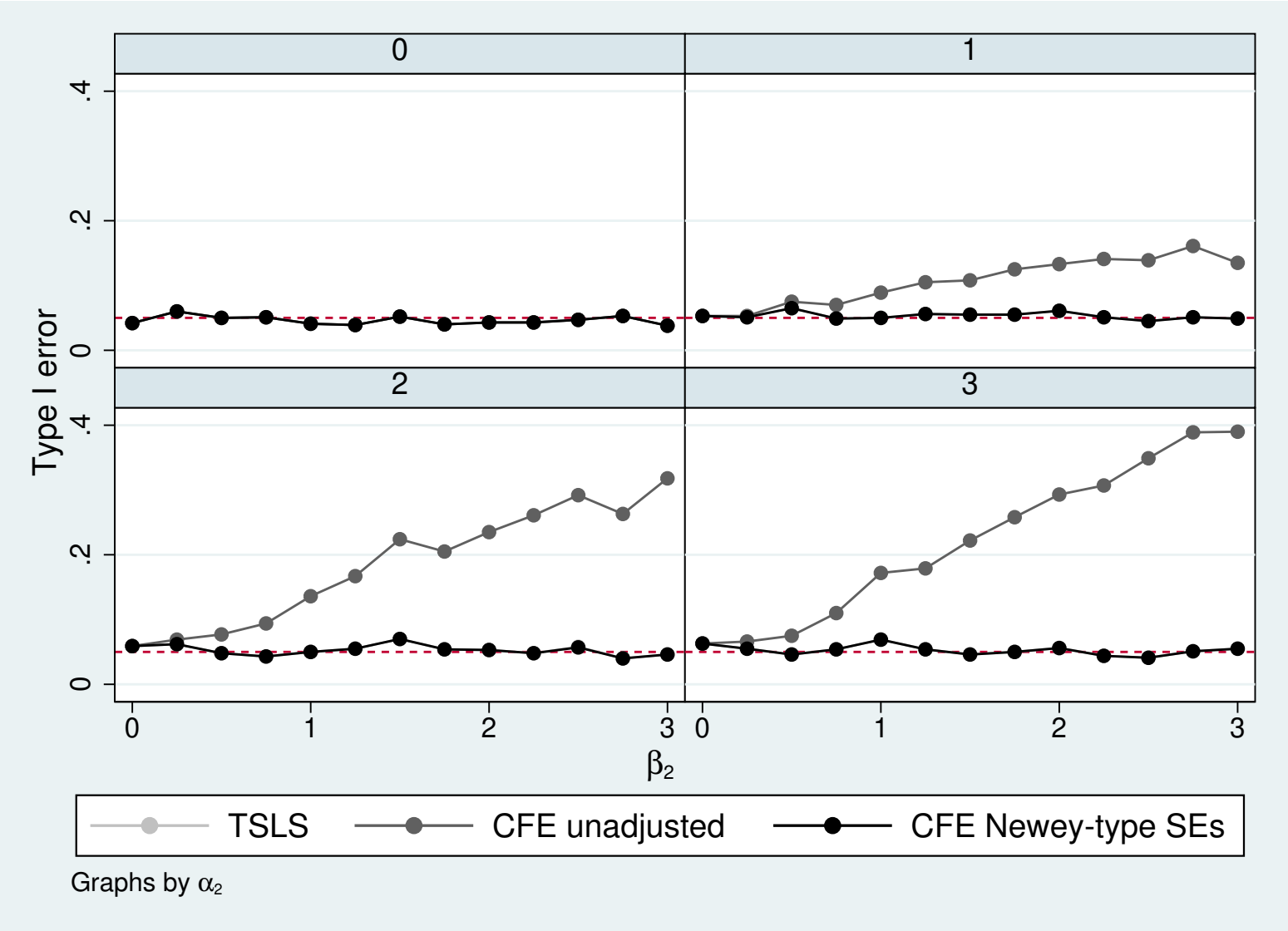
Coverage compared to uncorrected linear CFE SEs

Application to linear CFE



Coverage compared to corrected TSLS SEs

Application to linear CFE



Type I error compared to TSLS and uncorrected linear CFE SEs

Real data example I

- 17057 participants from 6 cohorts of European ancestry
- exposure: body-mass index
- instrument: externally weighted allele score
- continuous outcome: systolic blood pressure
- binary outcome: diabetes status

Real data example II

Estimator	Ancilliary statistics	Estimate (95% CI)
Direct		0.76 (0.70, 0.82)
TSLS (F=119, R^2=0.007)	SE=0.374	0.36 (-0.37, 1.10)
Linear CFE (unadjusted)	SE=0.372	0.36 (-0.37, 1.09)
Linear CFE (Newey-type)	SE=0.374	0.36 (-0.37, 1.10)

Table: Estimates of the causal effect of a one unit increase in body mass index on systolic blood pressure (mmHg) (All N=17 057).

Newey-type SE 0.5% larger.

Real data example III

Estimator	Ancilliary statistics	Odds ratio (95% CI)
Direct		1.14 (1.13, 1.15)
Standard IV (F=119 R^2=0.007)	SE=0.056, $z = 4.96$	1.32 (1.19, 1.48)
Logistic CFE (unadjusted)	SE=0.058, $z = 4.79$	1.32 (1.18, 1.48)
Logistic CFE (Newey-type)	SE=0.059, $z = 4.71$	1.32 (1.17, 1.48)
Logistic SMM	SE=.	1.41 (.,.)
Probit CFE	$z = 4.74$	0.15 (0.089, 0.214)

Table: Estimates of the causal odds ratios for diabetes for a one unit increase in body mass index (All N=17 057, SEs for logistic SMM did not converge).

Newey-type SE 2% larger.

Summary

- CFEs estimate marginal parameters
- Manual TSLS SEs require correction
- SEs of the manual two-step Probit CFE can be corrected (Newey, 1987)
- Logistic and linear CFE SEs can be corrected using this method – correct coverage, type I error (possibly other GLMs at second stage too)
- Probit and logistic CFE z-statistics slightly different (could correct logistic CFE SE using the Probit z-statistic but then matrix formula for estimate would be slightly incorrect)
- Doesn't overcome weak instrument problem (Davies et al., 2015).