A Stein-Like 2SLS Estimator

Bruce E. Hansen[*]
University of Wisconsin[†]

December 2014
Revised: June 2015

Abstract

Maasoumi (1978) proposed a Stein-like estimator for simultanous equations and showed that his Stein shrinkage estimator has bounded finite sample risk, unlike the 3SLS estimator. We revisit his proposal by investigating Stein-like shrinkage in the context of 2SLS estimation of a structural parameter. Our estimator follows Maasoumi (1978) in taking a weighted average of the 2SLS and OLS estimators, with the weight depending inversely on the Hausman (1978) statistic for exogeneity. Using a local-to-exogenous asymptotic theory, we derive the asymptotic distribution of the Stein estimator, and calculate its asymptotic risk. We find that if the number of endogenous variables exceeds two, then the shrinkage estimator has strictly smaller risk than the 2SLS estimator, extending the classic result of James and Stein (1961). In a simple simulation experiment, we show that the shrinkage estimator has substantially reduced finite sample median squared error relative to the standard 2SLS estimator.

[*] Research supported by the National Science Foundation. This paper is written for the *Econometric Reviews* special issue in honor of Essie Maasoumi. I thank the Co-Editors and two referees for helpful comments.

[†] Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706. behansen@wisc.edu

1 Introduction

In a particularly insightful paper, Maasoumi (1978) derives a Stein-like estimator for the reduced form coefficients of simultaneous equations. He shows that while the standard three-stage least squares (3SLS) and two-stage-least squares (2SLS) estimators possess no finite moments, his modified Stein-like estimator has thinner tails, finite moments, and thus bounded risk. His is the first paper to apply the idea of Stein shrinkage to simultaneous equations estimation, and the first to be motivated by the goal of producing an estimator with finite moments.

In the more recent literature, the focus has shifted to the structural coefficients in single equation instrumental variables models. While an enormous literature has been written on improvements on 2SLS estimates, to my knowledge there has been no meaningful follow-up to Maasoumi's insight that Stein-like shrinkage could improve the performance of the 2SLS estimator. This paper returns to Massoumi's idea, and investigates the performance of a Stein-like shrinkage estimator using modern asymptotic tools.

Just as in Maasoumi (1978) we consider a shrinkage estimator which is a weighted average of the ordinary least squares (OLS) and 2SLS estimators, with the weight inversely proportional to the Hausman (1978) statistic for exogeneity. We derive the asymptotic distribution of the shrinkage estimator using a local-to-exogeneity condition so that the distribution is continuous in the parameters. We calculate the asymptotic risk of the estimator and find that the risk is strictly less than the risk of the 2SLS estimator, so long as the number of right-hand-side endogenous variables exceeds two. This condition is analogous to the classic condition of Stein (1956) and James and Stein (1961) who found that shrinkage strictly reduces the risk of estimators of the mean in normal sampling when the dimension exceeds two.

The asymptotic theory used in this paper is similar to the local asymptotic theory used in Hansen (2015) who considers shrinkage estimators which combine unrestricted and resticted maximum likelihood estimators.

There are a number of limitations in the paper which can be investigated in future research. First, the analysis is confined to homoskedastic errors. This is not essential to the idea of shrinkage but greatly simplifies the calculations. A generalization to allow for heteroskedasticity and/or serial correlation would be valuable but more involved. Second, the analysis is confined to the 2SLS estimator. It would be useful to generalize to incorporate other estimators including LIML and GMM. Third, the risk analysis imposes the assumption that risk is measured by weighted squared error where the weight matrix is set equal to the inverse of the difference of the asymptotic variances of the 2SLS and OLS estimators. This might seem like an odd choice for weighted risk, but greatly simplifies the analysis. It would be useful to generalize the analysis to allow for a user-specific weight matrix, but again this would greatly complicate the derivations. Fourth, we do not derive nor explore minimax bounds on the estimation risk. These generalizations would be quite useful and valuable, but cannot be accomplished within the page limits of this contribution.

There are a number of related ideas in the previous literature. The idea of combining OLS and 2SLS estimator to reduce bias may have first been proposed by Sawa (1973). Phillips (1980) derived the exact distribution of the 2SLS estimator. Phillips (1984) derived the exact distribution of the Stein-rule estimator in linear regression. Ullah and Srivastava (1988) proposed a Stein-type estimator of the coefficients of structural simultaneous equations system and analyzed the distribution using small-sigma asymptotic methods. Kim and White (2001) studied asymptotic approximations to James-Stein-type estimators and gave conditions under which the Stein estimators preserve superiority when the sample size goes to infinity. Guggenberger (2010) examimes the asymptotic properties of a pretest estimator which conditions on the result of a first stage Hausman (1978) statistic. Our Stein-type estimator can be viewed as a smoothed version of the hard threshold Hausman pretest estimator. Chakravarty (2012) proposes a Stein-rule estimator similar to that studied in this paper, and discusses its properties using the theory of Kim and White (2001). Finally, Ditraglia (2014) has proposed combining OLS and 2SLS based on a plug-in estimate of the focused information criterion (as in Claeskens and Hjort, 2003).

The organization of the paper is as follows. In Section 2 we present the model and the Stein shrinkage estimator. In Section 3 we present the asymptotic distribution of the estimator under a local-to-exogeneity asymptotic assumption. In Section 4 we derive the asymptotic risk of the estimator. Section 5 presents a simulation experiment. Section 6 is a brief conclusion. Proofs of the theoretical results are presented in the appendix. The computer code used to generate the simulations, and a supplemental appendix, are available on the author's website.

2 A Stein 2SLS Estimator

Suppose we have a random sample $\{y_i, x_i, z_i : i = 1, ..., n\}$ with $z_i = (z_{1i}', z_{2i}')'$, where y_i is scalar, x_i is $m \times 1$, z_{1i} is $\ell \times 1$, and z_{2i} is $k \times 1$, with $k \geq m$. The variables are assumed to satisfy a standard instrumental variables equation

$$\begin{aligned} y_i &= x_i'\beta + z_{1i}'\gamma + e_i \qquad (1) \\ \mathbb{E}\left(z_i e_i\right) &= 0 \end{aligned}$$

The vector x_i is treated as endogenous, z_i as exogeneous, and the variables z_{2i} are valid excluded instruments. The parameter of interest is the coefficient β on the endogenous variables.

Two conventional estimators of equation (1) are ordinary least squares (OLS) and two-stage least squares (2SLS). Let $\widehat{\beta}_{OLS}$ denote the OLS estimator of β, let $\widehat{\beta}_{2SLS}$ denote the 2SLS estimator using z_i as instruments, and let $\widehat{V}_{OLS}$ and $\widehat{V}_{2SLS}$ denote the conventional covariance matrix estimators for $\widehat{\beta}_{OLS}$ and $\widehat{\beta}_{2SLS}$.

In large samples, 2SLS is typically preferred to OLS as it is consistent for β under endogeneity while OLS is inconsistent. However in small samples the 2SLS estimator can have a much larger variance so the OLS estimator can have better precision. Motivated by this observation, Maasoumi (1978) proposed a Stein-like weighted-average estimator with the weight depending on a specification test similar to the Hausman-Wu statistic

$$H_n = \left(\widehat{\beta}_{2SLS} - \widehat{\beta}_{OLS}\right)' \left(\widehat{V}_{2SLS} - \widehat{V}_{OLS}\right)^{-1} \left(\widehat{\beta}_{2SLS} - \widehat{\beta}_{OLS}\right).$$

We adopt his idea here to see if combination of OLS and 2SLS can result in improved estimation precision.

Specifically, we consider the following Stein-like estimator of β

$$\widehat{\beta}^* = w\widehat{\beta}_{OLS} + (1-w)\,\widehat{\beta}_{2SLS} \tag{2}$$

where

$$w = \begin{cases} \dfrac{\tau}{H_n} & \text{if } H_n \geq \tau \\ \\ 1 & \text{if } H_n < \tau \end{cases} \tag{3}$$

and τ is a shrinkage parameter. We recommend setting $\tau = m - 2$ when $m > 2$. This choice will be justified later. For now we leave τ as a free parameter so we can assess its impact on the performance of $\widehat{\beta}^*$. For $m \leq 2$ there is no strong theoretical guidance for choice of τ, but in the simulations which follow we set $\tau = 1$ for $m = 2$ and $\tau = 1/4$ for $m = 1$.

The restriction $w \in [0, 1]$ implicit in the definition (3) is identical to the "positive-part" restriction of the classic James-Stein estimator due to Baranchick (1970). As shown by Lemma 2 of Hansen (2014), positive-part trimming generically reduces estimation risk.

Notice that the weight (3) is inversely proportional to the Hausman statistic H_n, and recall that large values of H_n indicate the presence of endogeneity, and small values of H_n indicate exogeneity. Thus when there is evidence of endogeneity the estimator (2) puts more weight on 2SLS. Conversely, when there is no evidence of endogeneity the estimator (2) puts more weight on OLS. If H_n is sufficiently small (less than τ) then $w = 1$ and the Stein estimator (2) reduces to the OLS estimator.

3 Asymptotic Distribution

Our distribution theory is developed under a local asymptotic framework which is designed so that the Stein estimator has a non-degenerate asymptotic distribution. This can be achieved by specifying the model to be local to exogeneity. Specifically, first write the reduced form equation for the endogenous variable x_i as

$$\begin{aligned} x_i &= \Pi_1' z_{1i} + \Pi_2' z_{2i} + v_i \qquad (4) \\ \mathbb{E}\left(z_i v_i'\right) &= 0. \end{aligned}$$

Second, write the structural equation error e_i as a linear function of the reduced form error v_i and an orthogonal error ε_i

$$\begin{aligned} e_i &= v_i'\rho + \varepsilon_i \\ \mathbb{E}\left(v_i\varepsilon_i\right) &= 0 \end{aligned} \tag{5}$$

The equations (4) and (5) are without loss of generality as they can be defined by projection.

The variables x_i are exogenous if e_i and v_i are uncorrelated, or equivalently that the coefficient ρ is zero. We thus assume that this coefficient is local to zero, specifically

$$\rho = n^{-1/2}\delta \tag{6}$$

where the $m \times 1$ parameter δ indexes the degree of endogeneity. The variables x_i are exogenous if $\delta = 0$ and are (locally) endogenous when $\delta \neq 0$.

We now state our regularity conditions. Set $\Sigma = \mathbb{E}\left(v_iv_i'\right)$, $Q = \mathbb{E}\left(z_iz_i'\right)$ and $\sigma^2 = \mathbb{E}\left(\varepsilon_i^2\right)$.

Assumption 1 *$\mathbb{E}\left\|z_i\right\|^4 < \infty$, $\mathbb{E}\left\|v_i\right\|^4 < \infty$, $\mathbb{E}\varepsilon_i^4 < \infty$, $\mathbb{E}\left(\varepsilon_i^2 \mid z_i, x_i\right) = \sigma^2$, $rank(\Pi_2) = m$, $\Sigma > 0$, and $Q > 0$.*

Assumption 1 specifies that the variables have finite fourth moments (so that conventional central limit theory applies) and that the error ε_i is conditionally homoskedastic given the instruments and regressors. The latter assumption is used to simplify the asymptotic covariance expressions and is thus important for our analysis, but is not critical to the shrinkage technique. The rank condition on Π_2 ensures that the coefficient β is identified. The full rank condition on Σ is equivalent to assuming that x_i and z_i have no common elements.

Theorem 1 *Under Assumption 1,*

$$\sqrt{n}\begin{pmatrix} \widehat{\beta}_{OLS} - \beta \\ \widehat{\beta}_{2SLS} - \beta \end{pmatrix} \rightarrow_d h + \xi, \tag{7}$$

where

$$h = \begin{pmatrix} \sigma^{-2}V_1\Sigma\delta \\ 0 \end{pmatrix}$$

and $\xi \sim N(0, V)$ where

$$V = \begin{bmatrix} V_1 & V_1 \\ V_1 & V_2 \end{bmatrix}$$

with

$$\begin{aligned} V_1 &= \sigma^2\left(\mathbb{E}\left(x_ix_i'\right) - \mathbb{E}\left(x_iz_{1i}'\right)\left(\mathbb{E}\left(z_{1i}z_{1i}'\right)\right)^{-1}\mathbb{E}\left(z_{1i}x_i'\right)\right)^{-1} \\ V_2 &= \sigma^2\left(\mathbb{E}\left(x_iz_i'\right)\left(\mathbb{E}\left(z_iz_i'\right)\right)^{-1}\mathbb{E}\left(z_ix_i'\right) - \mathbb{E}\left(x_iz_{1i}'\right)\left(\mathbb{E}\left(z_{1i}z_{1i}'\right)\right)^{-1}\mathbb{E}\left(z_{1i}x_i'\right)\right)^{-1}. \end{aligned}$$

Furthermore, jointly with (7),

$$H_n \rightarrow_d (h+\xi)' P (h+\xi) \tag{8}$$

and

$$\sqrt{n}\left(\widehat{\beta}^* - \beta\right) \rightarrow_d G_2'\xi - \left(\frac{\tau}{(h+\xi)' P (h+\xi)}\right)_1 G'(h+\xi) \tag{9}$$

where $P = G(V_2 - V_1)^{-1} G'$, $G = (-I \quad I)'$, $G_2 = (0 \quad I)'$, *and* $(a)_1 = \min[1, a]$.

Theorem 1 presents the joint asymptotic distribution of the OLS and 2SLS estimators, the Hausman statistic, and the Stein estimator under the local exogeneity assumption. The joint asymptotic distribution of the OLS and 2SLS estimators is normal with a classic covariance matrix. The OLS estimator has an asymptotic bias when $\delta \neq 0$ but not the 2SLS estimator. The Hausman statistic has an asymptotic non-central chi-square distribution, with non-centrality parameter h depending on the local endogeneity parameter δ. The asymptotic distribution of the Stein estimator is a nonlinear function of the normal random vector in (7), and in particular is a function of the noncentrality parameter h.

4 Asymptotic Risk

We measure the efficiency of the estimators by asymptotic weighted mean-squared error. For any sequence of estimators T_n of β, and for some weight matrix W, we define the asymptotic risk as

$$R(T) = \lim_{\zeta\to\infty} \liminf_{n\to\infty} \mathbb{E} \min\left[n(T_n - \beta)' W (T_n - \beta), \zeta\right].$$

This is the expected scaled loss, trimmed at ζ, but in large samples ($n \to \infty$) and with arbitrarily negligible trimming ($\zeta \to \infty$). This definition of asymptotic risk is convenient as it is well defined and easy to calculate whenever the estimator has an asymptotic distribution, e.g.

$$\sqrt{n}(T_n - \beta) \rightarrow_d \psi \tag{10}$$

for some random variable ψ. For then (as shown by Lemma 6.1.14 of Lehmann and Casella (1998)) we have

$$R(T) = \mathbb{E}\left(\psi' W \psi\right) = \mathrm{tr}(W\mathbb{E}\left(\psi\psi'\right)). \tag{11}$$

Thus the asymptotic risk of any estimator T_n satisfying (10) can be calculated using (11).

It turns out to be convenient to set the weight matrix to equal $W = (V_2 - V_1)^{-1}$ and all our calculations will use this choice.

Theorem 2 *Under Assumption 1, if* $m > 2$ *and* $0 < \tau \leq 2(m-2)$ *then*

$$\begin{aligned} R(\widehat{\beta}_{2SLS}) &= \mathrm{tr}(WV_2) \\ R(\widehat{\beta}^*) &< \mathrm{tr}(WV_2) - \frac{\tau(2(m-2)-\tau)}{\sigma^{-4}\delta'\Sigma V_1 (V_2 - V_1)^{-1} V_1 \Sigma\delta + m} \end{aligned} \tag{12}$$

and thus the asymptotic risk of the Stein estimator is globally smaller than the asymptotic risk of the 2SLS estimator.

Theorem 2 shows that the asymptotic risk of the Stein estimator is strictly less than that of 2SLS for all parameter values, so long as m (the dimension of the endogenous variables x_i) exceeds 2. Theorem 2 holds under the mild regularity conditions of Assumption 1 and the local endogeneity framework (5)-(6). Since the inequality (12) is strict and holds for all values of localizing parameter δ, even very large values, this inequality shows that in a very real sense the Stein estimator strictly dominates 2SLS.

The assumption $m > 2$ is similar to Stein's (1956) classic condition for shrinkage. As shown by Stein (1956) the shrinkage dimension must exceed 2 in order for shrinkage to achieve global reductions in risk relative to unrestricted estimation.

The bound on the right-hand-side of (12) is minimized when $\tau = m - 2$, motiviating our recommendation for the shrinkage parameter, and with this choice the inequality $0 < \tau \leq 2(m-2)$ is trivally satisfied when $m > 2$.

To understand the magnitude of the risk improvement, define

$$a_m = \frac{\operatorname{tr}\left((V_2 - V_1)^{-1} V_2\right)}{m}$$

and

$$c_m = \frac{\sigma^{-4}\delta'\Sigma V_1 (V_2 - V_1)^{-1} V_1 \Sigma\delta}{m}.$$

The parameter a_m is a nonlinear function of the correlations between the regressors x_i and the instruments z_{2i} and satisfies $a_m \geq 1$. When the correlation between the regressors and instruments is high then a_m can be arbitrarily large. On the other hand when the correlation is weak (weak instruments) then a_m approaches one. The parameter c_m is a scalar measure of the strength of endogeneity δ. c_m is increasing as the magnitude of δ increases.

From (12) we can calculate

$$\begin{aligned}\frac{R(\widehat{\beta}^*)}{R(\widehat{\beta}_{2SLS})} &\leq 1 - \frac{m-2}{\operatorname{tr}\left((V_2 - V_1)^{-1} V_2\right)} \frac{m-2}{\sigma^{-4}\delta'\Sigma V_1 (V_2 - V_1)^{-1} V_1 \Sigma\delta + m} \\ &\simeq 1 - \frac{1}{a_m(c_m + 1)}\end{aligned}$$

with the approximation approaching equality as m increases. This shows that the percentage reduction in asymptotic risk achieved by the Stein estimator relative to the 2SLS estimator is approximately $100/(a_m(c_m + 1)$. This gain is highest when a_m is small (weak instruments), c_m is small (mild endogeneity), or m is large (as both a_m and c_m decrease with m). Thus we expect the Stein estimator to achieve particularly large reductions in risk when the instruments are weak, the degree of endogeneity is small, and/or the number of endogenous variables is large. On the other hand, the bound also suggests that the efficiency difference to be modest when instruments are strong or the degree of endogeneity is high.

One intuition for the risk reduction of Theorem 2 and the requirement $m > 2$ is that the OLS estimator can be viewed as a restricted GMM estimator which imposes m restrictions. Specifically, augment the model (1) with the additional equation

$$\mathbb{E}\left(x_i e_i\right) = \alpha \tag{13}$$

with α a free $m \times 1$ parameter. The GMM estimator of the system (1)-(13) under the homoskedasticity assumption is $\widehat{\beta}_{2SLS}$ as (13) adds no information. Now consider GMM estimation under the restriction $\alpha = 0$. This is equivalent to adding $\mathbb{E}\left(x_i e_i\right) = 0$ to (1). Under the homoskedasticity assumption this GMM estimator is $\widehat{\beta}_{OLS}$. This shows that the shrinkage estimator $\widehat{\beta}^*$ is equivalent to shrinking an unrestricted GMM estimator to a restricted GMM estimator. Since the number of restrictions exceeds 2 when $m > 2$, this satisfies the classic shrinkage result that Stein estimation reduces risk when the number of restrictions exceeds 2.

5 Simulation

Our simulation experiment uses a design similar to Donald and Newey (2001), Donald, Imbens and Newey (2009), and Kuersteiner and Okui (2010).

The observations (y_i, x_i, z_i) are generated by the process

$$\begin{aligned}
y_i &= x_i'\beta + \gamma + e_i \\
x_i &= \Pi z_i + \mu + v_i \\
\begin{pmatrix} e_i \\ v_i \end{pmatrix} &\sim N\left(0, \begin{pmatrix} 1 & \rho/\sqrt{m} & \cdots & \rho/\sqrt{m} \\ \rho/\sqrt{m} & 1 & 0 & 0 \\ \vdots & 0 & 1 & 0 \\ \rho/\sqrt{m} & 0 & 0 & 1 \end{pmatrix}\right) \\
z_i &\sim N(0, I_m)
\end{aligned} \tag{14}$$

Thus the instruments z_i, equation error e_i, and reduced form errors v_i are all $N(0,1)$, with the error e_i and elements of v_i having correlation $\rho/\sqrt{m}$, but all other correlations zero. The reason for setting the correlation equal to $\rho/\sqrt{m}$ is because then we can allow ρ to vary in $(-1,1)$. To see this, observe that the determinant of the covariance matrix in (14) is $1-\rho^2$, which is positive if and only if $|\rho| < 1$. Thus specifying the correlation as $\rho/\sqrt{m}$ with $\rho \in (-1,1)$ is a natural parameterization.

The distributions are invariant to β and γ so we set these parameters to zero. We also set $\mu = 0$ and set the $m \times m$ reduced form matrix as $\Pi = I_m d$ and the scale d set as $d = \sqrt{R^2/\left(1-R^2\right)}$ so that R^2 is the reduced form population R^2 for each x_{ji}.

We vary $n = \{100, 800\}$, $m = \{1,2,3,4\}$, $R^2 = \{0.01, 0.10, 0.40\}$ and ρ on a 40-point grid on $[0, 0.975]$. The parameter R^2 controls the strength of the instruments (small R^2 is the case of weak instruments) and the parameter ρ controls the degree of endogeneity ($\rho = 0$ is the case of exogenous regressors; large ρ is the case of strong endogeneity.)

Note that our experiment sets the dimension of z_i equal to that of x_i, so the 2SLS estimates are just-identified. Generalizing the simulation to add over-identification reduces the benefit of shrinkage but otherwise does not change the nature of the results.

We generated 50,000 samples for each configuration, and on each calculated $\widehat{\beta}_{OLS}$, $\widehat{\beta}_{2SLS}$ and $\widehat{\beta}^*$ where for the latter we set $\tau = 1/4$ for $m = 1$, $\tau = 1$ for $m = 2$, and $\tau = m - 2$ otherwise. We also calculated the Hausman pre-test estimator

$$\widehat{\beta}_{PT} = \widehat{\beta}_{OLS} 1 \left(H_n < c\right) + \widehat{\beta}_{2SLS} 1 \left(H_n \geq c\right)$$

where c is the 5% critical value from the χ^2_m distribution. This is the estimator examined by Guggenberger (2010) in the context of testing hypotheses on β. We also calculated the bias-corrected Sawa(1973) estimator, but its median squared error was nearly identical to that of 2SLS (most likely because our model is just-identified) so the results are not reported here.

To compare the estimators we calculated the median squared error (MSE) of each estimator, that is, for an estimator $\widehat{\beta}$,

$$R(\widehat{\beta}) = median \left(\left(\widehat{\beta} - \beta\right)' \left(\widehat{\beta} - \beta\right) \right) \tag{15}$$

We report the median squared error rather than the mean squared error because 2SLS estimators may not have finite variances in finite samples. This is common in the literature on the evaluation of estimation under simultaneity. While the asymptotic theory focused on mean squared error, we expect that the insights will carry over to the finite sample median squared error. Notice also that we evaluate based on an equally weighted squared error loss while Theorem 2 concerns a specific weighted squared error loss. We do this because while the weighted squared error is convenient for the theory, in practice we are more concerned with unweighted squared error.

To simplify presentation we present the relative median squared error, which is (15) divided by the median squared error of 2SLS. Thus values less than one indicate improved precision relative to 2SLS, and values greater than one indicate worse performance.

We present the results graphically, plotting the relative median squared error (15) as a function of the degree of endogeneity ρ. Each figure corresponds to a fixed set of $\{n, R^2, m\}$, with three lines representing the relative MSE of the estimators. Values less than one correspond to lower MSE than 2SLS.

All together, we generated 24 figures. To summarize the results, 5 representative figures are included here, with the full set of 24 figures posted as a supplemental appendix on the author's webpage. Figure 1 is the case $n = 100$, $R^2 = 0.01$ and $m = 1$. Figure 2 is the case $n = 100$, $R^2 = 0.40$ and $m = 1$. Figure 3 is the case $n = 100$, $R^2 = 0.01$ and $m = 2$. Figure 4 is the case $n = 100$, $R^2 = 0.10$ and $m = 2$. Figure 5 is the case $n = 800$, $R^2 = 0.40$ and $m = 4$. We select these cases as they are the most extreme, yet are quite representative of all cases examined. In particular, the case $R^2 = 0.01$ is quite important as it corresponds to an extreme weak instrument parameterization (and is a commonly used benchmark case in the simulation literature).

First consider the case of one endogenous regressor, $m = 1$. The 8 plots for this case look similar to either Figure 1 or Figure 2. OLS and the pre-test estimator have much smaller MSE than 2SLS for small values of ρ, but the ranking is reversed for large values of ρ. The MSE of the pre-test estimator is generally similar to OLS for small ρ and is similar to 2SLS for small ρ in many cases. For intermediate values of ρ the MSE of the pre-test estimator is typically much higher than 2SLS.

For small ρ the Stein estimator has lower MSE than 2SLS and can be quite close to that of OLS. For larger values of ρ, however, the Stein estimator has higher MSE than OLS, but its MSE is bounded unlike OLS. The region of dominance for OLS and the Stein estimator over 2SLS is greater for small values of R^2 and n. This can be seen by contrasting Figures 1 and 2. The plots for the $n = 800$ cases all look similar to Figure 2, where the Stein estimator achives some reduction in MSE relative to 2SLS for small values of ρ, but has higher MSE for moderate values of ρ.

Next consider the case of two endogenous regressors, $m = 2$. The 8 plots for this case look similar to Figures 2, 3, and 4. For large values of n and R^2, the plots are similar to Figure 2, where the Stein estimator has lower MSE than 2SLS for small ρ but the reverse holds for large ρ. For very small values of R^2, the plots are similar to Figure 3, where OLS and the Stein estimator are near equivalents and both have dramatically smaller MSE than 2SLS. What is happening here is that for very weak instruments, 2SLS has very high dispersion so OLS has smaller MSE, and the Stein estimator puts nearly all weight on the OLS estimator. For most cases, the plots are similar to Figure 4, where the Stein estimator uniformly dominates 2SLS, and has similar MSE to OLS for the small values of ρ where OLS has small MSE. Overall, the improvements in the Stein estimator over 2SLS are greatest in the cases of small sample sizes and weak instruments.

Finally consider the cases of three and four endogenous regressors $m = 3$ and $m = 4$. These are cases where Theorem 2 shows that the weighted asymptotic MSE of the Stein estimator is uniformly smaller than that of the 2SLS estimator. The 8 plots for this case look similar to Figures 4 and 5, and that the asymptotic uniform ranking holds in finite samples. Generally, the Stein estimator has much smaller MSE than 2SLS for small values of the endogeneity parameter ρ, and the estimators perform similarly for large values of ρ. The improvements of the Stein estimator are also greatest for the case of small samples (small n) and weak instruments (small R^2).

It is also instructive to examine the performance of the pre-test estimator. As is commonly found, the risk of the pre-test estimator is highly variable, with low MSE for small and very large ρ but very high MSE for intermediate values.

In summary, the simulation evidence provides strong finite sample confirmation of the predictions from the large sample approximations of Theorem 2. When the number of endogenous variables is three or larger, the Stein estimator uniformly dominates 2SLS. When the number of endogenous variables is less than 3, the Stein estimator has MSE which is either less than that of 2SLS or is not too much greater. The improvements achieved by the Stein estimator are particularly large in the empirically relevant context of small samples and weak instruments. The reductions in MSE due to Stein shrinkage are numerically large, perhaps surprisingly so.

6 Conclusion

Essie Maasoumi (1978) has made many important contributions to the field of econometrics. One of his most important theoretical contributions is his 1978 investigation of Stein shrinkage in simultaneous equations models. It is our pleasure to re-investigate this suggestion in the context of 2SLS estimation.

Our theory shows that a Stein-like shrinkage of 2SLS towards OLS using a weight inversely proportional to the classic Hausman statistic produces an estimator with reduced risk, specifically a reduction in asymptotic mean squared error, and reflected in finite samples as a reduction in median squared error.

7 Appendix

Proof of Theorem 1: Under the homoskedasticity assumption, the joint convergence (7) is a straightforward and standard calculation. The convergence in (8) follows immediately given the consistency of the covariance matrix estimates. (9) follows by the continuous mapping theorem. ■

Proof of Theorem 2: For convenience and without loss of generality assume $\sigma^2 = 1$.

Observe that $\sqrt{n}\left(\widehat{\beta}_{2SLS} - \beta\right) \rightarrow_d G_2'\xi \sim \mathrm{N}\left(0, V_2\right)$ under Theorem 1. Then (11) shows that

$$R(\widehat{\beta}_{2SLS}) = \mathbb{E}\left(\xi' G_2' W G_2' \xi\right) = \operatorname{tr}\left(W V_2\right). \tag{16}$$

Next, $\sqrt{n}\left(\widehat{\beta}^* - \beta\right) \rightarrow_d \psi$, where ψ is the limiting random variable in (9). Define an analogous random variable without positive part trimming

$$\psi^* = G_2'\xi - \left(\frac{\tau}{\left(\xi + h\right)' P \left(\xi + h\right)}\right) G' \left(\xi + h\right). \tag{17}$$

Then using (11) and the fact that the pointwise quadratic risk of ψ is strictly smaller than that of ψ^* (as shown, for example, by Lemma 2 of Hansen (2014)),

$$R(\widehat{\beta}^*) = \mathbb{E}\left(\psi' W \psi\right) < \mathbb{E}\left(\psi^{*\prime} W \psi^*\right). \tag{18}$$

Using (17), we calculate that

$$\begin{aligned} \mathbb{E}\left(\psi^{*\prime} W \psi^*\right) &= R(\widehat{\beta}_{2SLS}) + \tau^2 \mathbb{E}\left(\frac{1}{\left(\xi + h\right)' P \left(\xi + h\right)}\right) \\ &\quad - 2\tau \mathbb{E}\left(\eta(\xi + h)' G W G_2' \xi\right) \end{aligned} \tag{19}$$

where $\eta(x) = x/(x'Px)$. Since

$$\frac{\partial}{\partial x}\eta(\mathbf{x})' = \left(\frac{1}{x'Px}\right) I - \frac{2}{(x'Px)^2} Pxx',$$

$$G'VG = G_2'VG = V_2 - V_1 = W^{-1},$$

and

$$GWG_2'VP = GWG_2'VGWG' = GWG' = P,$$

then by Stein's Lemma (Lemma 1 of Hansen (2015) which is a matrix-notation version of Stein (1981))

$$\begin{aligned}
\mathbb{E}\left(\eta(\xi+h)'GWG_2'\xi\right) &= \mathbb{E}\operatorname{tr}\left(\frac{\partial}{\partial x}\eta\left(\xi+h\right)' GWG_2'V\right) \\
&= \mathbb{E}\left(\frac{\operatorname{tr}\left(GWG_2'V\right)}{\left(\xi+h\right)' P\left(\xi+h\right)}\right) \\
&\quad - 2\mathbb{E}\operatorname{tr}\left(\frac{P\left(\xi+h\right)\left(\xi+h\right)' GWG_2'V}{\left(\left(\xi+h\right)' P\left(\xi+h\right)\right)^2}\right) \\
&= \mathbb{E}\left(\frac{m-2}{\left(\xi+h\right)' P\left(\xi+h\right)}\right).
\end{aligned} \tag{20}$$

Thus (19) equals

$$R(\widehat{\beta}_{2SLS}) - \mathbb{E}\left(\frac{\tau\left(2\left(m-2\right)-\tau\right)}{\left(\xi+h\right)' P\left(\xi+h\right)}\right) \leq R(\widehat{\beta}_{2SLS}) - \frac{\tau\left(2\left(m-2\right)-\tau\right)}{\mathbb{E}\left(\xi+h\right)' P\left(\xi+h\right)} \tag{21}$$

where use has been made of Jensen's inequality and the assumption that $\tau \leq 2\left(m-2\right).$

We calculate that since $\operatorname{tr}\left(VP\right) = \operatorname{tr}\left(WG'VG\right) = m$,

$$\begin{aligned}
\mathbb{E}\left(\xi+h\right)' P\left(\xi+h\right) &= h'Ph + \operatorname{tr}\left(VP\right) \\
&= \sigma^{-4}\delta'\Sigma V_1\left(V_2 - V_1\right)^{-1} V_1\Sigma\delta + m.
\end{aligned}$$

Substituted into (21) we have established

$$R(\widehat{\beta}^*) \leq R(\widehat{\beta}_{2SLS}) - \frac{\tau\left(2\left(m-2\right)-\tau\right)}{\sigma^{-4}\delta'\Sigma V_1\left(V_2 - V_1\right)^{-1} V_1\Sigma\delta + m}$$

as claimed. ■

References

[1] Baranchick, A. J. (1970): "A family of minimax estiamtors of the mean of a multivariate normal distribution," *Annals of Mathematical Statistics*, 41, 642-645.

[2] Chakravarty, Tirthankar (2012): "Shrinkage estimators for structural parameters," working paper, UCSD.

[3] Claeskens, Gerda and Nils L. Hjort (2003): "The focused information criterion," *Journal of the American Statistical Association*, 98, 900-945.

[4] Ditraglia, Francis, J. (2014): "Using invalid instruments on purpose: Focused moment selection and averaging for GMM," working paper, University of Pennsylvania.

[5] Donald, Stephen G. and Whitney K. Newey (2001): "Choosing the number of instruments," *Econometrica*, 69, 1161-1191.

[6] Donald, Stephen G., Guido W. Imbens, and Whitney K. Newey (2009): "Choosing instrumental variables in conditional moment restriction models," *Journal of Econometrics*, 152, 28-36.

[7] Guggenberger, Patrik (2010): "The impact of a Hausman pretest on the asymptotic size of a hypothesis test," *Econometric Theory*, 26, 369-382.

[8] Hansen, Bruce E. (2014): "Shrinkage efficiency bounds," *Econometric Theory*, forthcoming.

[9] Hansen, Bruce E. (2015): "Efficient shrinkage in parametric models," working paper.

[10] Hausman, Jerry A. (1978): "Specification tests in econometrics," *Econometrica*, 46, 1251-1271.

[11] James W. and Charles M. Stein (1961): "Estimation with quadratic loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 361-380.

[12] Kim, Tae-Hwan and Halbert White (2001): "James-Stein-type estimators in large samples with application to the least absolute deviations estimator," *Journal of the American Statistical Association*, 96, 697-705.

[13] Kuersteiner, Guido and Ryo Okui (2010): "Constructing optimal instruments by first-stage prediction averaging," *Econometrica*, 78, 697-718.

[14] Maasoumi, Esfandia (1978): "A modified Stein-like estimator for the reduced form coefficients of simultaneous equations, *Econometrica*, 46, 695-703.

[15] Phillips, Peter C. B. (1980): "The exact distribution of instrumental variable estimators in an equation containing $n+1$ endogenous variables," *Econometrica*, 48, 861-878.

[16] Phillips, Peter C. B. (1983): "Exact small sample theory in the simultaneous equations model," in *Handbook of Econometrics, Vol I*, ed. by Z. Griliches and M. D. Intriligator, 449-516.

[17] Phillips, Peter C. B. (1984): "The exact distribution of the Stein-rule estimator," *Journal of Econometrics*, 25, 123-131.

[18] Sawa, Takamitsu (1973): "Almost unbiased estimator in simultaneous equations systems," *International Economic Review*, 14, 97-106.

[19] Stein, Charles M. (1956): "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," *Proc. Third Berkeley Symp. Math. Statist. Probab.*, 1, 197-206.

[20] Stein, Charles M. (1981): "Estimation of the mean of a multivariate normal distribution," *Annals of Statistics*, 9, 1135-1151.

[21] Ullah, Aman and V. K. Srivastava (1988): "On the improved estimation of structural coefficients," *Sankhya*, 50, 111-118.

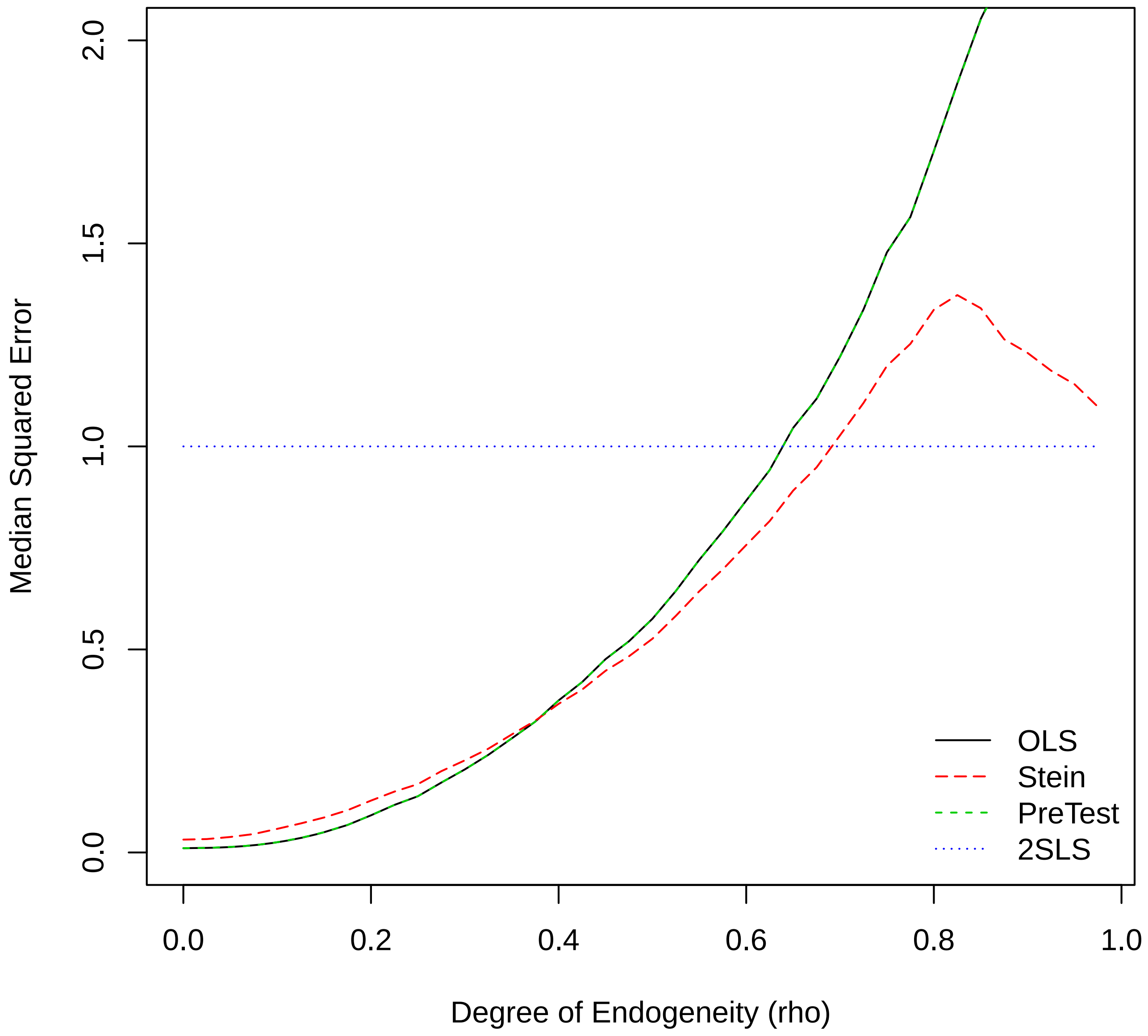

Figure 1: Relative Median Squared Error of OLS, Stein, and Pretest Estimators, $n = 100$, $R^2 = 0.01$, $m = 1$

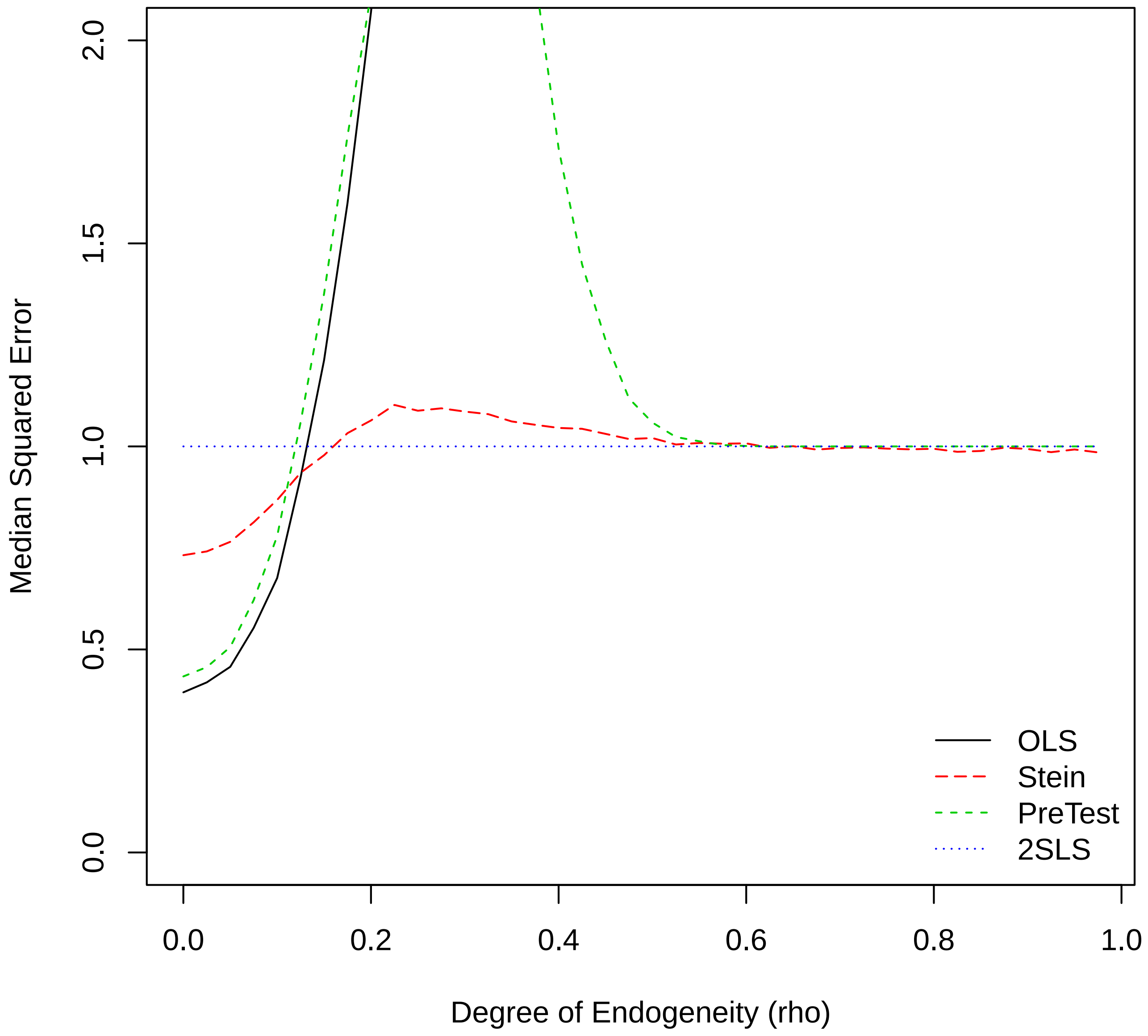

Figure 2: Relative Median Squared Error of OLS, Stein, and Pretest Estimators, $n = 100$, $R^2 = 0.40$, $m = 1$

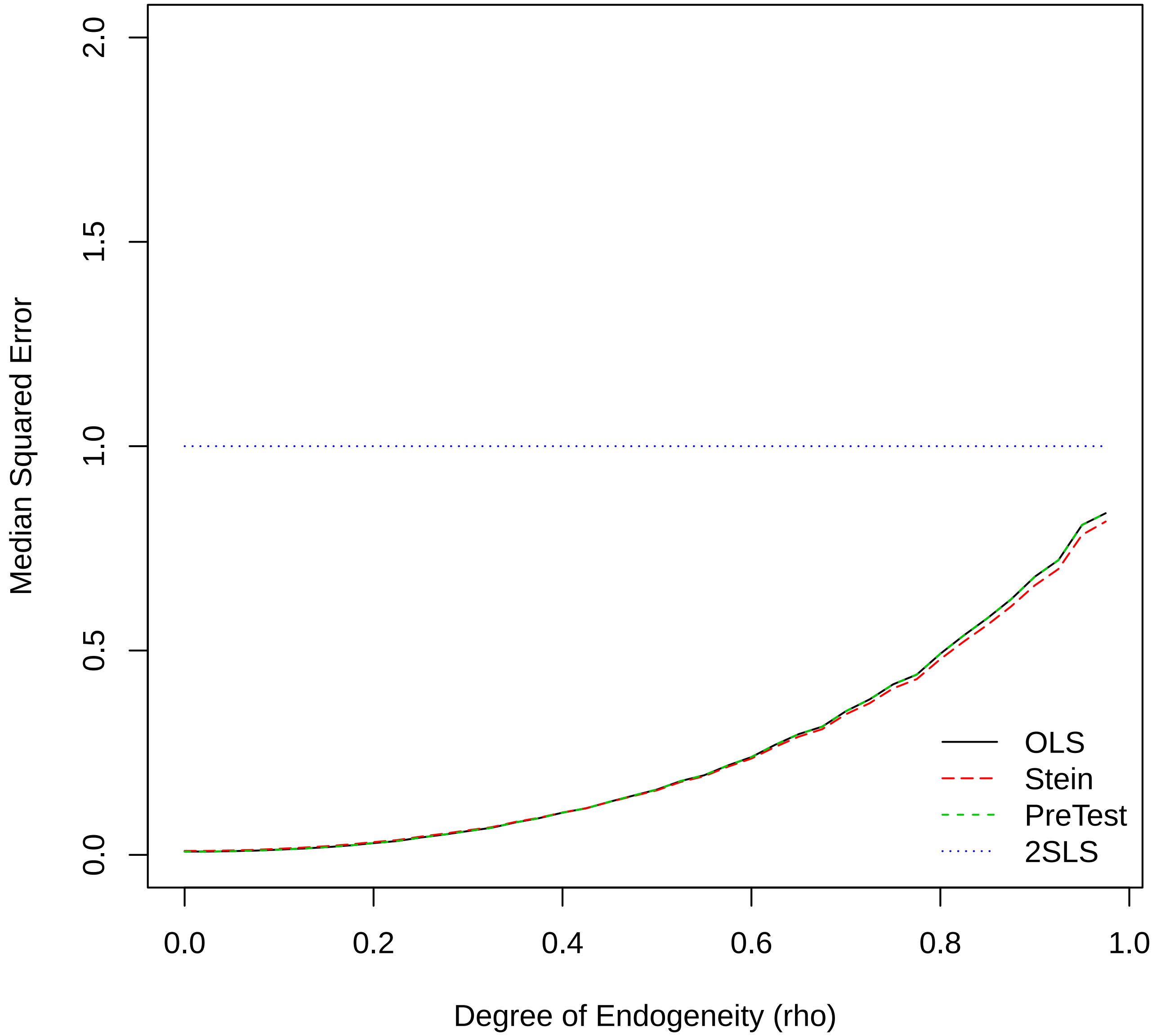

Figure 3: Relative Median Squared Error of OLS, Stein, and Pretest Estimators, $n = 100$, $R^2 = 0.01$, $m = 2$

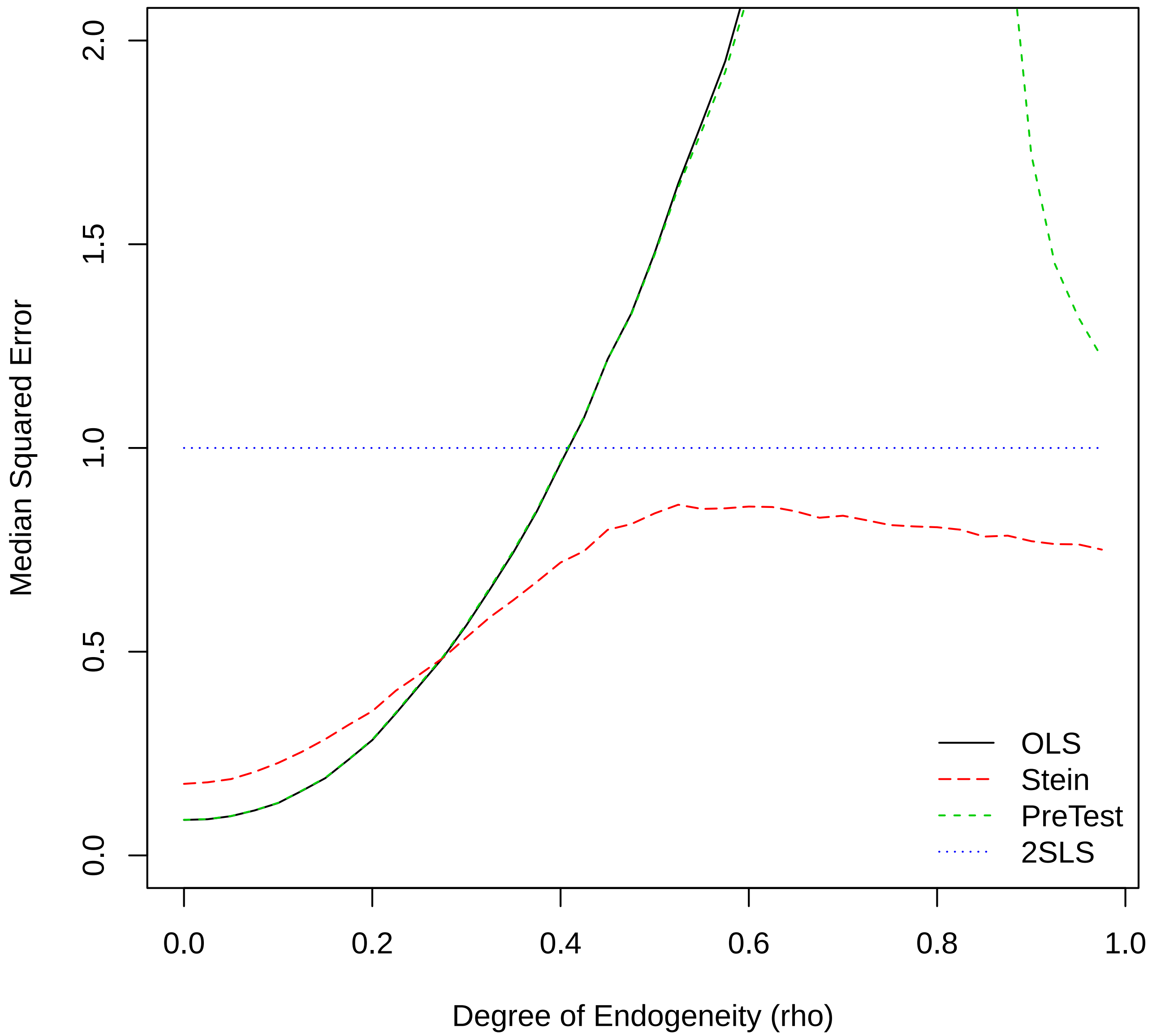

Figure 4: Relative Median Squared Error of OLS, Stein, and Pretest Estimators, $n = 100$, $R^2 = 0.10$, $m = 2$

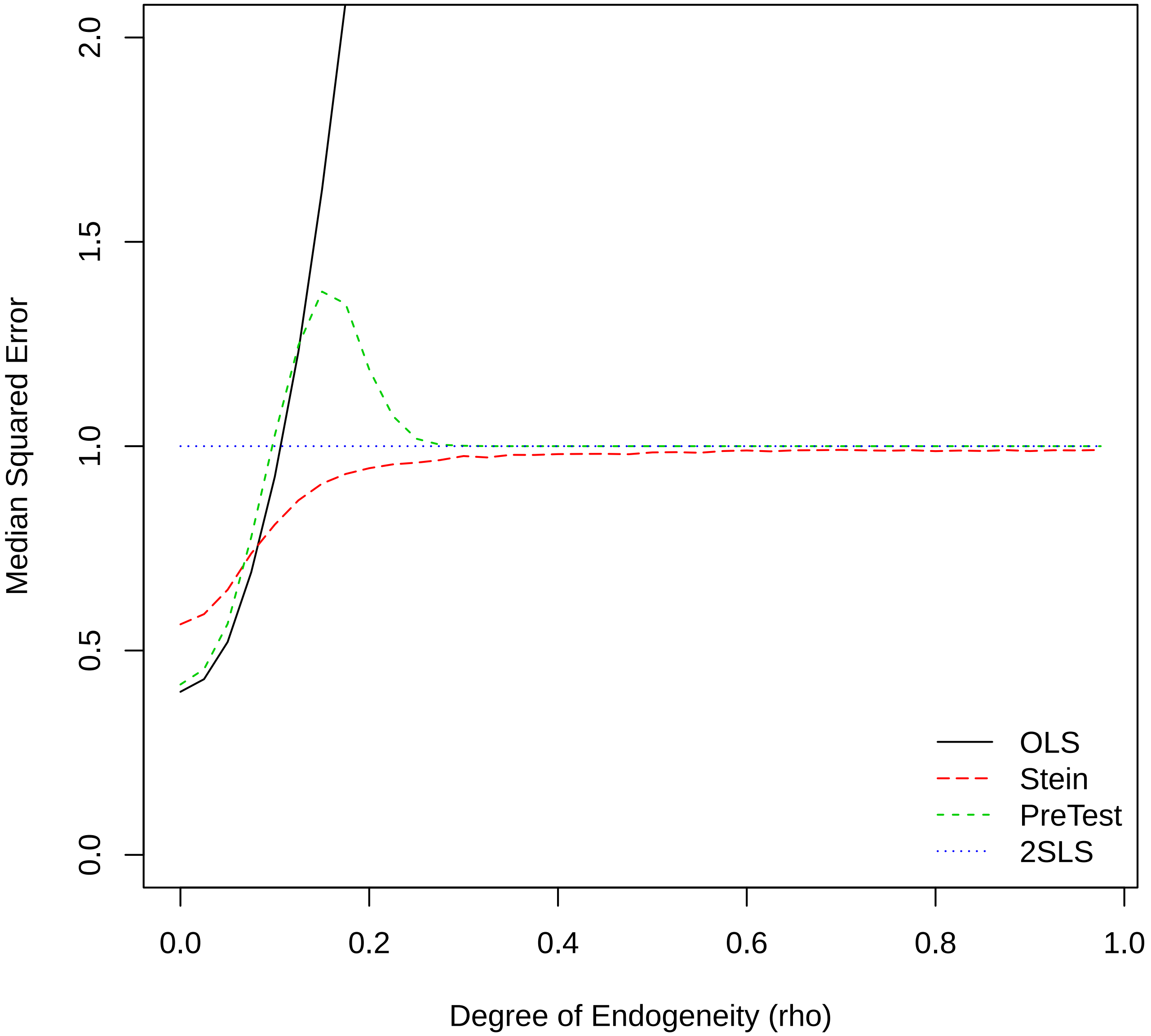

Figure 5: Relative Median Squared Error of OLS, Stein, and Pretest Estimators, $n = 800$, $R^2 = 0.40$, $m = 4$