Heteroskedasticity Correction and Dimension Reduction

Joris Pinkse*

November 2006

Abstract

This paper provides a nonparametric method of correcting for heteroskedasticity in linear regression models with independent and identically distributed (i.i.d.) observations. The new estimator requires an empiricist to select a small set (or index) of variables which are deemed to be the most important in explaining the presence of heteroskedasticity. The new estimator is the most efficient estimator in a wide class of estimators for which the heteroskedasticity correction can only depend on the variables chosen. The nonparametric correction uses k–nearest neighbor (KNN) estimation.

*Department of Economics, The Pennsylvania State University, 608 Kern Graduate Building, University Park 16802, joris@psu.edu, http://joris.econ.psu.edu. I thank Min Ahn, Don Andrews, Herman Bierens, Pierre Brochu, Bertrand Clarke, David Green, Lars Hansen, Ken Hendricks, Barry Ickes, Yuichi Kitamura, James MacKinnon, Daniel McFadden, Geert Ridder, Quang Vuong, and seminar participants at UBC, UC Berkeley, Queens, Carleton, the University of Texas, Rice university, the university of Pennsylvania, Tilburg University, Georgetown university and UCLA for helpful suggestions.

1 Motivation

This paper provides a nonparametric method of correcting for heteroskedasticity in linear regression models with independent and identically distributed (i.i.d.) observations. The new estimator requires an empiricist to select a small set of variables which are deemed to be the most important in explaining the presence of heteroskedasticity. The new estimator is the most efficient estimator in a wide class of estimators for which the heteroskedasticity correction can only depend on the variables chosen. The nonparametric correction uses k–nearest neighbor (KNN) estimation.

The presence of heteroskedasticity does not affect the consistency and asymptotic normality of least squares estimators, but it renders them inefficient and results in a more complicated form of the asymptotic variance matrix. In the linear regression model (LRM), for instance, the standard OLS t–statistics need to be corrected for valid inference, possibly by means of a White–correction.[1] Although cross–sections data sets in economics are often large, the model typically explains only a small fraction of the observed variation. A large error variance results both in inaccurate estimates and in low t–statistics (when the 'true' coefficient values are nonzero). Correcting for heteroskedasticity can improve both estimator accuracy and t–statistics.

If asymptotic efficiency were the only concern, the best method in a LRM would be to estimate the conditional variance function v nonparametrically, as suggested by Robinson (1987); see also Carroll (1982) and Delgado (1992). Robinson's nonparametric generalized least squares (GLS) procedure results in (asymptotically) efficient estimates under minimal conditions on v and the model variables. However, since the precision of nonparametric estimates decreases exponentially in the number of dimensions (i.e. the number d_x of regressors x_i) due to the *curse of dimensionality,* in most applications v is unlikely to be estimated accurately. Consequently, the optimal correction is unlikely to obtain in many applications in economics due to the typically large number of regressors.

It is possible to use only a vector h_i of smaller dimension than x_i in the estimation of v. Such a procedure would mitigate the curse of dimensionality problem, but it would not result in consistent estimation of v unless the conditional variance function does not depend on the variables left out. As a consequence, the resulting estimator $\hat{\theta}$ of the unknown parameter vector θ_0 could have an asymptotic variance larger than that of the OLS estimator.

The same problem applies to parametric GLS, as first proposed by Aitken (1935); see also e.g. Godfrey (1978), Goldfeld and Quandt (1965) and Harvey (1976). With parametric GLS, one

[1] See e.g. Eicker (1963), Huber (1967) and White (1980).

assumes a parametric form for v and estimates the associated parameters. Correct specification of v yields efficient estimates of θ_0, but misspecification can cause the GLS estimator to be less precise than the OLS estimator, even asymptotically. The problem is that economic theory is rarely informative about the choice of v, and there are typically few reasons to choose one particular function over another.

Cragg (1992) proposed a procedure in which the parametric form for v is plugged into the sandwich form of the asymptotic variance matrix of the corresponding weighted least squares (WLS) estimator, which is then optimized with respect to the v–parameters. The problem is that the choice of parameters which are optimal for one of the regression coefficients is typically suboptimal for another. Cragg hence proposed to minimize some scalar–valued function of the asymptotic variance matrix. His procedure guarantees an efficiency improvement over OLS in terms of the objective function specified, but the Cragg procedure is generally not optimal for every coefficient.

The method proposed in this paper is based on the supposition that even though there is usually little information about the functional form of v and that v can be a function of all elements of x_i, there is typically some information about which subset h_i of the regressors is responsible for most of the heteroskedasticity. For instance, growth rates of small firms tend to be more variable than those of large firms, but the precise manner in which the conditional variance of growth rates varies with firm size is unclear.[2] The same applies to cross–country growth rates, see e.g. the scatterplot of growth versus per capita GDP in figures 2.5 and 2.6 of Romer (1989). A third example is that of a Mincer (1970) equation, in which (log)earnings are regressed on experience and other covariates; more experience means higher earnings on average, but the variability in earnings also increases; see e.g. Mincer (1974), Chart 6.1, and Dooley and Gottschalk (1984).[3]

The procedure proposed here finds the optimal correction mechanism in a wide class of such procedures in which the correction factor can only depend on h_i. The procedure is of the WLS variety, albeit that the optimal weights are matrix–valued, not scalar–valued. If h_i only contains the constant term, the new procedure is equivalent to OLS. If, at the other extreme, $h_i = x_i$, then the new estimator is identical to the Robinson (1987) estimator, otherwise the proposed procedure can be less efficient than the Robinson procedure asymptotically, but it can be better in moderate size samples because of the curse of dimensionality problem of the Robinson method. In contrast to the Cragg (1992) estimator the procedure is optimal in all directions simultaneously.

[2] See e.g. Hall (1987), Amit et al. (1997), Petrunia (2002), for interesting studies on firm growth rates.

[3] These examples ignore the potential for right hand side endogeneity.

A maintained assumption in this paper is that one has a correctly specified model with a potential presence of heteroskedasticity. Although there are instances in which it can be difficult to differentiate between the presence of heteroskedasticity and model misspecification (see e.g. Hall, 1987), such issues are beyond the scope of this paper. And even though White–corrected t–statistics have sometimes been found to be close to uncorrected ones,[4] such a finding only suggests that using uncorrected t–statistics is innocuous; it neither implies an absence of heteroskedasticity nor the efficiency of OLS; see example E3 in an appendix.

The new procedure can be interpreted as the solution to an optimal instrument selection problem, albeit for an unusual set of moment conditions. There are several generic estimators available which can be used for such problems including Robinson (1991), Newey (1990,1993) Donald et al. (2003) and Kitamura et al. (2001), but the conditions required for those estimators do not apply to the present case.

An alternative possibility is to use an increasing sequence of moment conditions, an idea first proposed by Cragg (1983) specifically for the purpose of heteroskedasticity correction in the LRM. Knowledge about heteroskedasticity of the above–described kind helps one to choose which moment conditions to use. This issue is discussed in section 2.

Like Robinson (1987) and Newey (1990), I use KNN estimation, originally due to Fix and Hodges (1951),[5] albeit that the estimation problem here is different. KNN is a nonparametric estimation technique in which estimates are determined by averaging over a fixed number of neighboring observations instead of over all observations within a given distance, as with nonparametric kernel estimation. I use KNN estimation since the regularity conditions required are minimal — not even continuity of v is needed — and various regressor distributions can be accomodated naturally.[6] This is in contrast to other potential nonparametric estimators like kernel regression (used by Carroll, 1982) or series estimation (Newey, 1990). Further, since an average is always taken over the same number of observations, the KNN estimator variances are rarely a problem, thereby obviating the need for an awkward trimming procedure, as would be the case with kernel estimators. Computation of KNN estimates is however cumbersome and (computer–)time–consuming. For instance, if the number of neighbors $k_n = n^{4/5}$, with n the sample size, then if $n = 1,000$ one needs to determine

[4] This in fact forms the basis of White's (1980) well–known test of heteroskedasticity.

[5] In fact, the proofs in this paper are similar to those in Robinson (1987) and Newey (1990), especially Robinson (1987).

[6] The regularity conditions required are weak because pointwise convergence of the nearest neighbor estimates is not needed.

the 251 nearest neighbors of each of the 1,000 observations. Good algorithms exist, but the gain achieved by using these deteriorates with an increase in k_n and the dimension of the conditioning vector: x_i in the case of Robinson, h_i here. This is a secondary problem with a full Robinson–style correction. It is nevertheless practically feasible to use the proposed procedure on problems with 30,000 observations and 50 regressors. Computation is much faster if the elements in h_i have a discrete distribution.

The proposed procedure has several limitations. First, I only allow for i.i.d. data. Time series and spatial data could be permitted at the expense of stronger assumptions. One would then also want to explore the possibility of exploiting the dependence across observations to improve efficiency. A second limitation concerns the choice of k_n. In this paper it is assumed that k_n is fixed and chosen by the practitioner. The choice of k_n is 'discrete' and KNN estimators are far less sensitive to the choice of k_n than are kernel estimators to the choice of a bandwidth, due to the afore–mentioned tail variance problems of kernel estimators. Automatic choices of k_n are possible,[7] but their optimality will depend on further assumptions.

The outline of the paper is as follows. In section 2 I propose the new estimator and discuss its uses and benefits. The theoretical properties of the estimator are established in section 3, section 4 presents the results of a simulation study and suggestions for potential extensions are in section 5.

2 Heteroskedasticity Correction

In the linear regression model

$$y_i = x_i'\theta_0 + u_i, \quad i = 1, \ldots, n,$$

where $E(u_i|x_i) = 0$ a.s., $v(x) = E(u_i^2|x_i = x)$ is allowed to vary with x. The idea is to consider *matrix–weighted least squares estimators* of the form

$$\hat{\hat{\theta}} = \left(\sum_{i=1}^{n} \hat{A}_i x_i x_i' \right)^{-1} \sum_{i=1}^{n} \hat{A}_i x_i y_i,$$

where the $\hat{A}_i$'s are estimated weight matrices. The optimal $\hat{A}_i$ is an estimate of $A_i = x_i x_i'/(||x_i||^2 v_i)$ where $v_i = v(x_i)$.

Since v — and hence A — has a high–dimensional argument (x), the curse of dimensionality precludes accurate estimation of A without further assumptions. Instead, A is allowed to depend

[7] See e.g. Li (1984).

only on a vector of lower dimension h_i, here taken to be a subvector of x_i necessarily including the constant.

The optimal such correction matrix is given by (see L2 in the appendix).

$$A(h) = L(h)Q^+(h) = E(x_i x_i'|h_i = h)\big(E(u_i^2 x_i x_i'|h_i = h)\big)^+,$$

where a + superscript denotes a *Moore–Penrose inverse*. Estimation procedures which involve a sequence of Moore–Penrose inverses are unattractive. It is hence preferable to write A in a form that does not involve Moore–Penrose inverses if this is possible. Since h_i is a subvector of x_i, x_i can be partitioned as $x_i = [h_i', b_i']'$ such that for $t_i = [1, b_i']'$,

$$L(h) = \begin{bmatrix} hh' & h_i E(b_i'|h_i = h) \\ E(b_i|h_i = h)h' & E(b_i b_i'|h_i = h) \end{bmatrix} = \begin{bmatrix} h & 0 \\ 0 & I \end{bmatrix} E(t_i t_i'|h_i = h) \begin{bmatrix} h' & 0 \\ 0 & I \end{bmatrix} = H(h)S(h)H'(h),$$

$$Q(h) = \begin{bmatrix} E(u_i^2|h_i = h)hh' & h_i E(u_i^2 b_i'|h_i = h) \\ E(u_i^2 b_i|h_i = h)h' & E(u_i^2 b_i b_i'|h_i = h) \end{bmatrix} = \begin{bmatrix} h & 0 \\ 0 & I \end{bmatrix} E(u_i^2 t_i t_i'|h_i = h) \begin{bmatrix} h' & 0 \\ 0 & I \end{bmatrix} = H(h)T(h)H'(h),$$

for implicitly defined S, T, H functions. Provided that $T(h) > 0$,

$$Q^+(h) = \big(H'(h)\big)^+ T^{-1}(h)H^+(h), \text{ where } H^+(h) = \begin{bmatrix} \frac{h'}{||h||^2} & 0 \\ 0 & I \end{bmatrix}.$$

Thus,

$$A(h) = H(h)S(h)T^{-1}(h)H^+(h), \tag{1}$$

and

$$A(h)x = H(h)S(h)T^{-1}(h)t, \tag{2}$$

for any vectors $x = [h', b']', t = [1, b']'$.

Let $A_i = A(h_i)$ and similarly for T, S, L, Q, H. It will be shown that

$$\sqrt{n}(\hat{\hat{\theta}} - \theta_0) \xrightarrow{d} N\Big(0, \big(E(L_1 Q_1^+ L_1)\big)^{-1}\Big). \tag{3}$$

In the case of homoskedasticity, $T(h) = Eu_i^2 S(h)$, such that $A(h)x = H(h)H^+(h)x/Eu_i^2 = H(h)t/Eu_i^2 = x$, which means that the infeasible version of $\hat{\hat{\theta}}$ would be the OLS estimator. So efficiency improvements only arise when there is heteroskedasticity.

There are a few important points to be made here. First, in all but a few cases no scalar WLS estimators will optimize the asymptotic variances of all regression coefficient estimators simultaneously. In particular if scalar weights $1/E(u_i^2|h_i)$ are chosen — as one would do when the

(infeasible) Robinson–GLS procedure is used with h_i instead of x_i — then the asymptotic variance of the scalar–WLS estimator can be worse than that of the OLS estimator as the following example demonstrates.

E1 *Let $x_i = [h_i, b_i]'$ where h_i, b_i are independent Bernoulli with probability $1/2$. Suppose that $E(u_i^2|h_i, b_i) = I(h_i + b_i \neq 0) + a^{-1}I(h_i + b_i = 0)$ for some $a > 0$. Then the OLS and Robinson (conditioning only on h) variance matrices are*

$$\frac{4}{3}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \quad \frac{4}{(2a+1)^2}\begin{bmatrix} 3a^2+2a+1 & -2a^2-1 \\ -2a^2-1 & 4a^2+2 \end{bmatrix}.$$

The difference between the Robinson and OLS variance matrices is

$$\frac{4(a-1)^2}{3(2a+1)^2}\begin{bmatrix} 1 & -2 \\ -2 & 4 \end{bmatrix} \geq 0.$$

The above difference is only zero when $a = 1$. Note that the difference is singular, which in this case means that $2\hat{\theta}_h + \hat{\theta}_b$ has the same asymptotic variance irrespective of which estimator is used. In the worst case scenario the ratio of variances for $\hat{\beta}_b$ (when $a \downarrow 0$) is $(8/3)/(8) = 1/3$ in favor of OLS, but this ratio can be made arbitrarily large by changing the Bernoulli probabilities. □

One scenario in which the Robinson–procedure using h_i is fully efficient is when b_i is independent of both h_i and u_i, in which case it would be better to perform separate regressions of y_i on h_i and y_i on $t_i = [1, b_i']'$ instead of regressing y_i on both h_i and b_i.

Efficiency improvements obtain even if h_i is independent of both u_i and z_i since $E(t_i t_i'|h_i = h)$ and $E(u_i^2 t_i t_i'|h_i = h)$ will then not vary with h but will also not be diagonal; see the following example.

E2 *Let $h_i = [1, \tilde{h}_i]'$ where $\tilde{h}_i$ is standard normal and b_i be a random variable taking the values of 1 and -1 with probability $1/2$. Let further h_i, b_i be independent and $v(x_i) = v(b_i) = (4 - 2b_i)/3$. Then*

$$A_i = \begin{bmatrix} \frac{h_i h_i'}{||h_i||^2} & \frac{h_i}{2} \\ \frac{h_i'}{2||h_i||^2} & 1 \end{bmatrix}, \quad A_i x_i = \begin{bmatrix} (1 + b_i/2)h_i \\ 1/2 + b_i \end{bmatrix}.$$

The proposed and OLS estimators can then be written as

$$\left(\frac{3}{2}\sum_{b_i=1}\begin{bmatrix} h_i h_i' & h_i \\ h_i' & 1 \end{bmatrix} + \frac{1}{2}\sum_{b_i=-1}\begin{bmatrix} h_i h_i' & -h_i \\ -h_i' & 1 \end{bmatrix}\right)^{-1}\left(\frac{3}{2}\sum_{b_i=1}\begin{bmatrix} h_i y_i \\ y_i \end{bmatrix} + \frac{1}{2}\sum_{b_i=-1}\begin{bmatrix} h_i y_i \\ -y_i \end{bmatrix}\right),$$

and

$$\left(\sum_{b_i=1}\begin{bmatrix} h_ih_i' & h_i \\ h_i' & 1 \end{bmatrix}+\sum_{b_i=-1}\begin{bmatrix} h_ih_i' & -h_i \\ -h_i' & 1 \end{bmatrix}\right)^{-1}\left(\sum_{b_i=1}\begin{bmatrix} h_iy_i \\ y_i \end{bmatrix}+\sum_{b_i=-1}\begin{bmatrix} h_iy_i \\ -y_i \end{bmatrix}\right).$$

The new estimator hence manages to reduce the weight on the observations for which $b_i=-1$, i.e. the high variance observations. The resulting asymptotic variances are

$$\frac{1}{3}\begin{bmatrix} 4 & 0 & -2 \\ 0 & 3 & 0 \\ -2 & 0 & 4 \end{bmatrix} \text{ and } \frac{1}{3}\begin{bmatrix} 4 & 0 & -2 \\ 0 & 4 & 0 \\ -2 & 0 & 4 \end{bmatrix}.$$

In this example, therefore, only the variance on the coefficient estimator of $\tilde{h}_i$ is reduced. But note also that despite the independence of h_i and (u_i, b_i), the proposed estimator here coincides with the infeasible GLS estimator. □

When h_i is independent of both u_i, b_i, however, it is still better to drop (the nonconstant portion of) h_i from the regression altogether; only if all elements of h_i other than the constant have mean zero is efficiency the same. So there is no gain from adding 'phantom' regressors to the model in order to improve efficiency.

The new estimator can be viewed as an optimal instrument selection problem when the conditional moment conditions are

$$E\big(x_i(y_i - x_i'\theta_0)|h_i\big) = 0 \text{ a.s.}.$$

Indeed, the optimal unconditional moment condition for this problem is

$$E\Big(E(x_ix_i'|h_i)\big(E(u_i^2x_ix_i'|h_i)\big)^+x_i(y_i - x_i'\theta_0)\Big) = 0.$$

Besides the procedure for which I provide theoretical results, it is hence possible to achieve an asymptotically equally efficient estimator by using an IV–estimator with instruments of the form $z_i = \Upsilon_n(h_i)x_i$ for some matrix–valued function Υ_n whose number of rows increases slowly with the sample size. This possibility is based on Cragg's (1983) estimator and is explored in the simulation study.

3 Theory

I now establish the asymptotic properties of my estimator.

A1 *For some* $p_h, p_t, p_u > 0$ *satisfying* $2/p_h + 6/p_t + 2/\min\{p_u, p_x\} < 1$, $E||h_1||^{p_h} < \infty, E||t_1||^{p_t} < \infty, E|u_1|^{p_u} < \infty$.

In case $h_i = x_i$ (full Robinson correction), $t_i = 1$ and A1, like Robinson (1988), requires existence of moments greater than four for regressors and errors. If h_i is a strict subvector of x_i then the condition is satisfied e.g. when regressors and errors have moments greater than ten or when the t–regressors have moments greater than eighteen and h–regressors and errors have sixth moments. Existence of moments is not usually a concern in empirical applications with cross sectional data, but an implication of A1 is that regressors with fat–tailed distributions are better put in h than in t.

I use nearest neighbor weights w_{ij} here, i.e. $w_{ij} > 0$ for the k_n observations j closest to observation i (including observation i itself) and $w_{ij} = 0$, otherwise. In the case of ties, the positive weight(s) is (are) randomly allocated among the observations at equal distance to observation i. The weights are further such that for some $C_w > 0$, $\max_j w_{ij} \leq C_w/k_n$. The presumption is that k_n is chosen and nonrandom; no automatic selection mechanism is provided here.

A2 $\lim_{n\to\infty} n^{2/3}/k_n = 0$ *and* $\lim_{n\to\infty} k_n/n = 0$.

Finally, we make some simplifying regularity assumptions.

A3 *(i)* $E(x_1x_1') > 0$, *(ii)* $\inf_x v(x) > 0$ *and (iii)* $\inf_h E(t_it_i'|h_i = h) > 0$.

Part (i) excludes the possibility of collinearity and (ii) and (iii) together guarantee the invertibility of T_i. Part (ii) is also part of Robinson's (1988) assumptions. Part (iii), which is trivially satisfied when $h_i = x_i$, is equivalent to assuming that $\inf_h V(b_i|h_i = h) > 0$. This condition can be violated, e.g. when b_i is wage income and h_i is number of hours worked: when $h_i = 0$, $b_i = 0$, also. Violation of (iii) will be infrequent but can be resolved by moving variables between h_i and b_i, which admittedly is less than ideal.

T1 *If assumptions A1–A3 hold, then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\Big(0, \big(E(L_1Q_1^+L_1)\big)^{-1}\Big).$$

4 Simulations

To evaluate the performance of the new estimator I have conducted an extensive simulation exercise. The purpose here is not to discover which correction strategy is best in a typical empirical case,

since the premise is that no knowledge on the form of v is available. The experiments have instead been designed to determine under which circumstances the proposed estimator should be used instead of various alternatives.

The entries in each of the tables in appendix F is the simulated root mean square error (RMSE) for each method, i.e. the square root of $R^{-1}\sum_{r=1}^{R}||\hat{\beta}_{(r)}-\beta||^2$, where $\hat{\beta}_{(r)}$ is the vector of regression coefficient estimates in replication r and $R = 1,000$ is the number of replications. Note that the measure is the total distance between two vectors; large differences in individual coefficient estimates may thus be masked.

In all tables a number of estimates is computed with a variety of choices of input parameters. There are entries for the new procedure using KNN estimation (labeled 'Pinkse'), the Cragg–procedure using an increasing number of instruments as suggested in section 2 (labeled 'Cragg') and the Robinson and Harvey estimators where the shadings in the tables indicate which entries in the tables correspond to which estimators. The OLS estimator is a special case of all estimators and was given its own shading. The (top)heading in each table indicates the model type and model parameters, including the number of observations n, the number of regressors d_x and the number of regressors d_r that enter the conditional variance function; these are always taken to be the first d_r regressors and always include a constant. To make it easier to compare numbers across tables with different numbers of observations, the error variance is multiplied by $n/100$, such that if the RMSE decreases proportional to $\sqrt{n}$ the entries in the table should remain constant. The regressors always include a constant and both the remaining regressors and errors were drawn from normal distributions, unless otherwise noted.

Every column in each table corresponds to a different choice of the input parameter d_h for each estimator, where 'small' means $d_h = [(d_x+4)/3]$ and 'large' means $d_h = [(2d_x+2)/3]$, where $[\cdot]$ denotes the largest integer less than or equal to $\cdot$. For the new and Robinson estimators the number of neighbors is chosen either 'small' ($k_n \approx n^{4/5}$) or 'large' ($k_n \approx n^{9/10}$), except in the cases in which binary regressors were used; see the description of those tables below. The first row of entries in each table shows the results for the infeasible version of each estimator. For the Cragg estimator the number of instruments used is $d_h\big(d_x-(d_h-1)/2\big)$, for the d_h–value indicated at the top of the column. The instruments are all interactions between the first d_h regressors and all other regressors (including squares of nonconstant and nonbinary regressors).

Tables 1 and 2 correspond to the case of homoskedasticity, so all estimators have the same asymptotic efficiency. Still, the OLS estimator is expected to perform better than the others since

it does not have the same 'overhead;' it does not require the estimation of first stage estimates. The OLS estimator indeed does best but all estimators are pretty close.

Tables 3–6 report the case of Harvey's (1976) model, which uses the conditional variance function

$$v(x) = e^{2x'\alpha},$$

where α is a vector of unknown parameters. The results in tables 3–6 correspond to the case where the α–parameters are chosen to equal one for $j = 1, \ldots, d_r$. As could be expected the Harvey estimator does usually (but not always!) best when $d_h = d_r$. My estimator does worse than Robinson's for $d_h = 2$ when $d_r = 2$, which is due to the additional overhead involved with my estimator. When $d_h = 2$ and $d_r = 3$ my estimator typically does better. Both nonparametric correction methods, however, perform far below what the results for the corresponding infeasible estimators would suggest. Although A3 is violated by the current choice of v, it appears more likely that first stage estimation error is the culprit here. The discrepancy does become smaller as the sample size increases, but not as fast as one would have hoped. The results of tables 3–6 also suggest that the choice of k_n is not of paramount importance here.

There is some evidence of a curse of dimensionality in tables 4 and 6. Even when d_r is large, choosing a small d_h typically does better than choosing $d_h = d_x$. The performance of the full Robinson correction is however much better than curse of dimensionality arguments would suggest. This is due to two reasons: in view of the discrepancy between the performance of the infeasible and feasible versions of my estimator it appears that the nonparametric weight matrices are poorly estimated and the Robinson estimator effectively reverts to the OLS estimator when the conditional variance function is poorly estimated; it simply 'flattens out' the conditional variance function.

In tables 7–10 the conditional variance is 0.01 if the first $d_r - 1$ slope regressors have values less than -1 and 1 otherwise. The new estimator clearly outperforms the full nonparametric correction when $d_r = 2$, but fails to do so when $d_r = 3$. However, it outperforms the Robinson estimator with too small a choice of d_h in that case.

To show that OLS can beat the reduced Robinson estimator even in large samples I created an experiment designed to beat the Robinson estimator. The results are tabulated in table 11. In the experiment all regressors are binary and the nonparametric corrections use this fact.[8] The conditional variances were chosen specifically to make the reduced Robinson estimator look bad and in this scenario it indeed performs poorly; worse than the OLS estimator.

[8] This is done to reduce the computational burden.

The discrepancy in the performance between the infeasible and feasible versions of the new estimator is substantial and is so for any reasonable choice of k_n, which suggests that the A–matrices are poorly estimated. This is most likely due to the fact that $\hat{T}_i^{-1}$. A better procedure for estimating the A–matrices would thus be valuable.

Finally, the simulation results suggest that for the purpose of nonparametric heteroskedasticity correction, estimating optimal instruments works better than adding instruments.

5 Extensions

There are several issues that were not addressed in this paper. First, there may be multiple reasonable choices of h_i. If in doubt regarding the best choice of h_i, one can simply try multiple ones and make a comparison of the estimated asymptotic variances. Such a procedure is acceptable as long as one choice of h_i strictly dominates other choices in terms of the asymptotic variance. An alternative possibility would be to form a linear combination of estimates $\hat{\theta}_{(1)}, \ldots, \hat{\theta}_{(R)}$ which are based on different choices of h_i. If the $\hat{\theta}_{(r)}$'s are stacked into a long vector $\tilde{\theta}$, then the optimal linear combination is given by

$$\Lambda'\tilde{\theta}, \text{ where } \Lambda = \tilde{V}^{-1}\aleph(\aleph'\tilde{V}^{-1}\aleph)^{-1},$$

with $\tilde{V}$ the asymptotic variance matrix of $\tilde{\theta}$ and $\aleph = [I \cdots I]'$. The asymptotic variance of the linear combination would then be

$$(\aleph'\tilde{V}^{-1}\aleph)^{-1},$$

which, however, is still not guaranteed to achieve the same asymptotic efficiency as the full nonparametric correction.

A second issue concerns the discrepancy in performs between the infeasible and feasible versions of my estimator. One may try alternatives to the nearest neighbor estimator. A more promising avenue, however, would be to define $\hat{\theta}^*$ as a solution to

$$\sum_{i=1}^{n} \hat{A}_i(\hat{\theta}^*) x_i x_i' \hat{\theta}^* = \sum_{i=1}^{n} \hat{A}_i(\hat{\theta}^*) x_i y_i,$$

where $\hat{A}_i(\theta) = H_i \hat{S}_i \hat{T}_i^{-1}(\theta) H_i^+$ and $\hat{T}_i(\theta)$ is a nonparametric estimator of $E\big((y_i - x_i'\theta)^2 t_i t_i' | h_i\big)$.

A further possibility is to condition on an index of regressors instead of a subvector. This would be appealing since it would then be possible to always do at least as well as parametric GLS by conditioning on the hypothesized function. Due to the nondifferentiability of nearest neighbor

estimators using an index would be challenging with the current methodology. But there are alternative nonparametric estimators, which would be more tractable for this purpose.

Finally, the current methodology can be extended to models with endogenous regressors and ones involving nonlinearity. A linear instrumental variables version of the current estimator is a trivial extension, but nonlinearity can be tricky since it does usually not afford the simplification used here which allowed me to avoid the use of estimates of Moore–Penrose inverses.

6 Conclusions

In this paper I propose an estimator for the regression coefficients in a linear regression model in the presence of heteroskedasticity of unknown form when there is some prior knowledge about the most important variables entering the asymptotic variance function. The new estimator is asymptotically the most efficient estimator in its class and the expectation was that due to the curse of dimensionality problem the estimator would outperform a full nonparametric correction in finite samples. The simulation results suggest that it can indeed outperform the full nonparametric correction, although its performance falls short of what its theoretical properties would suggest. Nevertheless, if the number of regressors potentially entering the conditional variance function is large then the proposed procedure is worth using.

References Cited

Aitken, A. (1935), "On least squares and linear combination of observations," *Proceedings of the Royal Society of Edinburgh* 55, 42–48.

Amit, R., J. Brander, K. Hendricks and D. Whistler (1998), "The engine of growth hypothesis: on the relationship between firm size and employment growth," *UBC working paper.*

Eicker, F. (1963), "Asymptotic normality and consistency of the least squares estimator for families of linear regressions," *Annals of Mathematical Statistics* 34, 447–456.

Carroll, R. (1982), "Adapting for heteroscedasticity in linear models," *Annals of Statistics* 10, 1224–1233.

Cragg, J. (1983), "More efficient estimation in the presence of heteroscedasticity of unknown form," *Econometrica* 51, 751–764.

Cragg, J. (1992), "Quasi–Aitken estimation for heteroskedasticity of unknown form," *Journal of Econometrics* 54, 179–202.

Delgado, M. (1992), "Semiparametric generalized least squares estimation in the multivariate nonlinear regression model," *Econometric Theory* 8, 203–222.

Donald, S., G. Imbens and W. Newey (2003), "Empirical likelihood estimation and consistent tests with conditional moment restrictions," *Journal of Econometrics* 117, 59–93.

Dooley, M.D. and P. Gottschalk (1984), "Earnings inequality among males in the United States: Trends and the effect of labor force growth," *Journal of Political Economy* 92, 59–89.

Dunne, P. and A. Hughes (1994), "Age, size, growth and survival: UK companies in the 1980s," *Journal of Industrial Economics* 42–2, 115–140.

Fix, E. and J. Hodges (1951), "Discriminatory analysis, nonparametric discrimination, consistency properties," Randolf Field, Texas, Project 21–49–004, report 4.

Godfrey, L. (1978), "Testing for multiplicative heteroskedasticity," *Journal of Econometrics* 8, 227–236.

Goldfeld, S. and R. Quandt (1965), "Some tests for homoskedasticity," *Journal of the American Statistical Association* 60, 539–547.

Hall, B. (1987), "The relationship between firm size and firm growth in the US manufacturing sector," *Journal of Industrial Economics* 35, 583–606.

Harvey, A. (1976), "Estimating regression models with multiplicative heteroscedasticity," *Econometrica* 44, 461–465.

Huber, P. (1967), "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics,* volume I, University of California Press: Berkeley, California, 221–233.

Kitamura, Y., Tripathi, G. and H. Ahn (2004), "Empirical Likelihood–based inference in conditional moment restriction models," *Econometrica* 72, 1667–1714.

Li, K. (1984), "Consistency of cross–validated nearest neighbor estimates in nonparametric regression," *Annals of Statistics* 12, 230–240.

Mincer, J. (1970), "The distribution of labor incomes: a survey. With special reference to the human capital approach," *Journal of Economic Literature* 8, 1–26.

Mincer, J. (1974), "Schooling, experience and earnings," Columbia University Press, New York.

Newey, W. (1990), "Efficient instrumental variables estimation of nonlinear models," *Econometrica* 58, 809–837.

Newey, W. (1993), "Efficient estimation of models with conditional moment restrictions," in *Handbook of Statistics* 11, G.S. Maddala, C.R. Rao and H.D. Vinod, eds., North Holland, Amsterdam.

Petrunia, R. (2002), "Start–up conditions and the post–entry experience of new firms," *UBC working paper.*

Robinson, P. (1987), "Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form," *Econometrica* 55, 875–891.

Robinson, P. (1991), "Best Nonlinear Three-Stage Least Squares Estimation of Certain Econometric Models," *Econometrica* 59, 755–786.

Romer, P. (1989), "Capital accumulation in the theory of long–run growth," in *Modern business cycle theory,* R. Barro ed., Harvard University Press, Cambridge Massachusetts, 51–127.

Stone, C. (1977), "Consistent nonparametric regression," *Annals of Statistics* 5, 595–645.

White, H. (1980), "A heteroskedasticity–consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica* 48, 817–838.

A Examples

E3 *This example shows how* $E(u_1^2 x_1 x_1') = E(u_1^2)E(x_1 x_1')$ *can hold in the presence of heteroskedasticity.*

Let $x_i = [1, \tilde{x}_i]'$, where $\tilde{x}_i$ takes the values $0, \pm 1, \pm 2$ with probabilities $p_0 = 3/8, p_{\pm 1} = 1/4, p_{\pm 2} = 1/16$. Suppose further that $E(u_i^2|x_i) = I(|\tilde{x}_i| \neq 1) + tI(|\tilde{x}_i| = 1)$, for some $t \geq 0$. Then

$$Eu_i^2 = \frac{t+1}{2}, \quad E(x_i x_i') = I, \quad E(u_i^2 x_i x_i') = \frac{t+1}{2} I, \quad E\Big(\frac{x_i x_i'}{E(u_i^2|x_i)}\Big) = \frac{t+1}{2t} I.$$

Hence $E(u_i^2 x_i x_i') = Eu_i^2 E(x_i x_i')$. The OLS and efficient GLS asymptotic variance matrices are $\big((t+1)/2\big)I$ and $\big(2t/(t+1)\big)I$, respectively. The difference between the OLS and GLS variance matrices is hence $\big\{(t-1)^2/\big(2(t+1)\big)\big\}I$, which only equals zero if $t = 1$, i.e. when there is homoskedasticity. Note also that the ratio of the variances, i.e. $(t+1)^2/(4t)$ can be arbitrarily large if t is taken to ∞ or 0. □

E4 *This example also shows how OLS can be more efficient than reduced Robinson.*

Let $x_i = [h_i', z_i']'$ with $h_i = [1, \tilde{h}_i]'$ where $\tilde{h}_i$ is Bernoulli and $z_i \sim N(0, I)$. Suppose further that $E(u_i^2|h_i, z_i) = \tilde{h}_i + (1 - \tilde{h}_i)t\psi(z_i)$, with $E\big(\psi(z_i)z_i\big) = 0$ and $E\psi^2(z_i) = 1$. Then the OLS and Robinson variances of $\hat{\beta}_h$ are the same and equal

$$\begin{bmatrix} \frac{t}{1-p^*} & -\frac{t}{1-p^*} \\ \frac{-t}{1-p^*} & \frac{t}{1-p^*} + \frac{1}{p^*} \end{bmatrix}.$$

The OLS and Robinson variances of $\hat{\beta}_z$ are not the same (note that $\text{Cov}(\hat{\beta}_z, \hat{\beta}_h) = 0$ in both cases), but are ($\Omega = E\big(\psi(z_i)z_i z_i'\big)$)

$$p^* I + t(1-p^*)\Omega \text{ and } \frac{t^2 p^* I + t(1-p^*)\Omega}{\big((t-1)p^*+1\big)^2},$$

respectively. For $t = 1$ the two are the same since then Robinson=OLS. The difference between the Robinson and OLS variance matrices for $\hat{\beta}_z$ is

$$\frac{(t-1)p^*(1-p^*)\Big(\big\{(1-p^*)+t(1+p^*)\big\}I - t\big\{(t-1)p^*+2\big\}\Omega\Big)}{\big\{(t-1)p^*+1\big\}^2}.$$

Now suppose that $\psi(z_i) = (\sum_{j=1}^{d_z} z_{ij}^a)/(dEz_{ij}^a)$ for some even integer $a \geq 2$. Then $\Omega = (a/d+1)I$. For some $t < 1$, choose $a = (1-t)d/t$, such that $\Omega = t^{-1}I$, the OLS variance is I and the Robinson variance is

$$\frac{t^2 p^* + (1-p^*)}{\big((t-1)p^*+1\big)^2} = \frac{1-(1-t)p^*(1+t)}{1-(1-t)p^*(1+t)-(1-t)p^*\big(1-(1-t)p^*\big)} \geq 1.$$

The Robinson variance goes to $1/(1-p^*)$ as $t \to 0$, which goes to ∞ as $p^* \to 1$.

Alternatively, one can pick $p^* = 1/(t+1)$ such that the Robinson asymptotic variance is $(t+1)^2/(4t)$. For a fixed choice of a, the Robinson asymptotic variance is then $(a+2d_z)^2/\big(4d_z(a+d_z)\big)$. To get an asymptotic variance of $v^* > 1$, one should pick $a = 2d(v^*-1)\big(1+\sqrt{v^*/(v^*-1)}\big)$. □

B A, L, Q, S, T, M

L1 *(i)* $L_i = H_i S_i H_i'$, *(ii)* $Q_i = H_i T_i H_i'$, *(iii)* $Q_i^+ = (H_i')^+ T_i^{-1} H_i^+$, *(iv)* $A_i = H_i M_i H_i^+$, *(v)* $A_i x_i = H_i M_i t_i$. **Proof:** First (i). I have

$$\begin{aligned} L_i = E(x_i x_i'|h_i) &= E\left(\begin{bmatrix} h_i h_i' & h_i b_i' \\ b_i h_i' & b_i b_i' \end{bmatrix}|h_i\right) \\ &= \begin{bmatrix} h_i & 0 \\ 0 & I \end{bmatrix} E\left(\begin{bmatrix} 1 & b_i' \\ b_i & b_i b_i' \end{bmatrix}|h_i\right)\begin{bmatrix} h_i' & 0 \\ 0 & I \end{bmatrix} = H_i E(t_i t_i'|h_i) H_i' = H_i S_i H_i'. \end{aligned}$$

(ii) follows similarly. Now (iii). Since

$$H_i^+ = \begin{bmatrix} \frac{h_i'}{||h_i||^2} & 0 \\ 0 & I \end{bmatrix},$$

it's clear that the stated expression of Q_i^+ satisfies the four requirements of a Moore Penrose inverse. (iv) then follows from the fact that $H_i^+ H_i = I$ and the definition of M_i. Finally, (v) is implied by the fact that $H_i^+ x_i = t_i$. □

L2 *$A(h) = L(h)Q^+(h)$ is the optimal choice of weight matrix.* **Proof:** Note that for arbitrary matrix–valued weight matrix function $A(h)$, the asymptotic variance of the matrix–weighted least squarest estimator is

$$V_A = \big(E(A_1x_1x_1')\big)^{-1}E(A_1u_1^2x_1x_1'A_1')\big(E(x_1x_1'A_1')\big)^{-1} = \big(E(A_1L_1)\big)^{-1}E(A_1Q_1A_1')\big(E(L_1A_1')\big)^{-1}$$
$$= \big(E(A_1H_1S_1H_1')\big)^{-1}E(A_1H_1T_1H_1'A_1')\big(E(H_1S_1H_1'A_1')\big)^{-1}.$$

For the choice $A_i = L_iQ_i^+ = H_iS_iT_i^{-1}H_i^+$ it is by L1 equal to $V_* = \big(E(H_1S_1T_1^{-1}S_1H_1')\big)^{-1}$. Now let $\boldsymbol{H} = [H_1|H_2|\cdots|H_n] \in \mathbb{R}^{d_x\times nd_t}$, $\boldsymbol{T},\boldsymbol{S} \in \mathbb{R}^{nd_t\times nd_t}$ be block–diagonal matrices with blocks T_i, S_i respectively, and let $\boldsymbol{A} = [A_1'|\cdots|A_n']' \in \mathbb{R}^{nd_x\times d_x}$, such that

$$V_*^{-1} - V_A^{-1} = (\boldsymbol{HST}^{-1}\boldsymbol{S}'\boldsymbol{H}' - \boldsymbol{HSH}'\boldsymbol{A}'(\boldsymbol{AHTH}'\boldsymbol{A}')^{-1}\boldsymbol{AHS}'\boldsymbol{H}')/n + o_p(1). \tag{4}$$

But the first RHS term times n in (4) is

$$\boldsymbol{HST}^{-1}\boldsymbol{S}'\boldsymbol{H}' - \boldsymbol{HSH}'\boldsymbol{A}'(\boldsymbol{AHTH}'\boldsymbol{A}')^{-1}\boldsymbol{AHS}'\boldsymbol{H}' = \boldsymbol{HST}^{-1/2}\Big(I - \boldsymbol{T}^{1/2}\boldsymbol{A}'(\boldsymbol{AHTH}'\boldsymbol{A}')^{-1}\boldsymbol{AT}^{1/2}\Big)\boldsymbol{T}^{-1/2}\boldsymbol{S}'\boldsymbol{H}' \geq 0,$$

since the expression in large brackets is an orthogonal projection matrix. □

C Technical Lemmas

L3 *For any sequence $\{B_i\}$ for which $E(B_i|\mathcal{H}, B_1, \ldots, B_{i-1}) = 0$, then*

$$\forall p \geq 1 : \max_i E||B_i||^p < \infty \Rightarrow E\Big|\Big|\sum_j w_{ij}B_j\Big|\Big|^p = O(k_n^{\max\{-p/2,1-p\}}),$$

Proof: First suppose $p \geq 2$. Note that by the Burkholder (1973) and Jensen inequalities,

$$E\left(\Big|\Big|\sum_j w_{ij}B_j\Big|\Big|^p|\mathcal{H}\right) \leq E\left(\Big(\Big|\Big|\sum_j w_{ij}B_j\Big|\Big|^2\Big)^{p/2}|\mathcal{H}\right) \overset{A2}{\leq} C_w^{p/2}k_n^{-p/2}E\left(\Big(\sum_j w_{ij}||B_j||^2\Big)^{p/2}|\mathcal{H}\right)$$
$$\leq C_w^{p/2}k_n^{-p/2}E\left(\sum_j w_{ij}||B_j||^p|\mathcal{H}\right) \leq C_w^{p/2}k_n^{-p/2}\sum_j w_{ij}E\big(||B_j||^p|\mathcal{H}\big). \tag{5}$$

Take expectations. The case $1 \leq p \leq 2$ follows from the Von Bahr–Esseen (1965) inequality since

$$E\left(\Big|\Big|\sum_j w_{ij}B_j\Big|\Big|^p|\mathcal{H}\right) \leq 2\sum_j E\left(\Big|\Big|w_{ij}B_j\Big|\Big|^p|\mathcal{H}\right) \leq 2C_w^{p-1}k_n^{1-p}\sum_j w_{ij}E\big(||B_j||^p|\mathcal{H}\big). \tag{6}$$

Take expectations. □

L4 *Let Ψ_{nij} be matrix–valued and such that (i) for all $i \neq j$: $E(\Psi_{nij}|X, y_i) = E(\Psi_{nij}|X, y_j) = 0$ a.s., (ii) $\max_i E||\Psi_{nii}|| = o(n^{-1})$ and $\max_{i\neq j} E||\Psi_{nij}||^2 = o(n^{-2})$, then (iii) $\sum_{i,j=1}^n \Psi_{nij} = o_p(1)$.*
Proof: Note that

$$\sum_{i,j=1}^n \Psi_{nij} = \sum_{i=1}^n \Psi_{nii} + \sum_{i=1}^n\sum_{j\neq i}^n \Psi_{nij}. \tag{7}$$

The first RHS term in (7) is $o_p(1)$ by requirement (ii). The second RHS term is $o(1)$, also, since

$$E\left\{\left(\sum_{i=1}^n\sum_{j\neq i}^n \Psi'_{nij}\right)\left(\sum_{i=1}^n\sum_{j\neq i}^n \Psi_{nij}\right)\right\} \overset{\text{req.(i)}}{=} \sum_{i=1}^n\sum_{j\neq i}^n \big(E(\Psi'_{nij}\Psi_{nij}) + E(\Psi'_{nij}\Psi_{nji})\big),$$

such that

$$E\Big|\Big|\sum_{i=1}^n\sum_{j\neq i}^n \Psi_{nij}\Big|\Big|^2 \leq 2n^2 \max_{i\neq j} E||\Psi_{nij}||^2 \overset{\text{req.(iii)}}{=} o(1). \quad \square$$

L5 *Let $\Xi_1,\ldots,\Xi_J$ be random matrices for which $E||\Xi_j||^{\pi_j} < \infty$ for some $\pi_1,\ldots,\pi_J > 0$, and $\prod_{j=1}^J \Xi_j$ is well–defined. Then for any $\pi_1^*,\ldots,\pi_J^* > 0$,*

$$\sum_{j=1}^J \frac{\pi_j^*}{\pi_j} \leq 1 \Rightarrow E\prod_{j=1}^J ||\Xi_j||^{\pi_j^*} \leq \prod_{j=1}^J \big(E||\Xi_j||^{\pi_j}\big)^{\pi_j^*/\pi_j} < \infty.$$

Proof: Let $\delta_j = \pi_j^*/\pi_j$. Then by repeated use of the Hölder inequality,

$$\begin{aligned}
E\Big|\Big|\prod_{j=1}^J ||\Xi_j||^{\pi_j^*}\Big|\Big| &\leq \Big(E\prod_{j=1}^{J-1} ||\Xi_j||^{\frac{\pi_j^*}{1-\delta_J}}\Big)^{1-\delta_J}\Big(E||\Xi_J||^{\pi_J}\Big)^{\delta_J} \\
&\leq \Big(E\prod_{j=1}^{J-2} ||\Xi_j||^{\frac{\pi_j^*}{1-\delta_{J-1}-\delta_J}}\Big)^{1-\delta_{J-1}-\delta_J}\Big(E||\Xi_{J-1}||^{\pi_{J-1}}\Big)^{\delta_{J-1}}\Big(E||\Xi_J||^{\pi_J}\Big)^{\delta_J} \leq \cdots \\
&\leq \Big(E||\Xi_1||^{\frac{\pi_1^*}{1-\delta_2-\ldots-\delta_J}}\Big)^{1-\delta_2-\ldots-\delta_J}\prod_{j=2}^J\Big(E||\Xi_j||^{\pi_j}\Big)^{\delta_j} \leq \Big(E||\Xi_1||^{\pi_1}\Big)^{\delta_1}\prod_{j=2}^J\Big(E||\Xi_j||^{\pi_j}\Big)^{\delta_j},
\end{aligned}$$

provided that $\delta_1/(1-\delta_2-\ldots-\delta_J) \leq 1$, i.e. that $\sum_{j=1}^J \delta_j \leq 1$, which was assumed. $\square$

D Nonparametric Approximation

In the remainder μ_{ni} is one of $u_i/\sqrt{n}$ and x_i/n.

D.1 $\bar{A} - A$

L6 $\exists\epsilon > 0 : P\big(\min_i \lambda_{\min}(\bar{T}_i) < \epsilon\big) = 0$. **Proof:** $P\big(\min_i \lambda_{\min}(\bar{T}_i) < \epsilon\big) \leq P\big(\min_i \lambda_{\min}(T_i) < \epsilon\big) = 0$. $\square$

L7 *(i) $M_i \leq \varsigma^{-1}I$ a.s., (ii) $\bar{M}_i \leq \varsigma^{-1}I$ a.s..* **Proof:** Note that by A3, $T_i = E\big(v(x_i)t_it_i'|h_i\big) \geq \varsigma S_i$, which gives (i). (ii) follows similarly since $\bar{T}_i = \sum_j w_{ij}T_j \geq \varsigma\sum_j w_{ij}S_j = \varsigma\bar{S}$. $\square$

L8 *(i) $E||\bar{T}_i - T_i||^{p_t} = o(1)$, (ii) $E||\bar{S}_i - S_i||^{p_S} = o(1)$.*
Proof: First (i), $E||\bar{T}_i - T_i||^{p_t} = E\big|\big|\sum_j w_{ij}(T_j - T_i)\big|\big|^{p_t} \leq E\big(\sum_j w_{ij}||T_j - T_i||^{p_t}\big) = o(1)$, by Robinson (1987), lemma 1. The proof for (ii) is similar. $\square$

L9 *For all $p > 0$, $E||\bar{T}_i^{-1} - T_i^{-1}||^p = o(1)$.* **Proof:** Let ε_n be such that $||\bar{T}_i - T_i||/\varepsilon_n = o_p(1)$ and $\varepsilon_n \to 0$; such ε_n exist by L8. Now, for any $p > 0$,

$$E||\bar{T}_i^{-1} - T_i^{-1}||^p = E\big(||\bar{T}_i^{-1} - T_i^{-1}||^p I(||\bar{T}_i - T_i|| > \varepsilon_n)\big) + E\big(||\bar{T}_i^{-1} - T_i^{-1}||^p I(||\bar{T}_i - T_i|| \leq \varepsilon_n)\big)$$
$$\overset{L6}{\leq} \epsilon^{-p} P\big(||\bar{T}_i - T_i|| > \varepsilon_n\big) + E\big(||\bar{T}_i^{-1}(T_i - \bar{T}_i)T_i^{-1}||^p (||\bar{T}_i - T_i|| \leq \varepsilon_n)\big) \overset{L6,L8}{\leq} \epsilon^{-p} o(1) + \epsilon^{-2p}\varepsilon_n^{2p} = o(1). \quad \square$$

L10 *$E||\bar{M}_i - M_i||^{p_S} = o(1)$.*

Proof: Let ε_n be such that $\varepsilon_n \to 0$ and $||\bar{T}_i^{-1} - T_i^{-1}||/\varepsilon_n = o_p(1)$, which is possible by L9. Using the expansion $\bar{M}_i - M_i = (\bar{S}_i - S_i)\bar{T}_i^{-1} + S_i(\bar{T}_i^{-1} - T_i^{-1})$, write

$$E||\bar{M}_i - M_i||^{p_S} = E\big(||\bar{M}_i - M_i||^{p_S} I(||\bar{T}_i^{-1} - T_i^{-1}|| \leq \varepsilon_n)\big) + E\big(||\bar{M}_i - M_i||^{p_S} I(||\bar{T}_i^{-1} - T_i^{-1}|| > \varepsilon_n)\big)$$
$$\overset{L7}{\leq} E\big(||(\bar{S}_i - S_i)\bar{T}_i^{-1} + S_i(\bar{T}_i^{-1} - T_i^{-1})||^{p_S} I(||\bar{T}_i^{-1} - T_i^{-1}|| \leq \varepsilon_n)\big) + 2\varsigma^{-p_S} P\big(||\bar{T}_i^{-1} - T_i^{-1}|| > \varepsilon_n\big)$$
$$\overset{L6,L9}{\leq} E\big(\epsilon^{-1}||\bar{S}_i - S_i|| + ||S_i||\varepsilon_n\big)^{p_S} + o(1) \overset{L8}{=} o(1). \quad \square$$

L11 $\sum_{i=1}^n H_i(\bar{M}_i - M_i)t_i\mu_{ni}' = o_p(1)$. **Proof:** Note that

$$E\Big|\Big|\sum_{i=1}^n H_i(\bar{M}_i - M_i)t_i\mu_{ni}'\Big|\Big|^2 = \sum_{i=1}^n E\big||H_i(\bar{M}_i - M_i)t_i\mu_{ni}'\big|\big|^2$$
$$\overset{L5}{\leq} n\big(E||H_i||^{p_h}\big)^{2/p_h}\big(E||\bar{M}_i - M_i||^{p_S}\big)^{2/p_S}\big(E||t_i||^{p_t}\big)^{2/p_t}\big(E||\mu_{ni}||^{p_\mu}\big)^{2/p_\mu} \overset{L10}{=} o(1),$$

since $2/p_h + 2/p_S + 2/p_t + 2/p_\mu \leq 1$ by A1. $\square$

D.2 $\hat{A} - \bar{A}$

I now break up the difference $\hat{M}_i - \bar{M}_i$ into two parts by

$$\hat{M}_i - \bar{M}_i = \big((\tilde{S}_i - \bar{S}_i) + \bar{M}_i(\bar{T}_i - \hat{T}_i)\big)\big(\hat{T}_i^{-1} - \bar{T}_i^{-1}\big) + \big(\bar{M}_i(\bar{T}_i - \tilde{T}_i) + (\tilde{S}_i - \bar{S}_i)\big)\bar{T}_i^{-1} + \bar{M}_i(\tilde{T}_i - \hat{T}_i)\bar{T}_i^{-1} = \Delta_{1i} + \Delta_{2i} + \Delta_{3i}. \tag{8}$$

D.2.1 Sums involving Δ_{1i}

L12 *(i) $E||\tilde{T}_i - \bar{T}_i||^{p_t} = O(k_n^{-p_t/2})$, (ii) $E||\tilde{S}_i - \bar{S}_i||^{p_S} = O(k_n^{-p_S/2})$.* **Proof:** I show (i); (ii) is similar. Since $E(u_j^2 t_j t_j' - T_j|\mathcal{H}) = 0$ a.s., $E||\tilde{T}_i - \bar{T}_i||^{p_t} = E\big|\big|\sum_j w_{ij}(u_j^2 t_j t_j' - T_j)\big|\big|^{p_t} \overset{L3}{=} O(k_n^{-p_t/2})$. $\square$

L13 *(i) $\max_i ||\tilde{T}_i - \bar{T}_i|| = O_p(n^{1/p_t}k_n^{-1/2})$ and (ii) $\max_i ||\tilde{S}_i - \bar{S}_i|| = O_p(n^{1/p_S}k_n^{-1/2})$.* **Proof:** I show (i) where (ii) follows identically. For any $\epsilon_T > 0$,

$$P\big(n^{-1/p_t}k_n^{1/2}\max_i ||\tilde{T}_i - \bar{T}_i|| > \epsilon_T\big) \leq \sum_{i=1}^n P\Big(n^{-1/p_t}k_n^{1/2}||\tilde{T}_i - \bar{T}_i|| > \epsilon_T\Big)$$
$$\overset{\text{Markov}}{\leq} n^{-1}k_n^{p_t/2}\epsilon_T^{-p_t}\sum_{i=1}^n E||\tilde{T}_i - \bar{T}_i||^{p_t} \overset{L12}{=} O(1). \quad \square$$

L14 $\max_i ||\hat{T}_i - \tilde{T}_i|| = O_p(n^{1/2}k_n^{-1})$, **Proof:** I have

$$\begin{aligned}
\max_i ||\hat{T}_i - \tilde{T}_i|| &= \max_i \Big|\Big|\sum_{j=1}^n w_{ij}(\hat{u}_j^2 - u_j^2)t_j t_j'\Big|\Big| \\
&\leq \max_i \Big|\Big|\sum_{j=1}^n w_{ij}(\hat{u}_j - u_j)^2 t_j t_j'\Big|\Big| + 2\max_i \Big|\Big|\sum_{j=1}^n w_{ij}u_j(\hat{u}_j - u_j)t_j t_j'\Big|\Big| \\
&\leq C_w k_n^{-1}||\hat{\theta} - \theta_0||^2 \sum_{j=1}^n ||t_j||^2 \cdot \Big|\Big|\frac{\partial g_j}{\partial \theta'}(\theta^*)\Big|\Big|^2 + C_w k_n^{-1}||\hat{\theta} - \theta_0|| \sum_{j=1}^n \Big|\Big|\frac{\partial g_j}{\partial \theta'}(\theta^*)\Big|\Big| |u_j| \cdot ||t_j||^2 \\
&\leq C_w k_n^{-1}||\hat{\theta} - \theta_0||^2 \max_j ||t_j||^2 \sum_{j=1}^n \Big|\Big|\frac{\partial g_j}{\partial \theta'}(\theta^*)\Big|\Big|^2 + C_w k_n^{-1}||\hat{\theta} - \theta_0|| \left(\sum_{j=1}^n \Big|\Big|\frac{\partial g_j}{\partial \theta'}(\theta^*)\Big|\Big|^2 \sum_{j=1}^n |u_j|^2 \cdot ||t_j||^4\right)^{1/2} \\
&= O_p(k_n^{-1}n^{-1}n^{2/p_t}n) + O_p(k_n^{-1}n^{-1/2}n) = O_p(k_n^{-1}n^{1/2}),
\end{aligned}$$

by A1. □

L15 $\max_i ||\hat{T}_i^{-1} - \bar{T}_i^{-1}|| = O_p(n^{1/p_t}k_n^{-1/2} + n^{1/2}k_n^{-1})$. **Proof:** Write $\hat{T}_i^{-1} - \bar{T}_i^{-1} = \bar{T}_i^{-1}(\bar{T}_i - \hat{T}_i)\bar{T}_i^{-1}\big(I - (\bar{T}_i - \hat{T}_i)\bar{T}_i^{-1}\big)^{-1}$, and use L6, L13 and L14. □

L16 $\sum_{i=1}^n H_i\Delta_{1i}t_i\mu_{ni}' = o_p(1)$. **Proof:** Note first that

$$\begin{aligned}
\max_i ||\Delta_{1i}|| &\leq \big(\max_i ||\tilde{S}_i - \bar{S}_i|| + \max_i ||\bar{M}_i|| \max_i ||\hat{T}_i - \bar{T}_i||\big) \max_i ||\hat{T}_i^{-1} - \bar{T}_i^{-1}|| \overset{\text{L7,L13,L14,L15}}{=} O_p\big(n^{2/p_t}k_n^{-1} + nk_n^{-2}\big), \\
E\big(||H_i|| \cdot ||t_i|| \cdot ||\mu_{ni}||\big) &\leq n^{-1/2}E\big(||H_i|| \cdot ||t_i|| \cdot (||u_i|| + ||x_i||)\big) = O(n^{-1/2}),
\end{aligned}$$

by L5 since $1/p_h + 1/p_t + \max\{1/p_u + 1/p_x\} \leq 1$ by A1. Thus,

$$\begin{aligned}
\Big|\Big|\sum_{i=1}^n H_i\Delta_{1i}t_i\mu_{ni}'\Big|\Big| &\leq \max_i ||\Delta_{1i}|| \sum_{i=1}^n ||H_i|| \cdot ||t_i|| \cdot ||\mu_{ni}|| \\
&= O_p\Big(\big(n^{2/p_t}k_n^{-1} + nk_n^{-2}\big)n^{1/2}\Big) = O_p\big(n^{2/p_t+1/2}k_n^{-1} + n^{3/2}k_n^{-2}\big) \overset{\text{A2}}{=} o_p(1). \quad \square
\end{aligned}$$

D.2.2 Sums involving Δ_{2i}

L17 $n^{-1}\sum_{i=1}^n H_i\Delta_{2i}t_i x_i' = o_p(1)$. **Proof:** Note that

$$\begin{aligned}
E\Big|\Big|\sum_{i=1}^n H_i\Delta_{2i}t_i x_i'/n\Big|\Big| &\leq E\big|\big|H_i\Delta_{2i}t_i x_i'\big|\big| \leq E\Big(||H_i|| \cdot ||\bar{T}_i^{-1}|| \cdot ||t_i|| \cdot ||x_i|| \cdot \big(||\bar{M}_i|| + 1\big)\big(||\tilde{S}_i - \bar{S}_i|| + ||\tilde{T}_i - \bar{T}_i||\big)\Big) \\
&\overset{\text{L6,L7}}{\leq} \frac{\epsilon + 1}{\epsilon\varsigma}E\Big(||H_i|| \cdot ||t_i|| \cdot ||x_i||\big(||\tilde{S}_i - \bar{S}_i|| + ||\tilde{T}_i - \bar{T}_i||\big)\Big) \\
&\overset{\text{Schwarz}}{\leq} 2\frac{\epsilon + 1}{\epsilon\varsigma}\Big(E\big(||H_i||^2||t_i||^2||x_i||^2\big)\Big)^{1/2}\big(E||\tilde{S}_i - \bar{S}_i||^2 + E||\tilde{T}_i - \bar{T}_i||^2\big)^{1/2} \quad (9)
\end{aligned}$$

The first RHS factor in (9) is finite by L5 since $2/p_h + 2/p_t + 2/p_x \leq 1$ by A1 and the second RHS factor is $O(k_n^{-1}) = o(1)$ by L12.

L18 $n^{-1/2}\sum_{i=1}^n H_i\Delta_{2i}t_iu_i = o_p(1)$. **Proof:** Take $\Psi_{nij} = n^{-1/2}w_{ij}H_i\big(\bar{M}_i(T_j - u_j^2t_jt_j') + (t_jt_j' - S_j)\big)\bar{T}_i^{-1}t_iu_i$ in L4, such that $n^{-1/2}\sum_{i=1}^n H_i\Delta_{2i}t_iu_i = \sum_{i,j=1}^n \Psi_{nij}$. I now verify the conditions of L4: (i) is clearly true. For (ii), note that

$$\max_i E||\Psi_{nii}|| \overset{\text{A2,L6,L7}}{\leq} 4C_w n^{-1/2}k_n^{-1}\varsigma^{-1}(\varepsilon^{-1}+1)\max_i E\Big(||H_i||\big(u_i^2||t_i||^2+||t_i||^2\big)||t_i||^2|u_i|\Big) = O(n^{-1/2}k_n^{-1}) \overset{\text{A2}}{=} o(n^{-1}),$$

since the RHS expectation is bounded by L5 because $1/p_h + 3/p_u + 4/p_t \leq 1$ by A1. □

D.2.3 Sums involving Δ_{3i}

L19 $\sum_{i=1}^n ||\hat{T}_i - \tilde{T}_i||^2 = o_p(1)$. **Proof:** Noting that $\hat{u}_j^2 - u_j^2 = (\hat{u}_j - u_j)^2 + 2u_j(\hat{u}_j - u_j) = \big(x_j'(\theta_0 - \hat{\theta})\big)^2 + 2u_jx_j'(\theta_0 - \hat{\theta})$,

$$\begin{aligned}\sum_{i=1}^n ||\hat{T}_i - \tilde{T}_i||^2 &= \sum_{i=1}^n \left|\left|\sum_{j=1}^n w_{ij}\Big\{\big(x_j'(\theta_0 - \hat{\theta})\big)^2 + 2u_jx_j'(\theta_0 - \hat{\theta})\Big\}t_jt_j'\right|\right|^2 \\ &\overset{\text{Minkowski}}{\leq} 2\sum_{i=1}^n \left|\left|\sum_{j=1}^n w_{ij}\Big\{\big(x_j'(\theta_0 - \hat{\theta})\big)^2 t_jt_j'\Big\}\right|\right|^2 + 8\sum_{i=1}^n \left|\left|\sum_{j=1}^n w_{ij}\Big\{\big(u_jx_j'(\theta_0 - \hat{\theta})\big)t_jt_j'\Big\}\right|\right|^2. \quad (10)\end{aligned}$$

The first RHS term in (10) is bounded by

$$2||\hat{\theta} - \theta_0||^4\sum_{i=1}^n\Big|\Big|\sum_{j=1}^n w_{ij}||x_j||^2||t_j||^2\Big|\Big|^2 \overset{\text{A2}}{=} O_p(n^{-1}k_n^{-2})\Big|\Big|\sum_{j=1}^n ||x_j||^2||t_j||^2\Big|\Big|^2 = O_p(nk_n^{-2}) \overset{\text{A2}}{=} o_p(1).$$

Now the second RHS term in (10). It is bounded by

$$8||\hat{\theta} - \theta_0||^2\sum_{i=1}^n\Big|\Big|\sum_{j=1}^n w_{ij}(u_jx_j \otimes t_jt_j')\Big|\Big|^2 \overset{\text{L3}}{=} O_p(n^{-1}\times n\times k_n^{-2}) = O_p(k_n^{-2}) = o_p(1). \quad \square$$

L20 $\big|\big|\sum_{i=1}^n H_i\Delta_{3i}t_i\mu_{ni}'\big|\big| = o_p(1)$. **Proof:** Note that

$$\begin{aligned}\Big|\Big|\sum_{i=1}^n H_i\bar{M}_i(\tilde{T}_i - \hat{T}_i)\bar{T}_i^{-1}t_i\mu_{ni}'\Big|\Big|^2 &\leq \Big(\sum_{i=1}^n ||H_i||\cdot||\bar{M}_i||\cdot||\bar{T}_i^{-1}||\cdot||t_i\mu_{ni}'||\cdot||\tilde{T}_i - \hat{T}_i||\Big)^2 \\ &\overset{\text{Schwarz}}{\leq} \Big(\sum_{i=1}^n |H_i||^2||\bar{M}_i||^2||\bar{T}_i^{-1}||^2||t_i||^2||\mu_{ni}||^2\Big)^{1/2}\Big(\sum_{i=1}^n ||\tilde{T}_i - \hat{T}_i||^2\Big)^{1/2}. \quad (11)\end{aligned}$$

The second RHS factor in (11) is $o_p(1)$ by L19. The square of the first RHS factor is by L6 and L7 bounded by

$$\varsigma^{-2}\epsilon^{-2}\sum_{i=1}^n ||H_i||^2||t_i||^2||\mu_{ni}||^2 = O(1),$$

because $nE\big(||H_1||^2||t_1||^2||\mu_{n1}||^2\big) \leq E\big(||H_1||^2||t_1||^2(u_1^2+||x_1||^2)\big) \overset{\text{L5}}{<} \infty$ since $2/p_h+2/p_t+\max\{2/p_u, 2/p_x\} \leq 1$ by A1. □

E Main Result

Proof of T1: I need to show that

$$\left(n^{-1}\sum_{i=1}^{n}\hat{A}_i x_i x_i'\right)^{-1} n^{-1/2}\sum_{i=1}^{n}\hat{A}_i x_i u_i \xrightarrow{d} N\Big(0,\big(E(L_1 Q_1^{+} L_1)\big)^{+}\Big).$$

For this it suffices to show that

$$\left(n^{-1}\sum_{i=1}^{n}A_i x_i x_i'\right)^{-1} n^{-1/2}\sum_{i=1}^{n}A_i x_i u_i \xrightarrow{d} N\Big(0,\big(E(L_1 Q_1^{+} L_1)\big)^{+}\Big), \tag{12}$$

$$\sum_{i=1}^{n}(\bar{A}_i - A_i)x_i\mu_{ni}' = o_p(1), \tag{13}$$

$$\sum_{i=1}^{n}(\hat{A}_i - \bar{A}_i)x_i\mu_{ni}' = o_p(1). \tag{14}$$

By standard application of the Khinchine, Lindeberg–Levy, Slutsky and Cramér theorems the LHS in (12) has a limiting mean zero normal distribution with variance $\big(E(A_1x_1x_1')\big)^{-1}E(A_1u_1^2x_1x_1'A_1')\big(E(x_1x_1'A_1')\big)^{-1}$. So (12) holds because

$$E(A_1x_1x_1') = EE(A_1x_1x_1'|h_1) = E(A_1L_1) = E(L_1Q_1^{+}L_1),$$
$$E(A_1u_1^2x_1x_1'A_1') = EE(A_1u_1^2x_1x_1'A_1'|h_1) = E(A_1Q_1A_1') = E(L_1Q_1^{+}L_1).$$

Now (13). The LHS is by L1 equal to $\sum_{i=1}^{n}H_i(\bar{M}_i - M_i)t_i\mu_{ni}'$. Apply L11. Finally (14). The LHS can by (8) be written as

$$\sum_{j=1}^{3}\sum_{i=1}^{n}H_i\Delta_{ji}t_i\mu_{ni}'.$$

Apply lemmas L16, L17, L18 and L20. □

F RMSE Tables

Table 1: homoskedastic $d_x = 5$, normal

1a: $n = 100$									**1b:** $n = 200$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	0.224	0.224	0.224	0.224	0.224	0.224	0.224	0.224	0.229	0.229	0.229	0.229	0.229	0.229	0.229	0.229
k_n =small	0.224	0.237	0.235	0.231	0.227	0.226	0.228	0.227	0.229	0.239	0.236	0.233	0.230	0.230	0.231	0.231
k_n =large	0.224	0.230	0.228	0.228	0.225	0.225	0.225	0.225	0.229	0.232	0.232	0.231	0.229	0.229	0.230	0.229
Cragg Harvey	0.224	0.234	0.240	0.241	0.241	0.238	0.234	0.229	0.229	0.235	0.237	0.238	0.240	0.238	0.235	0.232

1c: $n = 500$									**1d:** $n = 1000$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	0.224	0.224	0.224	0.224	0.224	0.224	0.224	0.224	0.227	0.227	0.227	0.227	0.227	0.227	0.227	0.227
k_n =small	0.224	0.228	0.228	0.226	0.224	0.225	0.225	0.224	0.227	0.229	0.228	0.227	0.227	0.227	0.228	0.228
k_n =large	0.224	0.226	0.225	0.224	0.224	0.224	0.224	0.224	0.227	0.228	0.227	0.227	0.227	0.227	0.227	0.227
Cragg Harvey	0.224	0.226	0.228	0.229	0.227	0.227	0.226	0.225	0.227	0.228	0.228	0.229	0.230	0.229	0.228	0.228

Table 2: homoskedastic $d_x = 10$, normal

2a: $n = 100$									**2b:** $n = 200$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	0.336	0.336	0.336	0.336	0.336	0.336	0.336	0.336	0.327	0.327	0.327	0.327	0.327	0.327	0.327	0.327
k_n =small	0.336	0.363	0.361	0.355	0.339	0.339	0.340	0.340	0.327	0.346	0.344	0.339	0.329	0.329	0.329	0.329
k_n =large	0.336	0.348	0.350	0.346	0.337	0.337	0.337	0.337	0.327	0.335	0.335	0.333	0.328	0.328	0.328	0.327
Cragg Harvey	0.336	0.355	0.370	0.372	0.379	0.368	0.355	0.343	0.327	0.340	0.353	0.361	0.356	0.348	0.338	0.331

2c: $n = 500$									**2d:** $n = 1000$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	0.328	0.328	0.328	0.328	0.328	0.328	0.328	0.328	0.322	0.322	0.322	0.322	0.322	0.322	0.322	0.322
k_n =small	0.328	0.341	0.337	0.335	0.329	0.329	0.328	0.330	0.322	0.331	0.327	0.326	0.323	0.323	0.323	0.323
k_n =large	0.328	0.332	0.331	0.331	0.328	0.328	0.328	0.328	0.322	0.325	0.325	0.325	0.322	0.322	0.322	0.322
Cragg Harvey	0.328	0.333	0.341	0.347	0.340	0.336	0.332	0.329	0.322	0.326	0.330	0.334	0.331	0.328	0.325	0.323

Table 3: Harvey $d_x = 5$, $d_r = 2$, normal

3a: $n = 100$									**3b:** $n = 200$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	0.788	0.150	0.150	0.150	0.150	0.150	0.150	0.150	0.855	0.136	0.136	0.136	0.136	0.136	0.136	0.136
k_n =small	0.788	0.275	0.295	0.314	0.326	0.303	0.276	0.243	0.855	0.246	0.264	0.281	0.296	0.274	0.249	0.223
k_n =large	0.788	0.322	0.362	0.399	0.432	0.403	0.365	0.313	0.855	0.290	0.334	0.367	0.401	0.372	0.335	0.284
Cragg Harvey	0.788	0.474	0.461	0.456	0.240	0.234	0.225	0.217	0.855	0.531	0.521	0.516	0.214	0.210	0.203	0.198

3c: $n = 500$									**3d:** $n = 1000$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	0.790	0.133	0.133	0.133	0.133	0.133	0.133	0.133	0.834	0.120	0.120	0.120	0.120	0.120	0.120	0.120
k_n =small	0.790	0.192	0.211	0.226	0.239	0.221	0.201	0.183	0.834	0.171	0.187	0.205	0.221	0.203	0.184	0.166
k_n =large	0.790	0.234	0.277	0.307	0.335	0.306	0.273	0.230	0.834	0.215	0.254	0.289	0.319	0.289	0.255	0.211
Cragg Harvey	0.790	0.544	0.538	0.530	0.151	0.151	0.149	0.147	0.834	0.569	0.550	0.545	0.136	0.134	0.133	0.131

Table 4: Harvey $d_x = 10$, $d_r = 2$, normal

4a: $n = 100$									**4b:** $n = 200$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	1.033	0.220	0.220	0.220	0.220	0.220	0.220	0.220	1.003	0.187	0.187	0.187	0.187	0.187	0.187	0.187
k_n =small	1.033	0.445	0.470	0.503	0.546	0.496	0.427	0.359	1.003	0.356	0.399	0.436	0.477	0.423	0.354	0.295
k_n =large	1.033	0.459	0.526	0.591	0.675	0.624	0.542	0.434	1.003	0.387	0.459	0.531	0.604	0.547	0.458	0.359
Cragg Harvey	1.033	0.510	0.530	0.572	0.379	0.362	0.348	0.334	1.003	0.504	0.489	0.491	0.286	0.276	0.263	0.253

4c: $n = 500$									**4d:** $n = 1000$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2		2	small	large	d_x	large	small	2
Infeasible	1.013	0.168	0.168	0.168	0.168	0.168	0.168	0.168	1.007	0.157	0.157	0.157	0.157	0.157	0.157	0.157
k_n =small	1.013	0.278	0.320	0.368	0.406	0.353	0.288	0.249	1.007	0.247	0.287	0.336	0.381	0.330	0.270	0.230
k_n =large	1.013	0.315	0.398	0.476	0.535	0.473	0.385	0.298	1.007	0.290	0.366	0.447	0.515	0.453	0.364	0.282
Cragg Harvey	1.013	0.553	0.511	0.496	0.212	0.206	0.200	0.196	1.007	0.587	0.554	0.548	0.183	0.178	0.175	0.173

Table 5: Harvey $d_x = 5$, $d_r = 3$, normal

5a: $n = 100$									**5b:** $n = 200$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	2.282	0.379	0.110	0.110	0.110	0.110	0.110	0.441	2.686	0.355	0.089	0.089	0.089	0.089	0.089	0.412
k_n =small	2.282	0.553	0.423	0.462	0.490	0.447	0.403	0.676	2.686	0.537	0.349	0.382	0.414	0.375	0.324	0.657
k_n =large	2.282	0.669	0.566	0.661	0.763	0.688	0.602	0.878	2.686	0.683	0.482	0.568	0.650	0.584	0.492	0.852
Cragg Harvey	2.282	0.732	0.634	0.633	0.469	0.485	0.487	1.130	2.686	0.790	0.673	0.667	0.499	0.528	0.525	1.227

5c: $n = 500$									**5d:** $n = 1000$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	2.697	0.335	0.078	0.078	0.078	0.078	0.078	0.394	2.587	0.298	0.070	0.070	0.070	0.070	0.070	0.366
k_n =small	2.697	0.449	0.245	0.265	0.279	0.254	0.224	0.566	2.587	0.413	0.192	0.208	0.220	0.195	0.178	0.528
k_n =large	2.697	0.590	0.347	0.409	0.470	0.400	0.322	0.709	2.587	0.561	0.266	0.332	0.391	0.320	0.253	0.682
Cragg Harvey	2.697	0.824	0.688	0.682	0.372	0.365	0.362	1.104	2.587	0.886	0.790	0.779	0.249	0.232	0.224	0.798

Table 6: Harvey $d_x = 10$, $d_r = 3$, normal

6a: $n = 100$									**6b:** $n = 200$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	2.712	0.562	0.159	0.159	0.159	0.159	0.159	0.608	2.962	0.496	0.123	0.123	0.123	0.123	0.123	0.544
k_n =small	2.712	0.830	0.736	0.786	0.917	0.817	0.667	0.941	2.962	0.708	0.566	0.633	0.727	0.609	0.472	0.797
k_n =large	2.712	0.923	0.848	0.996	1.338	1.176	0.965	1.187	2.962	0.838	0.684	0.843	1.073	0.918	0.690	1.018
Cragg Harvey	2.712	0.908	0.858	0.975	0.649	0.667	0.757	1.533	2.962	0.881	0.700	0.762	0.526	0.570	0.573	1.462

6c: $n = 500$									**6d:** $n = 1000$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	2.903	0.452	0.102	0.102	0.102	0.102	0.102	0.497	2.877	0.427	0.087	0.087	0.087	0.087	0.087	0.471
k_n =small	2.903	0.614	0.404	0.476	0.543	0.441	0.328	0.705	2.877	0.570	0.316	0.392	0.459	0.361	0.277	0.664
k_n =large	2.903	0.777	0.530	0.695	0.861	0.699	0.488	0.882	2.877	0.740	0.433	0.611	0.785	0.616	0.421	0.842
Cragg Harvey	2.903	0.936	0.670	0.666	0.378	0.387	0.374	1.050	2.877	0.985	0.718	0.681	0.258	0.250	0.238	0.850

Table 7: Area $d_x = 5$, $d_r = 2$, normal

7a: $n = 100$									**7b:** $n = 200$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	0.101	0.026	0.026	0.026	0.026	0.026	0.026	0.026	0.103	0.027	0.027	0.027	0.027	0.027	0.027	0.027
k_n =small	0.101	0.037	0.042	0.047	0.052	0.047	0.041	0.034	0.103	0.034	0.039	0.044	0.049	0.045	0.039	0.032
k_n =large	0.101	0.044	0.054	0.062	0.068	0.063	0.055	0.042	0.103	0.039	0.050	0.057	0.064	0.058	0.051	0.038
Cragg Harvey	0.101	0.073	0.070	0.068	0.047	0.047	0.047	0.047	0.103	0.078	0.076	0.074	0.047	0.046	0.047	0.046

7c: $n = 500$									**7d:** $n = 1000$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	0.102	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.108	0.026	0.026	0.026	0.026	0.026	0.026	0.026
k_n =small	0.102	0.030	0.036	0.041	0.046	0.042	0.036	0.030	0.108	0.030	0.034	0.039	0.044	0.040	0.034	0.029
k_n =large	0.102	0.035	0.045	0.054	0.059	0.054	0.046	0.034	0.108	0.033	0.043	0.053	0.060	0.053	0.044	0.033
Cragg Harvey	0.102	0.078	0.078	0.078	0.048	0.048	0.048	0.048	0.108	0.085	0.084	0.084	0.050	0.050	0.049	0.049

Table 8: Area $d_x = 10$, $d_r = 2$, normal

8a: $n = 100$									**8b:** $n = 200$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	0.143	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.140	0.037	0.037	0.037	0.037	0.037	0.037	0.037
k_n =small	0.143	0.059	0.065	0.074	0.089	0.080	0.065	0.049	0.140	0.050	0.057	0.068	0.081	0.072	0.058	0.043
k_n =large	0.143	0.064	0.078	0.094	0.108	0.101	0.085	0.060	0.140	0.055	0.072	0.087	0.101	0.093	0.076	0.051
Cragg Harvey	0.143	0.077	0.069	0.072	0.067	0.066	0.065	0.064	0.140	0.085	0.071	0.066	0.062	0.062	0.061	0.061

8c: $n = 500$									**8d:** $n = 1000$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	0.142	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.136	0.036	0.036	0.036	0.036	0.036	0.036	0.036
k_n =small	0.142	0.043	0.053	0.064	0.076	0.068	0.053	0.040	0.136	0.041	0.050	0.062	0.071	0.062	0.051	0.039
k_n =large	0.142	0.047	0.067	0.083	0.097	0.087	0.069	0.046	0.136	0.044	0.063	0.078	0.089	0.079	0.064	0.043
Cragg Harvey	0.142	0.090	0.083	0.075	0.062	0.062	0.061	0.061	0.136	0.086	0.085	0.083	0.062	0.062	0.062	0.062

Table 9: Area $d_x = 5$, $d_r = 3$, normal

9a: $n = 100$									**9b:** $n = 200$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	0.136	0.105	0.031	0.031	0.031	0.031	0.031	0.115	0.139	0.107	0.031	0.031	0.031	0.031	0.031	0.116
k_n =small	0.136	0.094	0.068	0.080	0.090	0.081	0.068	0.116	0.139	0.097	0.060	0.074	0.086	0.075	0.061	0.119
k_n =large	0.136	0.110	0.092	0.105	0.113	0.106	0.094	0.124	0.139	0.111	0.087	0.101	0.111	0.103	0.090	0.125
Cragg Harvey	0.136	0.125	0.115	0.113	0.082	0.083	0.083	0.132	0.139	0.129	0.120	0.118	0.080	0.080	0.081	0.142

9c: $n = 500$									**9d:** $n = 1000$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	0.138	0.106	0.030	0.030	0.030	0.030	0.030	0.116	0.143	0.106	0.030	0.030	0.030	0.030	0.030	0.118
k_n =small	0.138	0.098	0.053	0.066	0.079	0.068	0.054	0.118	0.143	0.102	0.050	0.064	0.076	0.065	0.050	0.121
k_n =large	0.138	0.109	0.079	0.095	0.105	0.097	0.081	0.124	0.143	0.108	0.077	0.094	0.104	0.095	0.078	0.126
Cragg Harvey	0.138	0.128	0.119	0.119	0.082	0.082	0.082	0.156	0.143	0.133	0.123	0.124	0.080	0.080	0.081	0.161

Table 10: Area $d_x = 10$, $d_r = 3$, normal

10a: $n = 100$									**10b:** $n = 200$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	0.193	0.154	0.043	0.043	0.043	0.043	0.043	0.160	0.187	0.148	0.042	0.042	0.042	0.042	0.042	0.155
k_n =small	0.193	0.119	0.107	0.129	0.152	0.141	0.115	0.161	0.187	0.116	0.091	0.116	0.141	0.127	0.097	0.156
k_n =large	0.193	0.150	0.135	0.157	0.171	0.165	0.148	0.172	0.187	0.146	0.125	0.147	0.161	0.154	0.132	0.165
Cragg Harvey	0.193	0.157	0.115	0.109	0.120	0.120	0.120	0.178	0.187	0.163	0.126	0.112	0.107	0.108	0.108	0.186

10c: $n = 500$									**10d:** $n = 1000$							
	OLS	Pinkse				Robinson			OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2	1	2	small	large	d_x	large	small	2
Infeasible	0.192	0.149	0.041	0.041	0.041	0.041	0.041	0.157	0.184	0.145	0.041	0.041	0.041	0.041	0.041	0.152
k_n =small	0.192	0.126	0.080	0.110	0.135	0.118	0.084	0.160	0.184	0.128	0.073	0.102	0.123	0.105	0.075	0.154
k_n =large	0.192	0.151	0.116	0.147	0.161	0.151	0.123	0.169	0.184	0.145	0.105	0.138	0.150	0.140	0.109	0.161
Cragg Harvey	0.192	0.170	0.140	0.131	0.103	0.103	0.104	0.199	0.184	0.163	0.137	0.135	0.099	0.099	0.099	0.208

Table 11: OLS dominates Robinson $n = 10,000$ $R = 1000$, $d_x = 5$

11a: $d_r = 2$	OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2
Infeasible	1.699	1.589	0.189	0.173	0.173	0.201	0.979	1.963
Feasible	1.699	1.559	0.189	0.173	0.173	0.201	0.977	1.962
Cragg Harvey	1.709	1.709	1.713	1.715	3.277	2.848	2.447	2.160

11b: $d_r = 3$	OLS	Pinkse				Robinson		
$d_h =$	1	2	small	large	d_x	large	small	2
Infeasible	2.398	1.434	0.212	0.184	0.184	0.218	2.707	2.380
Feasible	2.398	1.408	0.212	0.185	0.185	0.219	2.704	2.379
Cragg Harvey	2.398	2.398	2.399	2.396	3.487	2.758	2.514	2.450