MORE EFFICIENT ESTIMATION UNDER NON-NORMALITY WHEN HIGHER MOMENTS DO NOT DEPEND ON THE REGRESSORS, USING RESIDUAL-AUGMENTED LEAST SQUARES*

Kyung So Im
University of Central Florida

Peter Schmidt
Michigan State University

August 1999
This revision, May, 2000

Abstract

Suppose that the Gauss-Markov assumptions hold, so least squares is best linear unbiased. Under normality, least squares is efficient. However, if the errors are not normal, we can hope to find extra efficiency by examining higher order moments. We can gain efficiency from knowledge of higher order moments of the errors, but also just from the assertion that these moments do not depend on the regressors. Thus, for example, the assumption of no conditional heteroskedasticity leads to more efficient estimation except when the third moment of the errors is zero, and similar statements hold for higher-order moments. In this paper we show how the assumption that higher moments do not depend on the regressors can be exploited in a GMM framework, and we provide very simple estimators that are equivalent to GMM estimators. These simple estimators can be calculated by linear regressions which have been augmented with functions of the least squares residuals.

*This is a revision of the paper "Least Squares Approach to Non-Normal Errors," 1996, DAE working paper No. 9603, University of Cambridge. The first author is grateful to DAE, ESRC, and Barton School of Business for the financial support. The second author gratefully acknowledges the financial support of the National Science Foundation. We thank James MacKinnon, Hashem Pesaran, Stephen Satchell and Jeffrey Wooldridge for their comments on the earlier version.

1 INTRODUCTION

In this paper we consider the efficiency gains that are possible in least squares regression from consideration of higher order moments of the errors. These efficiency gains can be realized in a GMM framework if the moments are known, but also if the moments are simply assumed to be unrelated to the regressors, and should be possible except when the errors are normal.

To be more precise, consider first the case that y_i, $i = 1, ..., N$, is a random sample from a distribution with mean μ and variance σ^2, which corresponds to the "intercept only" case of a regression. Define $\mu_j = E(y - \mu)^j$, $j = 2, 3, ...$, and assume for the moment that μ_j is finite for all j. The sample mean is based on the moment condition $E(y - \mu) = 0$, and is the efficient estimator under normality. Suppose now that we add the second moment condition $E[(y - \mu)^2 - \sigma^2] = 0$. If σ^2 is unknown the GMM estimator of μ is still the sample mean. But if σ^2 is known, we get a different estimator, and this estimator is more efficient than the sample mean unless $\mu_3 = 0$. Similarly, knowledge of the third moment allows us to improve on the sample mean unless $\mu_4 = 3\sigma^4$. Of course, under normality $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$, so we don't improve on the sample mean. We show that this result extends to higher moments, and that in fact the only distribution for which all higher moments exist but are uninformative for the mean is the normal distribution. When efficiency gains are possible, we show that they can be achieved by regressing y_i on an intercept and certain functions of the least squares residuals. For example, when σ^2 is known, the linearized GMM estimator can be calculated as the intercept in a regression of y_i on $[1, (e_i^2 - \sigma^2)]$, where $e_i = y_i - \bar{y}$; and when the third moment is known, the appropriate function of e_i to include in the regression is $e_i^3 - \mu_3 - 3\hat{\sigma}^2 e_i$, where $\hat{\sigma}^2$ is the usual error variance estimate. We call these residual augmented least squares (RALS) estimators. Note that we are augmenting the regression (of y_i on intercept) with constructed variables that are uncorrelated with the intercept, but correlated with the error. Following Wooldridge (1993) or Qian and Schmidt (1999), we improve efficiency by including variables with these properties (uncorrelated with the regressors but correlated with the errors) in the regression. Specifically, the ratio of the asymptotic variance of the augmented estimator to the asymptotic variance of the original estimator (the sample mean) equals $(1 - R^2)$, where R is the multiple correlation coefficient between the error and the set of augmenting variables. Thus in constructing these simple estimators we have also highlighted a simple method of quantifying their efficiency gain.

Now proceed to the regression case, so that $y_i = x_i'\beta + \epsilon_i$. The OLS estimator comes from the moment conditions: $E[x_i(y_i - x_i'\beta)] = 0$. Suppose that we add the conditions that: $E\{x_i[(y_i - x_i'\beta)^2 - \sigma^2]\} = 0$, which reflect an assumption of no conditional heteroskedasticity. It could be assumed that σ^2 is known, but it need not be, so long as it does not depend on x. The condition for an efficiency gain from these extra moment conditions is still just that $\mu_3 \neq 0$. If the third conditional moment, say $\mu_3(x)$, depends on x, this condition is slightly more subtle, but $\mu_3(x) = 0$ for all x is sufficient for there to be no possible efficiency gain. Thus we can obtain a more efficient GMM estimator by imposing no conditional heteroskedasticity if $\mu_3 \neq 0$. In the case that $\mu_3(x)$ and $\mu_4(x)$ do not depend on x, we can obtain a simple RALS estimator that is asymptotically equivalent to the GMM estimator, and we can easily quantify the efficiency gain. Furthermore, when $\mu_3(x)$ and $\mu_4(x)$ do not depend on x, we show that the RALS estimator is as efficient as the GMM estimator based on the *conditional* moment restrictions: $E[(y - x'\beta) \mid x] = 0, E[(y - x'\beta)^2 - \sigma^2 \mid x] = 0$. The quantification of the efficiency gain and the ability to achieve it with a RALS estimator, but not the existence of the efficiency gain, depend on the assumption that the third and fourth moments do not depend on x. A similar analysis applies to the use of moment conditions that assert that third or higher order moments do not depend on x, and this analysis can also perhaps be extended to cover features of the distribution of y other than moments.

An earlier paper by MaCurdy (1982) pursued essentially the same ideas as this paper does, in that information on higher moments is used in a GMM framework. However, the method of analysis is somewhat different, and the specific moment conditions used are not the same, so that the estimators considered in the two papers are different. Correspondingly the RALS estimators of this paper do not appear in MaCurdy (1982). Another paper that discusses some of the same ideas is Newey (1993). He considers moment conditions that arise when the conditional variance is known up to some parameters, and this encompasses the case of no conditional heteroskedasticity. However, because we consider a more specialized case, we are able to achieve some new results and simpler estimators.

The plan of the paper is as follows. In section 2 we discuss the problem of estimation of the mean in some detail. In section 3 we discuss the linear regression model under the assumption of no conditional heteroskedasticity. In section 4 we briefly discuss higher order moments. Some simulation results are given in section 5, and section 6 contains our concluding remarks. Some proofs are relegated to the appendix.

2 ESTIMATION OF THE SAMPLE MEAN

In this section we consider the case that $y_i, i = 1, ..., N$, is a random sample from a distribution with mean μ, which is the parameter of interest. Define $\mu_j = E(y - \mu)^j$, and $\sigma^2 = \mu_2$. We assume that μ_j is finite for all j; however, the validity of the asymptotics for the GMM estimator based on knowledge of μ_j only requires that μ_{2j} be finite. The sample mean is based on the moment equation $E(y_i - \mu) = 0$, and is the efficient estimator under normality. We are interested in whether and how we can improve the efficiency of estimation of μ if we assume that we know the value of μ_j for some integer $j \geq 2$.

We first address the question of when knowledge of μ_j improves the efficiency of estimation of μ. To do so, we consider the pair of moment conditions:

$$E(y_i - \mu) = 0, \tag{1A}$$

$$E[(y_i - \mu)^j - \mu_j] = 0. \tag{1B}$$

The GMM estimator of μ using both (1A) and (1B) must be no less efficient asymptotically than the GMM estimator using (1A) only, which of course is the sample mean. But sometimes the additional set of moment conditions (1B) does not increase efficiency, in the sense that the GMM estimator using both (1A) and (1B) is no more efficient than the sample mean, in which case we can say that (1B) is redundant given (1A). Conveniently evaluated conditions for redundancy are given in Breusch et al. (1999) – hereafter BQSW – and can be used to prove the following result.

Proposition 1 *For a given value of $j \in \{2, 3, ...\}$, knowledge of μ_j fails to increase the efficiency of estimation of μ if and only if the following condition holds:*

$$\mu_{j+1} = \sigma^2 j \mu_{j-1}. \tag{2}$$

Proof. *We will adopt the general notation of BQSW. The moment conditions (1A) and (1B) are of the form $E[g(y, \theta)] = 0$, where $\theta = \mu$, $g = (g_1, g_2)'$, $D = E[\partial g(y, \theta)/\partial\theta']$, $C = E[g(y, \theta)g(y, \theta)']$, and D and C are partitioned into submatrices D_1 and D_2, and C_{11}, C_{12}, $C_{21} = C_{12}$ and C_{22}, corresponding to the partitioning of g. BQSW (Theorem 1, p. 94) and MaCurdy (1982, equation (5)) show that a necessary and sufficient condition for g_2 to be redundant given g_1 is:*

$$D_2 = C_{21}C_{11}^{-1}D_1. \tag{3}$$

In the present case we have: $D_1 = -1$, $D_2 = -j\mu_{j-1}$, $C_{11} = \sigma^2$ and $C_{21} = \mu_{j+1}$. Inserting these into (3) and rearranging gives (2). ■

A few special cases are instructive. Knowledge of σ^2 is useful unless $\mu_3 = 0$. Knowledge of μ_3 is useful unless $\mu_4 = 3\sigma^4$. It is not coincidental that these relationships between moments are true for the normal distribution, since under normality the sample mean is efficient, and knowledge of higher moments cannot be useful. Indeed, equation (2) is exactly the equation that gives the higher moments ($j \geq 3$) for the normal distribution. It is well known that no other distribution than the normal satisfies these relationships (i.e. has cumulants of order three or higher equal to zero). So except under normality some higher moment is non-redundant for estimation of μ. We state this formally as the following result.

Proposition 2 *Let y_i, $i = 1, ..., N$, be iid with mean μ, and define $\mu_j = E(y - \mu)^j$. Suppose that μ_j is finite for all positive integers j. Then knowledge of μ_j for all j = 2,3,... is redundant for estimation of μ if and only if y is normal.*

Now consider estimation of μ when μ_j is known. The following result gives a simple estimator that is as efficient as the GMM estimator based on (1A) and (1B).

Proposition 3 *Let y_i, $i = 1, ..., N$, be iid with mean μ. For a given value of $j \in \{2, 3, ...\}$, suppose that $\mu_j = E(y-\mu)^j$ is known, and that μ_{2j} is finite. Let $e_i = y_i - \bar{y}$, and define (for any integer k) $m_k = \frac{1}{N}\sum_{i=1}^{N} e_i^k$, so that m_k is the k^{th} central sample moment of y. Define the constructed variable*

$$w_{ij} = e_i^j - \mu_j - jm_{j-1}e_i. \tag{4}$$

Let $\hat{\mu}$ be the coefficient of the intercept in a regression of y_i on $[1, w_{ij}]$. Then $\hat{\mu}$ has the same asymptotic distribution as the GMM estimator based on the moment conditions (1A) and (1B).
Proof. *See Appendix.* ■

We will call the estimator given in Proposition 3 a residual augmented least squares (or RALS) estimator, since the regression of y_i on intercept, which would yield the sample mean, is augmented with a function of the residual e_i. To understand intuitively why such an augmentation increases efficiency of estimation, consider the following. Define $\epsilon_i = y_i - \mu$ and $\omega_{ij} = \epsilon_i^j - \mu_j - j\mu_{j-1}\epsilon_i$. The quantity ω_{ij} is unobservable, but the observable w_{ij} in (4) can be regarded as an estimate of it. Since $E(\omega_{ij}) = 0$, it is uncorrelated with the intercept. Since $E(\epsilon_i\omega_{ij}) = \mu_{j+1} - \sigma^2 j\mu_{j-1}$, it is correlated with the error so long as the redundancy condition (2) does not hold. From Wooldridge (1993) and Qian and Schmidt (1999), we increase efficiency of estimation by augmenting a regression with variables that are uncorrelated with the regressor but correlated with the error. By doing so, the relevant error variance now becomes the variance conditional on the augmenting variables, which is smaller than the original error variance. Equivalently, augmentation reduces the asymptotic variance of estimation by the factor $(1 - R^2)$, where R is the multiple correlation between the error (ϵ_i) and the set of augmenting variables. In the present context, this correlation is zero when the redundancy condition (2) holds, but it may be high when the errors are sufficiently non-normal.

3 LINEAR REGRESSION

We now consider estimation of the regression model

$$y_i = x_i'\beta + \epsilon_i, \tag{5}$$

where x_i is a $k \times 1$ vector of explanatory variables. We assume that the observations $z_i \equiv (y_i, x_i')'$ are iid. We assume that the following moment conditions hold:

$$E[x'(y - x'\beta)] = 0, \tag{6A}$$
$$E\{x \otimes [h(y - x'\beta) - K]\} = 0. \tag{6B}$$

(6A) is the least squares moment condition which asserts that x and ϵ are uncorrelated, and (6B) refers to some additional moment conditions that some funtion of the error is uncorrelated with x. $h(\cdot)$ is $Jx1$ vector of differentiable function, and K.is $Jx1$ vector of constant. Therefore, we have kJ additional moment conditions.

We assume that the distribution of $z = (y, x')'$ satisfies the regularity conditions such that the GMM estimators we consider are consistent and asymptotically normal with variance matrix of the usual form. Listings of such regularity conditions are widely available and will not be repeated here.

An alternative to (6A) and (6B) would be the conditional moment restrictions:

$$E[(y - x'\beta) \mid x] = 0, \tag{7A}$$
$$E\{x \otimes [h(y - x'\beta) - K] \mid x\} = 0. \tag{7B}$$

These assert that $E(y \mid x) = x'\beta$ and that $E[h(y - x'\beta) \mid x]$ is constant, and imply (6A)-(6B). More precisely, (6A) and (6B) imply that the "errors" $\epsilon = y - x'\beta$ and $h(\epsilon)$ are uncorrelated with x, while (7A) and (7B) imply that ϵ and $h(\epsilon)$ are uncorrelated with any measurable function of x.

Imposing (6A) alone leads to ordinary least squares. We seek to improve on ordinary least squares using the additional moments (6B).

itional heteroskedasticity.

Imposing (6A) alone leads to ordinary least squares. We seek to improve on ordinary least squares using the homoskedasticity assumption (6B). We do not assume that σ^2 is known, just that it does not depend on x. (Knowledge of σ^2 seems unlikely as a practical matter, but we will make a few comments on the case that σ^2 is known, at the end of this section.) The fact that (6B) adds k moment conditions but only one parameter makes efficiency gains possible, even when σ^2 is unknown; in that respect this case is different from the case of the previous section.

An alternative to (6A) and (6B) would be the conditional moment restrictions:

$$E[(y - x'\beta) \mid x] = 0, \tag{7A}$$
$$E\left\{\left[(y - x'\beta)^2 - \sigma^2\right] \mid x\right\} = 0. \tag{7B}$$

These assert that $E(y \mid x) = x'\beta$ and that $var(y \mid x)$ is constant, and imply (6A)-(6B). More precisely, (6A) and (6B) imply that the "errors" $\epsilon = y - x'\beta$ and $\eta = (y - x'\beta)^2 - \sigma^2$ are uncorrelated with x, while (7A) and (7B) imply that ϵ and η are uncorrelated with any measurable function of x. Other functions of x are possible and perhaps natural. For example, MaCurdy (1982) imposes no conditional heteroskedasticity by calculating $E(y^2) = (x'\beta)^2 + \sigma^2$, and then considers nonlinear least squares applied to the regression

$$y_i^2 = (x_i'\beta)^2 + \sigma^2 + error. \tag{8}$$

The first order conditions for this nonlinear regression can be regarded as the following moment conditions:

$$E\{x_i(x_i'\beta)[y_i^2 - \sigma^2 - (x_i'\beta)^2]\} = 0. \tag{9}$$

These moment conditions are very similar to those in (6B) above, since the "error" in (9) is $y_i^2 - \sigma^2 - (x_i'\beta)^2 = [(y_i - x_i'\beta)^2 - \sigma^2 + 2(x_i'\beta)\epsilon_i$. Thus, while (6A) and (6B) assert that x_i is uncorrelated with ϵ_i and η_i, (9) requires that $x_i(x_i'\beta)$ be uncorrelated with $\eta_i + 2(x_i'\beta)\epsilon_i$. There is no obvious basis for asserting that (9) is more or less natural than (6B). Of course, both would follow from the conditional moment restrictions (7A)-(7B). An interesting question, to which we will return later in this section, is under what circumstances GMM based on (6A)-(6B) is as efficient as GMM based on the conditional moment restrictions (7A)-(7B).

We first ask under what conditions imposition of (6B) increases efficiency of estimation of β, relative to ordinary least squares. As in the previous section, this depends on higher moments. As a matter of notation, we define $\mu_j(x) = E(\epsilon^j \mid x)$ for positive integers j. The application of GMM to (6A)-(6B) requires that $\mu_3(x)$ and $\mu_4(x)$ be finite. Then we have the following result.

Proposition 4 *The use of the moment conditions (6B) in addition to (6A) fails to increase the efficiency of estimation of β if $E[\epsilon xx'] = E[\mu_3(x)xx'] = 0$.*
Proof. *See Appendix.* ■

This result is, loosely speaking, a generalization of the last section's result that knowledge of σ^2 is useful unless $\mu_3 = 0$. We note the following. First, the condition of Proposition 4 is sufficient, not necessary, but we cannot identify any meaningful circumstances under which (6B) is redundant without this condition holding. Equation (A8) of the Appendix gives a necessary and sufficient condition, but it is not revealing. Second, if the conditional moment restriction $E(\epsilon \mid x) = 0$ holds, then $E(\epsilon xx') = 0$ and the redundancy condition becomes $E[\mu_3(x)xx'] = 0$. Clearly this condition reduces to $\mu_3 = 0$ if $\mu_3(x)$ does not depend on x. However, when $\mu_3(x)$ does depend on x, it is weaker than the condition $\mu_3(x) \equiv 0$, which is the condition for (7B) to be redundant given (7A), as we will discuss below.

The form of the GMM estimator depends on $\mu_3(x)$ and $\mu_4(x)$. The following result shows that, if the conditional moment restrictions (7A)-(7B) hold and if $\mu_3(x)$ and $\mu_4(x)$ do not depend on x, a simple RALS estimator is as efficient as the GMM estimator based on (6A)-(6B).

Proposition 5 *For* $i = 1,...,N$, *let* e_i *be the OLS residuals (from the regression of y on x); let* $\hat{\sigma}^2$ *be the usual error variance estimate; and define the constructed variable* $w_{i2} = e_i^2 - \hat{\sigma}^2$. *Define the RALS estimate of* β *to be the coefficients of* x_i *in a regression of* y_i *on* $[x_i', w_{i2}]$. *Suppose that the regression contains an intercept, that the conditional moment restrictions (7A) and (7B) hold, and that* $\mu_3(x)$ *and* $\mu_4(x)$ *are finite and do not depend on x. Then the RALS estimate has the same asymptotic distribution as the GMM estimate based on the moment conditions (6A) and (6B).*
Proof. *See Appendix.* ■

Chamberlain (1987) showed that, if the no conditional heteroskedasticity condition (7B) holds (but is not used as a source of moment conditions in estimation), the efficient GMM estimator using the conditional mean condition (7A) is ordinary least squares. Loosely speaking, when (7B) holds, the uncorrelatedness condition (6A) efficiently captures the conditional mean information in (7A). It is therefore natural to ask whether similar assumptions about higher moments might also imply the efficiency of simple estimators under the conditional moment restrictions (7A) and (7B). The answer is yes, as the next two results show.

Proposition 6 *Suppose that the conditional moment restrictions (7A) and (7B) hold and that* $\mu_3(x) = 0$ *for all x. Then the efficient GMM estimator of* β *given the conditional moment restrictions (7A) and (7B) is ordinary least squares.*
Proof. *See Appendix.* ■

Proposition 6 is not surprising, since it essentially extends the earlier result that second moment information is useless if the third moment is zero. The results of Newey (1993, p. 427) imply that (7B) is redundant given (7A) if the third conditional moment is zero for all x. Since ordinary least squares is efficient when (7B) holds but is not imposed, as shown by Chamberlain (1987), it should be efficient also when (7B) is imposed but redundant. The following result is a perhaps more important implication of the constancy of the third and fourth conditional moments.

Proposition 7 *Suppose that the regression contains an intercept, that the conditional moment restrictions (7A)-(7B) hold, and that* $\mu_3(x)$ *and* $\mu_4(x)$ *are finite and do not depend on x. Then the GMM estimator based on (6A)-(6B) is as efficient as the efficient GMM estimator based on the conditional moment restrictions (7A)-(7B).*
Proof. *See Appendix.* ■

Proposition 7 is a possible justification for the particular form of the moment conditions (6A)-(6B) and for the RALS estimator. Combining Propositions 5 and 7, we see that the RALS estimator efficiently exploits the conditional moment restrictions (7A) and (7B) when the third and fourth conditional moments are constant.

If (7A) and (7B) hold but $\mu_3(x)$ and $\mu_4(x)$ depend on x, then the efficient GMM estimator based on (7A) and (7B) dominates GMM based on (6A) and (6B), which dominates the RALS estimator and ordinary least squares. The extent to which GMM is more efficient than RALS would depend on how strongly $\mu_3(x)$ and $\mu_4(x)$ depend on x. Given that no conditional heteroskedasticity is assumed, it may be reasonable to choose a simple estimator (RALS) rather than a complicated estimator (GMM) whose efficiency gain depends on the way that $\mu_3(x)$ and $\mu_4(x)$ depend on x. In our view, the main implication for RALS of the dependence of $\mu_3(x)$ and $\mu_4(x)$ on x is that the conventionally-calculated standard errors (using the usual OLS formula) would be invalid. Rather, the usual heteroskedasticity-robust standard errors would be needed.

As noted above, the discussion so far in this section treats σ^2 as unknown. Although knowledge of σ^2 would seem unlikely in general, for completeness we will discuss this case briefly. Propositions 4, 6 and 7 still hold when σ^2 is known. If we modify the RALS estimator of Proposition 5 by defining $w_{i2} = e_i^2 - \sigma^2$ (instead of $e_i^2 - \hat{\sigma}^2$ as in Proposition 5), then Proposition 5 also still holds. Interestingly, under the assumptions of Proposition 5, knowledge of σ^2 improves the efficiency of estimation of β (by GMM based on (6A)-(6B) or RALS) so long as $\mu_3 \neq 0$, but the only difference is for the intercept. That is, for the non-constant regressors, the efficiency of estimation is the same whether σ^2 is known or unknown. That this is so is most easily seen from the last line of equation (A10) of the Appendix, in which changing $\hat{\sigma}^2$ to σ^2 affects only the intercept, but it does so in such a way as to affect the asymptotic distribution of the intercept.

Although the efficiency of the estimate of the intercept is not a major point, a precise comparison is possible. Let $x_i = (1, x_{i*}')'$ so as to distinguish intercept from non-constant regressors, and define $\Sigma_* = E(x_{i*}x_{i*}')$, $\mu_* = E(x_{i*})$ and $V_* = \Sigma_* - \mu_*\mu_*'$. Then a tedious calculation (available from the authors on request) yields the asymptotic variance of the estimated intercept when σ^2 is unknown as

$$\sigma^2 + \left[\sigma^2 - \frac{\mu_3^2}{(\mu_4 - \sigma^4)}\right] \mu_*' V_*^{-1} \mu_*, \tag{10}$$

whereas when σ^2 is known the asymptotic variance is

$$\left[\sigma^2 - \frac{\mu_3^2}{(\mu_4 - \sigma^4)}\right] (1 + \mu_*' V_*^{-1} \mu_*). \tag{11}$$

The difference between (10) and (11) equals $\mu_3^2/(\mu_4 - \sigma^4)$ and is positive when $\mu_3 \neq 0$. We can also compare (10) to the variance of the OLS estimated intercept, which is

$$\sigma^2(1 + \mu_*' V_*^{-1} \mu_*). \tag{12}$$

The difference between (12) and (10) equals $[\mu_3^2/(\mu_4 - \sigma^4)]\, \mu_*' V_*^{-1} \mu_*$, and is positive unless $\mu_3 = 0$ or $\mu_* = 0$.

4 THIRD MOMENTS

We now consider the use of moments of higher order than two. Specifically, we will consider the use of third moment first, then the use of second and third moments together. The analysis of fourth or higher order moments would follow very similar lines. It is interesting to see how our analysis of the assumption of no conditional heteroskedasticity generalizes to higher moments, and it is not transparent algebraically how things should work out when more than one set of moments (e.g. second and third) is considered.

The unconditional third moment assumption is:

$$E\left\{x_i\left[\left(y_i - x_i'\beta\right)^3 - \mu_3\right]\right\} = 0. \tag{6C}$$

The numbering (6C) is used to stress that this is a logical extension of the first and second moment assumptions (6A) and (6B) above. We treat μ_3 as unknown. Knowledge of μ_3 would matter only for the intercept (just as knowledge of σ^2 mattered only for the intercept in the previous section). Known μ_3 is perhaps potentially more relevant than knowledge of σ^2; in particular, in some circumstances one could imagine asserting $\mu_3 = 0$.

The conditional moment restriction analogous to (6C) is:

$$E\left\{\left[\left(y - x'\beta\right)^3 - \mu_3\right] \mid x\right\} = 0. \tag{7C}$$

As in the previous section, we are interested in circumstances in which the unconditional and conditional moment restrictions yield equally efficient estimators.

We first consider the case in which estimation is based on first and third moments; that is, on (6A) and (6C) or (7A) and (7C). This is not a terribly interesting case, because it seems implausible to assume (6C) without assuming (6B), or (7C) without (7B), but it could arise if, for example, we asserted $\mu_3 \equiv 0$ without an assumption of no conditional heteroskedasticity. The following results are the generalizations of Propositions 4 - 7 of the previous section. We state them without proof because the proofs are similar to those of the last section, and more importantly because they overlap substantially with the proofs for the case that all three moments are used.

Proposition 8 *The use of the moment conditions (6C) in addition to (6A) fails to increase the efficiency of estimation of β if*

$$E(\epsilon^4 xx') = 3E(\epsilon^2 xx')\left[E\left(xx'\right)\right]^{-1} E(\epsilon^2 xx'). \tag{13}$$

Proposition 9 *For i = 1,...,N, let e_i be the OLS residuals (from the regression of y on x); let $\hat{\sigma}^2$ be the usual error variance estimate; let $m_3 = \frac{1}{N}\sum_{i=1}^{N} e_i^3$ be the sample third moment (of the errors); and define the constructed variable $w_{i3} = e_i^3 - m_3 - 3\hat{\sigma}^2 e_i$. Define the RALS estimate of β to be the coefficients* *of x_i in a regression of y_i on $[x_i', w_{i3}]$. Suppose that the regression contains an intercept, that the conditional moment restrictions (7A), (7B) and (7C) hold, and that $\mu_4(x)$ and $\mu_6(x)$ are finite and do not depend on x. Then the RALS estimate has the same asymptotic distribution as the GMM estimate based on the moment conditions (6A) and (6C).*

Proposition 10 *Suppose that the conditional moment restrictions (7A), (7B) and (7C) hold and that $\mu_4(x) = 3\sigma^4$ for all x. Then the efficient GMM estimator of β given the conditional moment restrictions (7A) and (7C) is ordinary least squares.*

Proposition 11 *Suppose that the regression contains an intercept, that the conditional moment restrictions (7A), (7B) and (7C) hold, and that $\mu_4(x)$ and $\mu_6(x)$ are finite and do not depend on x. Then the GMM estimator based on (6A) and (6C) is as efficient as the efficient GMM estimator based on the conditional moment restrictions (7A) and (7C).*

Propositions 8 and 10 are obvious extensions of the condition that knowledge of μ_3 does not help in estimation of μ if $\mu_4(x) = 3\sigma^4$.

We now turn to the case of main interest, in which estimation is based on first, second and third moments; that is, on (6A), (6B) and (6C) or on (7A), (7B) and (7C). We first give the redundancy result for estimation based on the unconditional moments.

Proposition 12 *The use of the moment conditions (6B) and (6C) in addition to (6A) fails to increase the efficiency of estimation of β if*

$$E(\epsilon xx') = E\left[\mu_3\left(x\right)xx'\right] = 0 \text{ and } E(\epsilon^4 xx') = 3E(\epsilon^2 xx')\left[E\left(xx'\right)\right]^{-1}E(\epsilon^2 xx'). \tag{14}$$

Proof. *See Appendix.* ■

This result is basically just a combination of Propositions 4 and 8, in the sense that (14) just combines the conditions of those results. If the conditional moment restrictions (7A)-(7C) hold, and if $\mu_4(x)$ does not depend on x, these conditions reduce to $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$, which are the conditions that would arise in the case of estimation of the mean only, as in section 2. If ϵ is normal and independent of x, these conditions hold, and we do not improve on least squares, which is efficient.

Proposition 13 *Suppose that the conditional moment restrictions (7A), (7B) and (7C) hold, and that $\mu_3 = 0$ and $\mu_4(x) = 3\sigma^4$ for all x. Then the efficient GMM estimator of β given the conditional moment restrictions (7A), (7B) and (7C) is ordinary least squares.*
Proof. *See Appendix.* ■

Proposition 14 *Suppose that the regression contains an intercept, that the conditional moment restrictions (7A), (7B) and (7C) hold, and that $\mu_4(x)$, $\mu_5(x)$ and $\mu_6(x)$ are finite and do not depend on x. Then the GMM estimator based on (6A), (6B) and (6C) is as efficient as the efficient GMM estimator based on the conditional moment restrictions (7A), (7B) and (7C).*
Proof. *See Appendix.* ■

Proposition 13 basically combines Propositions 6 and 10 to give conditions under which ordinary least squares is efficient even given the information that the second and third conditional moments of the error are constant. These conditions are weaker than the assumption that ϵ is normal and independent of x, but perhaps not much weaker. Proposition 14 is arguably more interesting. It gives the conditions under which the unconditional moment conditions (6A), (6B) and (6C) efficiently capture the information in the conditional moment conditions (7A), (7B) and (7C), in the sense that the corresponding GMM estimators are equally efficient. These conditions would be satisfied if ϵ and x are independent. It should be noted that the conditions identified in Proposition 14 are not just a combination of the conditions in Propositions 7 and 11.

Our final result provides the RALS estimator that makes use of the assumption that both second and third moments do not depend on x.

Proposition 15 *Suppose that the regression contains an intercept, that the conditional moment restrictions (7A), (7B) and (7C) hold, and that $\mu_4(x)$, $\mu_5(x)$ and $\mu_6(x)$ are finite and do not depend on x. Let w_{i2} and w_{i3} be as defined in the statements of Propositions 5 and 9, and define the RALS estimator of β to be the coefficients of x_i in a regression of y_i on $[x_i', w_{i2}, w_{i3}]$. Then the RALS estimator has the same asymptotic distribution as the GMM estimator based on the moment conditions (6A), (6B) and (6C).*
Proof. *See Appendix.* ■

This result is in a logical sense a combination of Propositions 5 and 9 above. From an algebraic point of view it is really not a trivial combination, because the additional regressors w_{i2} and w_{i3} are not generally orthogonal. (They are orthogonal if $\mu_5 = 4\sigma^2\mu_3$.) However, whether w_{i2} and w_{i3} are orthogonal or not is not fundamentally important. The conditions that drive the efficiency gain from the RALS estimator are that the variables $[w_{i2}, w_{i3}]$ be uncorrelated with x_i but correlated with ϵ_i. As before, this reduces the asymptotic variance of estimation by the factor $(1 - R^2)$, where R is the multiple correlation coefficient between ϵ_i and $[w_{i2}, w_{i3}]$.

Since the conditions of Proposition 15 are the same as those of Proposition 14, combining the two results provides a useful efficiency result. The moment conditions (7A), (7B) and (7C) assert linearity of the regression, and that the second and third moments of y conditional on x do not depend on x. The efficient GMM estimator based on these conditions is in general complicated. However, when the higher moments of ϵ up to order six also do not depend on x, we can without loss of efficiency replace the conditional moment restrictions by the unconditional moment restrictions (6A), (6B) and (6C), and a simple RALS estimator is as efficient as the efficient GMM estimator.

5 EFFICIENCY CALCULATIONS AND MONTE CARLO RESULTS

In this section we provide some calculations and comparisons of asymptotic variances for various estimators, and we report the results of some Monte Carlo simulations that examine the relevance of these asymptotic results in finite samples.

We begin with the asymptotic efficiency comparisons. The asymptotic variance of the OLS estimate of β (normalized by $\sqrt{N}$) is $\sigma^2 E(xx')^{-1}$. Under the conditions of Proposition 15 (so that the moments of order two through six of the error conditional on x do not depend on x), the asymptotic variance of the RALS estimator is the same as the asymptotic variance of the efficient GMM estimator, and is of the form $\sigma_A^2 E(xx')^{-1}$, where

$$\sigma_A^2 = (\sigma^2 - \Sigma_{\epsilon w}\Sigma_{ww}^{-1}\Sigma_{w\epsilon}), \tag{15}$$

and where $w = w_2$ as in Proposition 5 when only second moments are used, whereas $w = (w_2, w_3)$ as in Proposition 15 when second and third moments are used. Thus the ratio of the asymptotic variance of the RALS estimator to the asymptotic variance of the OLS estimator is just σ_A^2/σ^2. For the case that only no conditional heteroskedasticity is imposed, we have $\sigma_A^2 = \sigma^2 - \mu_3^2/\left(\mu_4 - \sigma^4\right)$; for the case that both second and third moments are used, we have

$$\sigma_A^2 = \tag{16}$$
$$\sigma^2 - \frac{\mu_3^2\left(\mu_6 - 6\mu_4\sigma^2 + 9\sigma^6 - \mu_3^2\right) - 2\mu_3\left(\mu_4 - 3\sigma^4\right)\left(\mu_5 - 4\mu_3\sigma^2\right) + \left(\mu_4 - 3\sigma^4\right)^2\left(\mu_4 - \sigma^4\right)}{\left(\mu_4 - \sigma^4\right)\left(\mu_6 - 6\mu_4\sigma^2 + 9\sigma^6 - \mu_3^2\right) - \left(\mu_5 - 4\mu_3\sigma^2\right)^2}.$$

For the normal distribution, OLS is efficient and these ratios equal one. Table 1 gives these ratios of asymptotic variances for various non-normal distributions. We consider chi-squared with one, two, three, four, six and ten degrees of freedom; Student t with seven, eight and ten degrees of freedom; the double exponential distribution; and the Beta(2,2) distribution. We do not consider t distributions with degrees of freedom less than seven because the RALS/GMM estimator using second and third moments requires the sixth moment to be finite, which requires seven or more degrees of freedom; and there is not much point in considering the RALS/GMM estimator using only second moments for the t distribution, since there would be no efficiency gain over OLS.

Table 1 clearly shows the potential for non-trivial efficiency gains from asserting the constancy of second or third conditional moments. When only the second moment is used, i.e. we impose no conditional heteroskedasticity, there is no efficiency gain for the symmetric distributions, but there are considerable gains for all of the chi-squared distributions, even with degrees of freedom as high as ten. When constancy of second and third moments is imposed, we still have only modest gains in efficiency for the t distributions and the double exponential, but we have considerable gains for the chi-squared and Beta(2,2). In the chi-squared case, the gain from using the constancy of the third moment in addition to the second moment is not very large, however, except when the number of degrees of freedom is very small.

We now turn to Monte Carlo simulation to provide evidence on the finite sample performance of these estimators. The data are generated from a simple regression model:

$$y_i = \alpha + \beta x_{i*} + \epsilon_i,\ \ i = 1, ..., N. \tag{17}$$

We pick $\alpha = \beta = 1$; these values do not affect the results. The x_{i*} are chosen to be iid N(0,1). With this choice, and with the previously-used notation that $x_i = [1, x_{i*}]'$, we have $E(x_i x_i') = I_2$. Our errors will be drawn from different distributions, but will be normalized to have mean zero and variance one. (For example, for chi-squared with n degrees of freedom, we subtract n from the generated χ^2 variable, and then divide by the square root of $2n$.) With this normalization the variance of the OLS estimate of α or β will equal one, and the asymptotic variance of the RALS estimator of β will equal $\sigma_A^2 \leq 1$, as given above. We consider normal errors, and also errors following the same distributions as were considered in the asymptotic variance calculations above. We use the following values of sample size: $N = 50, 100, 1000, 5000$. The number of replications is 5000.

The pseudo-normal random numbers used in the experiments are generated by the built-in procedure in GAUSS, which uses a standard acceptance-rejection algorithm. The various non-normal errors are generated by appropriate transformations of independent standard normals. In particular, the generation of the Beta(2,2) random variables follows the method of Jambunathan (1954). See also Johnson and Kotz (1970, p. 38). The double exponential random variables are calculated by applying the inverse of the cdf to pseudo-uniform variables.

We consider three estimators: OLS; RALS, using second moments only (augmentation of x_i by w_{i2} only); and RALS, using second and third moments (augmentation of x_i by w_{i2} and w_{i3}). We could also have considered various types of GMM and linearized GMM estimators that are asymptotically equivalent to RALS, but these estimators are all much more complicated computationally.

Table 2 gives N times the variance of the RALS estimates of β. For OLS, N times variance equals one, and that is one standard of comparison. The other standard of comparison is the asymptotic variance, given in the last column ($N =$

∞), which is the same as is reported in Table 1. We intended to report also the mean (or bias) of the estimates, but do not do so because the estimates turned out to be essentially unbiased even for the smallest sample sizes.

The results in Table 2 seem to support the applicability of the asymptotic theory even for moderate sample sizes, such as $N = 100$. For $N \geq 100$, $N\cdot$Variance of RALS is generally quite similar to the asymptotic value given in the last column of the table. This has several obvious implications. First, there is little loss in finite samples in imposing conditions that are asymptotically redundant. That is, in cases in which there is no gain to using higher moments (i.e., under normality, or for the symmetric distributions when only second moments are used), the finite sample variance of RALS is only slightly larger than the variance of OLS. Second, where higher moments are asymptotically useful, there is almost always a gain in finite sample variance as well ($N\cdot$Variance of RALS is less than one), and the asymptotic variance does a reasonable job of predicting the size of the gain.

Efficiency of estimation is important, but it is also relevant to ask whether the RALS procedure leads to valid inference. In Table 3, we provide the frequency of rejection (empirical size) of the Wald test of the null hypothesis that β equals one (its true value). Nominal size is 5%. We give empirical size for the OLS estimator as well as the RALS estimators, since the OLS Wald tests will not have size equal to 5% in finite samples except under normality. For the RALS estimators, the Wald test is just the usual t-test applied to the augmented regression. This test is valid asymptotically but its finite sample size characteristics are unknown.

It is obvious from Table 3 that the empirical size of the test is usually quite close to 0.05. The largest size distortions are for small values of N and for the case that both second and third moments are used, but for $N \geq 100$ we would categorize these as not terribly serious. For example, for $N = 100$ the worst case is Beta(2,2), for which empirical size is 0.076, while for $N \geq 1000$, there are no size distortions serious enough to comment on.

The lack of serious size distortions is not surprising given the results of Table 2, which indicated that the finite sample variance is generally close to the asymptotic variance. Both of these optimistic outcomes simply indicate that asymptotic theory applies reasonably well to simple estimators such as our RALS estimator.

6 CONCLUDING REMARKS

This paper has asked when and by how much it is possible to improve on the ordinary least squares estimator by using moment conditions implied by the assumption that higher moments of the error do not depend on the regressors. The most important case is undoubtedly the one in which we consider second moments, so that the question is whether the assumption of no conditional heteroskedasticity is valuable for estimation of β. Basically it is, unless the third moment of the error is zero, and the efficiency gain from imposing no conditional heteroskedasticity depends on the magnitude of the third moment.

Since no conditional heteroskedasticity is naturally phrased as a statement about the conditional second moment of the error, it leads to conditional moment restrictions and potentially complicated GMM estimation. We give conditions under which very simple estimators are as efficient as the efficient GMM estimator. These conditions are straightforward: the third and fourth conditional moments of the error do not depend on the regressors. The simple efficient estimators just involve a regression augmented with functions of the least squares residuals. Similar results hold for the imposition of conditional moment restrictions based on higher order moments.

It is interesting to ask how general these results are. They would extend in a simple way to nonlinear regression, but not to simultaneous equations. It is not clear how easily they can be extended to features of the distribution of y given x other than moments (e.g., conditional quantiles).

The moment conditions we consider are obviously a subset of those implied by independence of the errors and regressors. If one is willing to assume independence, then adaptive estimation is possible, as shown by Bickel (1982) and Manski (1984), so that the relevant efficiency bound is the efficiency of the MLE (as if the distribution of the errors were known). The adaptive MLE estimators of Bickel and Manski involve nonparametric estimation of the score function, whereas Newey (1987) showed how to construct an efficient estimator using GMM with the number of moment conditions growing with sample size. Compared to these estimators, ours are less efficient under independence. The RALS/GMM estimators require certain moments to exist, which the adaptive estimators do not. However, if these moments exist, the RALS estimators rely on weaker assumptions and so should be more robust. Independence is a very strong assumption compared to no conditional heteroskedasticity, and one might not wish to pursue efficiency gains that depended on obscure features of the distribution. Furthermore, our RALS estimators are numerically simpler than the adaptive estimators, and so they might be expected to have better small sample properties. Hsieh and Manski (1987) and Newey (1988) provide some Monte Carlo evidence on the performance of the adaptive estimators in small samples, but their results are hard to compare to ours because of the usual differences in setup of experiments. For example, the only distributions used by both us and Hsieh and Manski are normal and Beta(2,2), and their regressor is binary while ours is normal. For sample size $N = 50$, in the normal case, their adaptive MLE is closer to the efficiency of OLS than our RALS estimators are. On the other hand, in the Beta(2,2) case, we do better than OLS and the adaptive MLE doesn't. Clearly further simulations would be needed to give even a partial answer to the question of how large a sample size is needed before the adaptive estimators are superior under independence. However, our intuition remains that there is always a setting in which simple estimators are worth having.

A Appendix

A.1 Proof of Proposition 3

Because we cannot express the GMM estimator in closed form, we will consider instead a linearized GMM (LGMM) estimator, using the initial consistent estimator $\bar{y}$. Subject to regularity conditions that are satisfied for this case, the LGMM estimator has the same asymptotic distribution as the GMM estimator.

To define notation, we consider the moment conditions $Eg(y_i, \mu) = 0$; their sample version, $\bar{g}(\mu) = \frac{1}{N}\sum_i g(y_i, \mu)$; and the sample version evaluated at the initial estimate, $\bar{g}(\bar{y})$, where

$$g(y_i, \mu) = \begin{bmatrix} y_i - \mu \\ (y_i - \mu)^j - \mu_j \end{bmatrix}, \; \bar{g}(\mu) = \begin{bmatrix} \bar{y} - \mu \\ \frac{1}{N}\sum_i (y_i - \mu)^j - \mu_j \end{bmatrix}, \; \bar{g}(\bar{y}) = \begin{bmatrix} 0 \\ m_j - \mu_j \end{bmatrix}. \tag{A1}$$

The LGMM estimator is

$$\tilde{\mu} = \bar{y} - \left(D'C^{-1}D\right)^{-1} D'C^{-1}\bar{g}(\bar{y}), \tag{A2}$$

where D and C are defined in the proof of Proposition 1. We will evaluate D and C at the true value of μ, so that $\tilde{\mu}$ is not a feasible estimator; but we simply need it as a standard of comparison. (Evaluating D and C at $\mu = \bar{y}$ would give a feasible estimator, with the same asymptotic distribution as $\tilde{\mu}$.)

In the present case, $D_1 = -1$, $D_2 = -j\mu_{j-1}$, $C_{11} = \sigma^2$, $C_{12} = \mu_{j+1}$ and $C_{22} = \mu_{2j} - \mu_j^2$. The determinant of C is $DET = \sigma^2\left(\mu_{2j} - \mu_j^2\right) - \mu_{j+1}^2$. Then we have

$$D'C^{-1}D = \left(\mu_{2j} - \mu_j^2 + \sigma^2 j^2 \mu_{j-1}^2 - 2j\mu_{j-1}\mu_{j+1}\right)/DET, \tag{A3a}$$

$$D'C^{-1}\bar{g}(\bar{y}) = -\left(\sigma^2 j\mu_{j-1} - \mu_{j+1}\right)\left(m_j - \mu_j\right)/DET, \tag{A3b}$$

$$\tilde{\mu} = \bar{y} - \frac{b_1}{b_2}\left(m_j - \mu_j\right), \tag{A3c}$$

where

$$b_1 = \mu_{j+1} - \sigma^2 j\mu_{j-1}, \tag{A4a}$$

$$b_2 = \mu_{2j} - \mu_j^2 + \sigma^2 j^2 \mu_{j-1}^2 - 2j\mu_{j-1}\mu_{j+1}. \tag{A4b}$$

We now consider RALS estimator, which is the intercept in the regression of y_i on $\left[1, \left(e_i^j - \mu_j\right) - jm_{j-1}e_i\right]$. Using standard regression notation we have

$$\frac{1}{N}X'X = \begin{bmatrix} 1 & \left(m_j - \mu_j\right) \\ \left(m_j - \mu_j\right) & \left(m_{2j} + \mu_j^2 - 2m_j\mu_j + j^2\hat{\sigma}^2 m_{j-1} - 2jm_{j-1}m_{j+1}\right) \end{bmatrix}, \tag{A5a}$$

$$\frac{1}{N}X'Y = \begin{bmatrix} \bar{y} \\ \frac{1}{N}\sum_i y_i \left(e_i^j - \mu_j\right) - jm_{j-1}\frac{1}{N}\sum_i e_i y_i \end{bmatrix} \quad \text{(A5b)}$$

$$= \begin{bmatrix} \bar{y} \\ m_{j+1} + \bar{y}\left(m_j - \mu_j\right) - jm_{j-1}\hat{\sigma}^2 \end{bmatrix},$$

using the facts that $\sum_i e_i y_i = \sum_i e_i^2$, $\sum_i y_i \left(e_i^j - \mu_j\right) = \sum_i \left(e_i + \bar{y}\right)\left(e_i^j - \mu_j\right) = \bar{y}\sum_i e_i^j + \sum_i e_i^{j+1} - N\bar{y}\mu_j$.

Then we have

$$\left(X'X\right)^{-1} X'Y = \frac{1}{D_*}\begin{bmatrix} \left(m_{2j} + \mu_j^2 - 2m_j\mu_j + j^2\hat{\sigma}^2 m_{j-1} - 2jm_{j-1}m_{j+1}\right) & -\left(m_j - \mu_j\right) \\ -\left(m_j - \mu_j\right) & 1 \end{bmatrix} \times$$

$$\begin{bmatrix} \bar{y} \\ m_{j+1} + \bar{y}\left(m_j - \mu_j\right) - jm_{j-1}\hat{\sigma}^2 \end{bmatrix}. \quad \text{(A6)}$$

where $D_* = m_{2j} + \mu_j^2 - 2m_j\mu_j + j^2\sigma^2 m_{j-1} - 2jm_{j-1}m_{j+1} - \left(m_j - \mu_j\right)^2$. We need only the first element, which we can write as:

$$\hat{\mu} = A\bar{y} - B\left(m_j - \mu_j\right), \quad \text{(A7a)}$$

where

$$A = \frac{m_{2j} + \mu_j^2 - 2m_j\mu_j + j^2\hat{\sigma}^2 m_{j-1} - 2jm_{j-1}m_{j+1}}{m_{2j} + \mu_j^2 - 2m_j\mu_j + j^2\hat{\sigma}^2 m_{j-1} - 2jm_{j-1}m_{j+1} - \left(m_j - \mu_j\right)^2}, \quad \text{(A7b)}$$

$$B = \frac{m_{j+1} - j\hat{\sigma}^2 m_{j-1} + \bar{y}\left(m_j - \mu_j\right)}{m_{2j} + \mu_j^2 - 2m_j\mu_j + j^2\hat{\sigma}^2 m_{j-1} - 2jm_{j-1}m_{j+1} - \left(m_j - \mu_j\right)^2}. \quad \text{(A7c)}$$

But $\text{plim}A = 1$, while $\text{plim}B = b_1/b_2$, with b_1 and b_2 defined in (A4a) and (A4b). It follows that $\hat{\mu}$ and $\tilde{\mu}$ have the same asymptotic distribution, and therefore $\hat{\mu}$ and the GMM estimator have the same asymptotic distribution.

A.2 Proof of Proposition 4

The moment conditions (6A)-(6B) are of the form $Eg\left(z,\theta\right) = 0$, where as before we partition $g\left(z,\theta\right) = \left[g_1\left(z,\theta\right)', g_2\left(z,\theta\right)'\right]'$. Now however we also partition θ into θ_1 $\left(=\beta\right)$ and θ_2 $\left(=\sigma^2\right)$. We ask when g_2 does not aid in the estimation of θ_1, a "partial redundancy condition" in the terminology of BQSW. Note that $g_1\left(z,\theta\right)$ depends on θ_1 only while g_2 depends on both θ_1 and θ_2. Theorem 8 (p.104) of BQSW applies to this case.

We partition the expected derivative matrix, D, and the variance matrix of the moment conditions, C, so that for $i,j = 1,2$, $D_{ij} = E\left[\partial g_i\left(z,\theta\right)/\partial\theta_j'\right]$ and $C_{ij} = E\left[g_i\left(z,\theta\right)g_j\left(z,\theta\right)'\right]$. Define $G_{21} = D_{21} - C_{21}C_{11}^{-1}D_{11}$ and $\Sigma_{22} = C_{22} - C_{21}C_{11}^{-1}C_{12}$. Then Theorem 8 of BQSW shows that g_2 is (partially) redundant for estimation of θ_1 if and only if

$$G_{21} = D_{22}\left(D_{22}'\Sigma_{22}^{-1}D_{22}\right)^{-1}D_{22}'\Sigma_{22}^{-1}G_{21}. \quad \text{(A8)}$$

Routine calculations yield: $D_{11} = -E(xx')$, $D_{21} = -2E(\epsilon xx')$, $C_{11} = E(\epsilon^2 xx')$, $C_{21} = E(\epsilon^3 xx') - \sigma^2 E(\epsilon xx') = E[\mu_3(x) xx'] - \sigma^2 E(\epsilon xx')$. Then we have

$$G_{21} = -2E(\epsilon xx') + \left[E[\mu_3(x) xx'] - \sigma^2 E(\epsilon xx')\right]\left[E\left(\epsilon^2 xx'\right)\right]^{-1} E(xx'). \quad \text{(A9)}$$

A sufficient condition for (A8) to hold is that $G_{21} = 0$. From (A9) we see that this is so if $E(\epsilon xx') = E[\mu_3(x) xx'] = 0$, which proves proposition 4. We finally note that $G_{21} = 0$ is sufficient but not necessary for (A8) to hold, but we could find nothing informative to say about cases in which (A8) holds with $G_{21} \neq 0$.

A.3 Proof of Proposition 5

Let $\hat{\beta}$ = ordinary least squares, $e_i = y_i - x_i'\hat{\beta} = OLS$ residual, $\hat{\sigma}^2 = \frac{1}{N}\sum_i e_i^2$ = error variance estimate. The RALS estimate is OLS of y_i on $[x_i', w_{i2}]'$ with $w_{i2} = e_i^2 - \hat{\sigma}^2$. As a matter of notation let $H = [X, W_2]$ represent the regressor matrix, and $W_2 = \mathbf{e}^2 - \hat{\sigma}^2 \mathbf{1}_N$, where $\mathbf{e}^2$ $(N \times 1)$ has i^{th} element e_i^2, and $\mathbf{1}_N$ is $N \times 1$ with each element equal to one.

We now define the "estimator"

$$\begin{aligned} \tilde{\beta} &= \hat{\beta} - \frac{\mu_3}{\mu_4 - \sigma^4}(X'X)^{-1}X'W_2 & \text{(A10)} \\ &= \hat{\beta} - \frac{\mu_3}{\mu_4 - \sigma^4}(X'X)^{-1}X'\mathbf{e}^2 + \begin{bmatrix} \frac{\mu_3}{\mu_4 - \sigma^4}\hat{\sigma}^2 \\ \mathbf{O}_{k-1} \end{bmatrix}. \end{aligned}$$

Here $\mathbf{O}_{k-1}$ is a $(k-1)$ dimensional vector of zeros, and the last equality follows if $\mathbf{1}_N$ is the first column of X. The "estimator" is infeasible - it depends on both μ_3 and μ_4 - but we will show that both the RALS and the LGMM estimators are asymptotically equivalent to it.

We start with the RALS estimator, which we write as

$$\dot{\beta} = (X'M_2X)^{-1}(X'M_2Y), \quad \text{(A11)}$$

with $M_2 = I_N - W_2(W_2'W_2)^{-1}W_2'$. We have $\frac{1}{N}X'W_2 \xrightarrow{p} 0$ while $N^{-1/2}X'W_2$ is asymptotically normal. The fact that $\frac{1}{N}X'W_2 \xrightarrow{p} 0$ implies $\dot{\beta}$ has the same asymptotic distribution as $\ddot{\beta}$, defined by

$$\ddot{\beta} = (X'X)^{-1}(X'M_2Y) = \hat{\beta} - (X'X)^{-1}X'W_2(W_2'W_2)^{-1}W_2'Y. \quad \text{(A12)}$$

We then calculate $\text{plim}N^{-1}W_2'W_2 = \mu_4 - \sigma^4$ and $\text{plim}N^{-1}W_2'Y = \text{plim}N^{-1}W_2'\epsilon = \mu_3$. Substituting $\mu_3/(\mu_4 - \sigma^4)$ for $(W_2'W_2)^{-1}W_2'Y$ does not affect the asymptotic distribution of $\ddot{\beta}$, and yields $\tilde{\beta}$ above.

We now turn to the GMM estimator. Again we will consider the LGMM estimator, where in the present case the initial consistent estimate is $\hat{\theta} = \left(\hat{\beta}', \hat{\sigma}^2\right)'$, where $\hat{\beta} = OLS$ and $\hat{\sigma}^2 = \frac{1}{N}\sum_i e_i^2$ (with $e_i = OLS$ residual). We define $\mu_x = E(x)$, $\Sigma_x = E(xx')$. Then we have the following:

$$D = -\begin{bmatrix} \Sigma_x & 0 \\ 0 & \mu_x \end{bmatrix}, \; C = \begin{bmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{bmatrix} \otimes \Sigma_x, \tag{A13a}$$

$$\bar{g}\left(\hat{\theta}\right) = \begin{bmatrix} 0 \\ N^{-1}X'W_2 \end{bmatrix}, \tag{A13b}$$

$$D'C^{-1}D = \frac{1}{DET}\begin{bmatrix} (\mu_4 - \sigma^4)\Sigma_x & -\mu_3\mu_x \\ -\mu_3\mu_x' & \sigma^2\mu_x'\Sigma_x^{-1}\mu_x \end{bmatrix} \equiv \frac{1}{DET}\begin{bmatrix} P & Q \\ R & S \end{bmatrix}, \tag{A13c}$$

$$D'C^{-1}\bar{g}\left(\hat{\theta}\right) = \frac{1}{DET}\begin{bmatrix} \mu_3 N^{-1}X'W_2 \\ -\sigma^2\mu_x'\Sigma_x^{-1}N^{-1}X'W_2 \end{bmatrix}, \tag{A13d}$$

where $DET = |C|$ will cancel later. The LGMM estimator is $\tilde{\theta} = \hat{\theta} - (D'C^{-1}D)^{-1} D'C^{-1}\bar{g}\left(\hat{\theta}\right)$. We are interested in $\tilde{\beta}$, so we need to use partitioned inversion. In terms of partitioning in (A13c), we have

$$J \equiv S - RP^{-1}Q = \frac{\sigma^2\mu_4 - \sigma^6 - \mu_3^2}{\mu_4 - \sigma^4}\mu_x'\Sigma_x^{-1}\mu_x, \tag{A14a}$$

$$\left(D'C^{-1}D\right)^{11} = \frac{1}{DET}\left(P^{-1} + P^{-1}QJ^{-1}RP^{-1}\right), \tag{A14b}$$

$$= \frac{1}{DET}\frac{1}{\mu_4 - \sigma^4}\left[\Sigma_x^{-1} + \frac{\mu_3^2}{\sigma^2\mu_4 - \sigma^6 - \mu_3^2}\Sigma_x^{-1}\mu_x\left(\mu_x'\Sigma_x^{-1}\mu_x\right)^{-1}\mu_x'\Sigma_x^{-1}\right],$$

$$\left(D'C^{-1}D\right)^{12} = -\frac{1}{DET}P^{-1}QJ^{-1} = \frac{1}{DET}\frac{\mu_3}{\sigma^2\mu_4 - \sigma^6 - \mu_3^2}\Sigma_x^{-1}\mu_x\left(\mu_x'\Sigma_x^{-1}\mu_x\right)^{-1} \tag{A14c}$$

Then the LGMM estimator is

$$\begin{aligned} \bar{\beta} &= \hat{\beta} - \left(D'C^{-1}D\right)^{11} \cdot [\text{first block of } D'C^{-1}\bar{g}\left(\hat{\theta}\right)] \\ &\quad - \left(D'C^{-1}D\right)^{12} \cdot [\text{second block of } D'C^{-1}\bar{g}\left(\hat{\theta}\right)]. \end{aligned} \tag{A15}$$

Using (A14b), (A14c), (A13d) and a lot of routine but messy algebra, we obtain

$$\bar{\beta} = \hat{\beta} - \frac{\mu_3}{\mu_4 - \sigma^4}\Sigma_x^{-1}N^{-1}X'W_2 + \frac{\mu_3}{\mu_4 - \sigma^4}\Sigma_x^{-1}\mu_x\left(\mu_x'\Sigma_x^{-1}\mu_x\right)^{-1}\mu_x'\Sigma_x^{-1}N^{-1}X'W_2. \tag{A16}$$

To simplify this, we observe that, if the first column of X is an intercept, then the first column of Σ_x equals μ_x. Then $\Sigma_x^{-1}\mu_x$ equals the first column of identity, and $\mu_x'\Sigma_x^{-1}\mu_x = 1$. Thus the matrix $\Sigma_x^{-1}\mu_x\left(\mu_x'\Sigma_x^{-1}\mu_x\right)^{-1}\mu_x'\Sigma_x^{-1}$ equals zero except for a "one" in the 1,1 position. Since the first element of $X'W_2$ equals zero if there is an intercept $\left[\sum_i\left(e_i^2 - \hat{\sigma}^2\right) = 0\right]$, the last term in (A16) equals zero.

To obtain $\tilde{\beta}$ in (A10), we simply replace Σ_x in the second term in (A16) by $N^{-1}X'X$, which does not affect the asymptotic distribution of the estimator. Then the N^{-1} terms cancel and we obtain (A10).

A.4 Proof of Propositions 6 and 7

We follow the method of analysis of Chamberlain (1987). Suppose that the conditional moment restrictions are of the form $E[g(z,\theta) \mid x] = 0$. Define $D(x) = E[\partial g(z,\theta)/\partial\theta' \mid x]$ and $C(x) = E\left[g(z,\theta)\, g(z,\theta)' \mid x\right]$. Then the efficient GMM estimator based on the conditional moment restrictions can be calculated as the GMM estimator based on the unconditional moment restrictions

$$E\left[D(x)' C(x)^{-1} g(z,\theta)\right] = 0. \tag{A17}$$

We now proceed to calculate these quantities for the case of the conditional moment restrictions (7A)-(7B). We have

$$\begin{aligned} g(z,\theta) &= \begin{bmatrix} y - x'\beta \\ (y - x'\beta)^2 - \sigma^2 \end{bmatrix},\ D(x) = \begin{bmatrix} -x' & 0 \\ 0 & -1 \end{bmatrix}, \\ C(x) &= \begin{bmatrix} \sigma^2 & \mu_3(x) \\ \mu_3(x) & \mu_4(x) - \sigma^4 \end{bmatrix}, \end{aligned} \tag{A18}$$

where as before $\mu_j(x) = E\left[(y - x'\beta)^j \mid x\right]$, $j = 3, 4$. Define the determinant $\Delta(x) = |C(x)| = \sigma^2[\mu_4(x) - \sigma^4] - \mu_3(x)^2$. Then we calculate

$$D(x)'C(x)^{-1}g(z,\theta) = \frac{1}{\Delta(x)} \begin{bmatrix} -[\mu_4(x) - \sigma^4]\, x\,(y - x'\beta) + \mu_3(x)\, x\left[(y - x'\beta)^2 - \sigma^2\right] \\ \mu_3(x)(y - x'\beta) - \sigma^2\left[(y - x'\beta)^2 - \sigma^2\right] \end{bmatrix}. \tag{A19}$$

The efficient GMM estimators come from these unconditional moment conditions. To prove proposition 6, we simply note that, when $\mu_3(x) = 0$, $\Delta(x) = \sigma^2[\mu_4(x) - \sigma^4]$ and we obtain

$$E\begin{bmatrix} -x(y - x'\beta)/\sigma^2 \\ -\left[(y - x'\beta)^2 - \sigma^2\right]/\Delta(x) \end{bmatrix} = 0. \tag{A20}$$

This is an exactly identified set of moment conditions and so the estimators (say $\tilde\beta, \tilde\sigma^2$) satisfy the sample version exactly. In particular, $\tilde\beta$ must satisfy

$$\sum_{i=1}^{N} x_i\left(y_i - x_i'\tilde\beta\right) = 0, \tag{A21}$$

the solution to which is obviously least squares.

To prove Proposition 7, we observe that, if x contains an intercept and $\mu_3(x)$ and $\mu_4(x)$ do not depend on x, so that $\Delta(x)$ does not depend on x, the moment conditions (A20) are implied by (6A) and (6B).

A.5 Proof of Proposition 12

Let g_1, g_2 and g_3 represent the moments in (6A), (6B) and (6C). According to BQSW (1999, Theorem 2), g_2 and g_3 are redundant given g_1 iff g_2 is redundant given g_1 and g_3 is redundant given g_1. Proposition 4 says that g_2 is redundant given g_1 if $E(\epsilon x x') = E(\mu_3(x) x x') = 0$, and Proposition 8 says that g_3 is redundant given g_1 if (10) holds. Combining these conditions, we obtain (11).

A.6 Proof of Propositions 13 and 14

We now let $g_1(z,\theta) = (y - x'\beta)$, $g_2(z,\theta) = \left[(y-x'\beta)^2 - \sigma^2\right]$, and $g_3(z,\theta) = \left[(y-x'\beta)^3 - \mu_3\right]$, with $\theta = (\beta', \sigma^2, \mu_3)'$. Using the same generic notation as in section A.4, we have

$$D(x) = \begin{bmatrix} -x' & 0 & 0 \\ 0 & -1 & 0 \\ -3\sigma^2 x' & 0 & -1 \end{bmatrix}, \tag{A22a}$$

$$C(x) = \begin{bmatrix} \sigma^2 & \mu_3 & \mu_4(x) \\ \mu_3 & \mu_4(x) - \sigma^4 & \mu_5(x) - \sigma^2\mu_3 \\ \mu_4(x) & \mu_5(x) - \sigma^2\mu_3 & \mu_6(x) - \mu_3^2 \end{bmatrix}. \tag{A22b}$$

Again following Chamberlain (1987), the efficient GMM estimator is based on the unconditional moment conditions $ED(x)' C(x)^{-1} g(z,\theta) = 0$. Suppressing, for the moment, the dependence of C on x, and of g_1, g_2 and g_3 on z and θ, we obtain

$$\begin{aligned} & D(x)' C(x)^{-1} g(z,\theta) \qquad\qquad (A23) \\ = & \begin{bmatrix} (C^{11} + 3\sigma^2 C^{31}) x g_1 + (C^{12} + 3\sigma^2 C^{32}) x g_2 + (C^{13} + 3\sigma^2 C^{33}) x g_3 \\ C^{21} g_1 + C^{22} g_2 + C^{23} g_3 \\ C^{31} g_1 + C^{32} g_2 + C^{33} g_3 \end{bmatrix}, \end{aligned}$$

where C^{ij} denotes the i, j^{th} element of C^{-1}.

Inspection of (A23) reveals that the first block reduces to $xg_1 = x(y - x'\beta)$, and implies that the estimate of β is OLS, if the following conditions hold:

$$C^{11}(x) + 3\sigma^2 C^{31}(x) \text{ does not depend on } x, \tag{A24a}$$

$$C^{12}(x) + 3\sigma^2 C^{32}(x) = 0, \tag{A24b}$$

$$C^{13}(x) + 3\sigma^2 C^{33}(x) = 0. \tag{A24c}$$

We impose the conditions of Proposition 12: $\mu_3(x) = 0$ and $\mu_4(x) = 3\sigma^4$, so that

$$C(x) = \begin{bmatrix} \sigma^2 & 0 & 3\sigma^4 \\ 0 & 2\sigma^4 & \mu_5(x) \\ 3\sigma^4 & \mu_5(x) & \mu_6(x) \end{bmatrix}. \tag{A25}$$

Then tedious calculation yields:

$$C^{11}(x) = \left[2\sigma^4\mu_6(x) - \mu_5(x)^2\right]/\mathrm{Det}(x), \tag{A26a}$$
$$C^{12}(x) = 3\sigma^4\mu_5(x)/\mathrm{Det}(x), \tag{A26b}$$
$$C^{13}(x) = -6\sigma^8/\mathrm{Det}(x), \tag{A26c}$$
$$C^{32}(x) = -\sigma^2\mu_5(x)/\mathrm{Det}(x), \tag{A26d}$$
$$C^{33}(x) = 6\sigma^6/\mathrm{Det}(x), \tag{A26e}$$
$$\mathrm{Det} = \sigma^2\left[2\sigma^4\mu_6(x) - \mu_5(x)^2\right] - 18\sigma^2. \tag{A26f}$$

Finally, it is easy to verify that $C^{11}(x) + 3\sigma^2 C^{31}(x) = \frac{1}{\sigma^2}$, which does not depend on x, while $C^{12}(x) + 3\sigma^2 C^{32}(x) = C^{13}(x) + 3\sigma^2 C^{33}(x) = 0$. Thus, (A24) is satisfied.

To prove Proposition 14, we note that the conditional moment restrictions (7A), (7B) and (7C), plus the condition that μ_4, μ_5 and μ_6 do not depend on x, ensure that $C(x)$ does not depend on x. Then, if the regression contains an intercept, the moment conditions (A23) are a linear combination of the moment conditions in (6A), (6B) and (6C).

A.7 Proof of Proposition 15

We will show that the asymptotic variance of the RALS estimator is the same as that of the GMM estimator. From Proposition 14, we can compare the asymptotic variance of RALS to the asymptotic variance of the GMM estimator using the moment restrictions (6A), (6B) and (6C).

Let $g_1 = x_i(y_i - x_i'\beta)$, $g_2 = x_i\left[(y_i - x_i'\beta)^2 - \sigma^2\right]$, $g_3 = x_i\left[(y_i - x_i'\beta)^3 - \mu_3\right]$, $g = (g_1', g_2', g_3')'$, $\theta = (\beta', \sigma^2, \mu_3)'$, $D = E(\partial g/\partial\theta')$ and $C = E[gg']$. Then the asymptotic variance of the GMM estimator, $\tilde{\theta}$, using the moment restrictions (6A), (6B) and (6C) is

$$Avar\sqrt{N}\tilde{\theta} = \left(D'C^{-1}D\right)^{-1}. \tag{A27}$$

Let $x_i = (1, x_{i*}')'$, $\mu_* = E(x_{i*})$ and $V_* = V(x_{i*})$, and

$$d = \begin{bmatrix} 1 \\ 0 \\ 3\sigma^2 \end{bmatrix} \text{ and } c = \begin{bmatrix} \sigma^2 & \mu_3 & \mu_4 \\ \mu_3 & \mu_4 - \sigma^2 & \mu_5 - \sigma^2\mu_3 \\ \mu_4 & \mu_5 - \sigma^2\mu_3 & \mu_6 - \mu_3^2 \end{bmatrix}.$$

Then, from (A27), after lengthy but straightforward algebra (available from the authors), we obtain for the GMM estimator of β :

$$Avar\left(\sqrt{N}\tilde{\beta}\right) = \begin{bmatrix} \frac{1}{\sigma^2} & \frac{1}{\sigma^2}\mu_*' \\ \frac{1}{\sigma^2}\mu_* & \left(d'c^{-1}d\right)V_* + \frac{1}{\sigma^2}\mu_*\mu_*' \end{bmatrix}^{-1}. \tag{A28}$$

Letting $\beta = (\alpha, \beta_*')'$, we have, for the GMM estimator of the slope coefficients β_*,

$$Avar\left(\sqrt{N}\tilde{\beta}_*\right) = \frac{1}{d'c^{-1}d}V_*^{-1}. \tag{A29}$$

Straightforward calculation gives

$$\frac{1}{d'c^{-1}d} = \tag{A30}$$
$$\sigma^2 - \frac{\mu_3^2\left(\mu_6 - \mu_3^2 - 6\sigma^2\mu_4 + 9\sigma^6\right) - 2\mu_3(\mu_4 - 3\sigma^4)\left(\mu_5 - 4\sigma^2\mu_3\right) + (\mu_4 - 3\sigma^4)^2(\mu_4 - \sigma^4)}{\left(\mu_4 - \sigma^4\right)\left(\mu_6 - \mu_3^2 - 6\sigma^2\mu_4 + 9\sigma^6\right) - \left(\mu_5 - 4\sigma^2\mu_3\right)^2}.$$

And we have, for the GMM estimator of the intercept,

$$Avar\left(\sqrt{N}\tilde{\alpha}\right) = \sigma^2 + \frac{1}{d'c^{-1}d}\mu_*'V_*^{-1}\mu_* \tag{A31}$$

Next we derive the asymptotic variance of the RALS estimator of the slope coefficient β_*. Let $\ddot{x}_i' = x_{i*} - \bar{x}_*$, where $\bar{x}_* = \frac{1}{N}\sum_i x_{i*}$. Then, the RALS estimator of β_* is obtained from a regression y_i on $(\ddot{x}_i', w_{2i}, w_{3i})'$, where $w_{2i} = e_i^2 - \hat{\sigma}^2$ and $w_{3i} = e_i^3 - m_3 - 3\hat{\sigma}^2 e_i$, defined in Proposition 5 and 9. Let $w_i = (w_{2i}, w_{3i})$. Then, we have for the RALS estimator $\dot{\beta} = [X'MX]^{-1} X'MY$, where $M = I_N - W(W'W)^{-1}W'$ and W is the $N \times 2$ matrix with w_i at the i^{th} row. Therefore,

$$\sqrt{N}\left(\dot{\beta}_* - \beta_*\right) = \tag{A32}$$
$$\left(\frac{1}{N}\sum_i \ddot{x}_i\ddot{x}_i'\right)^{-1}\left[\frac{1}{\sqrt{N}}\sum_i \ddot{x}_i\epsilon_i - \frac{1}{\sqrt{N}}\sum_i \ddot{x}_i w_i\left(\frac{1}{N}\sum_i w_i'w_i\right)^{-1}\frac{1}{N}\sum_i w_i'\epsilon_i\right] + o_p(1).$$

Consider the β estimator obtained by regressing y_i on $(x_i', \omega_i)'$, where $\omega_i = (\omega_{2i}, \omega_{3i})$ with $\omega_{2i} = \epsilon_i^2 - \sigma^2$ and $\omega_{3i} = \epsilon_i^3 - \mu_3 - 3\sigma^2\epsilon_i$. Let $\ddot{\beta}$ be the resulting estimator of β. Then we have the asymptotic variance

$$Avar\sqrt{N}\left(\ddot{\beta} - \beta\right) = \left(\sigma^2 - \Sigma_{\omega\epsilon}'\Sigma_{\omega\omega}^{-1}\Sigma_{\omega\epsilon}\right)E\left(xx'\right)^{-1}, \tag{A33}$$

where $\Sigma_{\omega\epsilon} = E\left(\omega'\epsilon\right)$, $\Sigma_{\omega\omega} = E\left(\omega'\omega\right)$. A straightforward calculation yields

$$\sigma^2 - \Sigma_{\omega\epsilon}'\Sigma_{\omega\omega}^{-1}\Sigma_{\omega\epsilon} = \frac{1}{d'c^{-1}d}. \tag{A34}$$

which is given in (A30). Now the proof for the slope coefficients β_* is complete if we show

$$\frac{1}{\sqrt{N}}\sum_i \ddot{x}_i\left(w_i - \omega_i\right) = o_p(1). \tag{A35}$$

But, it is shown in White (1980, Appendix) that

$$\frac{1}{\sqrt{N}}\sum_i \ddot{x}_i\left(w_{2i}-\omega_{2i}\right)=\frac{1}{\sqrt{N}}\sum_i \ddot{x}_i\left[\left(e_i^2-\epsilon_i^2\right)-\left(\hat{\sigma}^2-\sigma^2\right)\right]=o_p\left(1\right). \quad \text{(A36)}$$

Also, noting that $e_i=\epsilon_i-x_i'\left(\hat{\beta}-\beta\right)$, where $\hat{\beta}$ is the OLS estimator of β, we have

$$\begin{aligned} & \frac{1}{\sqrt{N}}\sum_i \ddot{x}_i\left(w_{3i}-\omega_{3i}\right) \\ = & \frac{1}{\sqrt{N}}\sum_i \ddot{x}_i\left\{\left[e_i^3-\epsilon_i^3\right]-3\left(\hat{\sigma}^2 e_i-\sigma^2\epsilon_i\right)-\left(m_3-\mu_3\right)\right\} \qquad \text{(A37)} \\ = & \frac{1}{\sqrt{N}}\sum_i \ddot{x}_i\left\{\left[-3\epsilon_i^2 x_i'\left(\hat{\beta}-\beta\right)+3\hat{\sigma}^2 x_i'\left(\hat{\beta}-\beta\right)\right]-3\left(\hat{\sigma}^2-\sigma^2\right)\epsilon_i\right\}+o_p\left(1\right) \\ = & o_p\left(1\right), \end{aligned}$$

which completes the proof for the slope estimators.

For the intercept estimator of RALS we have

$$\begin{aligned} \sqrt{N}\dot{\alpha} &= \sqrt{N}\left(\bar{y}-\bar{x}_*'\dot{\beta}_*\right) \qquad \text{(A38)} \\ &= \sqrt{N}\left[\left(\bar{y}-\bar{x}_*'\beta_*\right)-\bar{x}_*'\left(\dot{\beta}_*-\beta_*\right)\right], \end{aligned}$$

from which it is easy to deduce that

$$\begin{aligned} Avar\sqrt{N}\dot{\alpha} &= \sigma^2+\mu_*'\left[Avar\sqrt{N}\dot{\beta}\right]\mu_* \\ &= \sigma^2+\frac{1}{d'c^{-1}d}\mu_*'V_*^{-1}\mu_*, \qquad \text{(A39)} \end{aligned}$$

which is what we have in (A31).

References

[1] Bickel, P. (1982), "On Adaptive Estimation," *Annals of Statistics*, 10, 647-671.

[2] Breusch, T., H. Qian, P. Schmidt and D. J. Wyhowski (1999), "Redundancy of Moment Conditions," *Journal of Econometrics*, 91, 89-112.

[3] Chamberlain, G.C. (1987), "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305-335.

[4] Hsieh, D.A. and C.F. Manski (1987), "Monte Carlo Evidence on Adaptive Maximum Likelihood Estimation of a Regression," *Annals of Statistics*, 15, 541-551.

[5] Jambunathan, M. V. (1954), "Some Properties of Beta and Gamma Distributions," *Annals of Mathematical Statistics*, 25, 401-405.

[6] Johnson, N. L. and S. Kotz (1970), *Distributions in Statistics: Continuous Univariate Distributions-2*, Wiley, New York.

[7] MaCurdy, T.E. (1982), "Using Information on the Moments of Disturbances to Increase the Efficiency of Estimation," NBER Technical Paper 22, Cambridge, MA.

[8] Manski, C.F. (1984), "Adaptive Estimation of Non-linear Regression Models," *Econometric Reviews*, 3, 145-194.

[9] Newey, W.K. (1988), "Adaptive Estimation of Regression Models via Moment Restrictions," *Journal of Econometrics*, 38, 301-339.

[10] Newey, W.K. (1993), "Efficient Estimation of Models with Conditional Moment Restrictions," Chapter 16 in *Handbook of Statistics*, Volume 11, G.S. Maddala, C.R. Rao and H.D. Vinod, eds., Elsevier Science Publishers, New York.

[11] Qian, H. and P. Schmidt (1999), "Improved Instrumental Variables and Generalized Method of Moments Estimators," *Journal of Econometrics*, 91, 145-170.

[12] Wooldridge, J.M. (1993), "Efficient Estimation with Orthogonal Regressors," Econometric Theory, 9, 687.

[13] White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 725-746.