Making estimators aware of heteroskedasticity

Alecos Papadopoulos[*] & Mike G. Tsionas[†]

June 3, 2019

Abstract

In pursuit of efficiency, we propose a new way to construct least squares estimators, as the minimizers of an augmented objective function that takes into account selected properties of the error term. In this paper we focus on heteroskedasticity. We initially derive an infeasible estimator which we then approximate using residuals from a first-step regression to obtain the feasible HOLS estimator. This estimator is consistent and outperforms Ordinary Least Squares in terms of finite-sample Mean Squared Error and asymptotic efficiency in the case of heteroskedasticity, unconditional or conditional on the regressors, but also under homoskedasticity. Theoretical results are accompanied by simulations that support them.

Keywords: Heteroskedasticity, least squares, efficiency, excess kurtosis.

JEL classification: C13, C21, C31.

[*] Athens University of Economics and Business, Greece, papadopalex@aueb.gr

[†] Lancaster University Management School, LA1 4YX, U.K., m.tsionas@lancaster.ac.uk

1 Introduction and motivation

In the context of linear regression and least squares (LS) estimation we face two unknown entities: the coefficient vector and the error term. In order to proceed with estimation and inference on them we make certain assumptions on the error term as regards its statistical properties, like having a zero mean, or a constant or not variance, and maybe a distributional assumption. Regarding the coefficient vector, sometimes we may impose restrictions directly on its elements based on out-of-sample information, in which case we execute Restricted or Inequality Restricted Least Squares.

Least-squares estimation of the coefficient vector may proceed without regard for the properties of the other unknown, the error term, since least squares is a mathematical approximation method that does not need a statistical foundation (for the occasional conflicts this may create see Spanos 2010). After estimation, the statistical properties of the error term co-determine the statistical properties of the LS estimator, and so of statistical inference.

An important example where the properties of the error term enter the picture after estimation of the coefficient vector, is "heteroskedasticity-robust" Ordinary Least Squares (OLS). We can obtain OLS point estimates without factoring in heteroskedasticity, which we can subsequently take into account only for estimating the variance of the estimator. This is the approach initiated by White (1980), who, as Romano and Wolf (2017) put it, "changed the game with one of the most influential and widely-cited papers in econometrics". MacKinnon (2013) offers a recent assessment of how critical was the influence of this paper on theoretical advances in econometrics, while it is also true that White's approach has come to dominate empirical practice to the degree that in many cases practitioners don't even test for the existence of heteroskedasticity but use "heteroskedasticity-robust" standard errors regardless. This has asymptotic justification because the White variance-covariance matrix estimator is consistent and non-parametric, and it will converge to the true matrix whether we have conditional homoskedasticity, unconditional heteroskedasticity, or heteroskedasticity conditional on the regressors (subject to assumptions that guarantee that the relevant limiting magnitudes exist and are finite).

In contrast, and for the case of conditional heteroskedasticity, the Generalized LS (GLS) and the Weighted LS (WLS) estimators are examples of estimators that attempt to factor in heteroskedasticity and let it affect the point estimates themselves, and not just the variance of the estimator. They have fallen relatively out of use, but Romano & Wolf (2017), criticizing White's approach due to its unsatisfactory performance in "small to moderate samples" as they write, propose to combine WLS estimation with White's method as follows: first test for the presence of conditional heteroskedasticity. If the test indicates that the phenomenon exist, use WLS *together* with "heteroskedasticity-robust" standard errors (based on the WLS estimator).

In the present work we aim to obtain a consistent LS estimator that outperforms OLS in terms of efficiency both in finite samples and asymptotically. To achieve that, we take a new route that cross-breeds the two approaches mentioned but in a different way than Romano & Wolf's suggestion: we link the two unknown entities of the regression setup by making the estimator of the coefficient vector "sensitive" to properties of the error term, specifically to the *possible* existence of heteroskedasticity. We do this through an augmented objective function that the estimator minimizes, without assuming any particular level or form that heteroskedasticity may take, allowing also for the case of homoskedasticity. In this way both the estimates and the variance of the estimator are affected as in the GLS and WLS cases, but at the same time we treat heteroskedasticity non-parametrically, as in White's approach. Admittedly, we are guilty of the same sin that one could accuse White's approach for: that it treats heteroskedasticity conditional on the regressors as a hurdle to be sidestepped, and not as a problem that at the same time is an opportunity. Conditioning reduces the probability space, there is information in conditioning that could improve our inference as regards prediction.

Our estimator is obtained in two steps: in the first step we use OLS residuals to calculate both an efficiency-optimization parameter *and* to transform the dependent variable, and in the second step the estimator is obtained as OLS applied on the transformed sample. For this reason we will label it "HOLS". HOLS is a consistent estimator under mild conditions, and it dominates OLS in terms of asymptotic efficiency and finite-sample Mean Squared Error (MSE). Importantly, the efficiency gains exist *even under homoskedasticity*, making the HOLS estimator a general-purpose tool not confined to be used only when heteroskedasticity is thought to be present.

The only partial precursor of our work that we were able to find in the literature is Gourieroux, Monfort & Renault (1996). They focus on the case of heteroskedasticity conditional on the regressors, and through a different route they obtain some results that we also arrived at. But their estimator is confined to work with symmetric error distributions, while the HOLS estimator can accomodate error skewness. Also, they develop a very general framework of GMM and Instrumental Variables estimation that produces a rather complicated environment not very friendly for wide empirical adoption. In contrast, the HOLS estimator is very easily and transparently implemented. We will discuss some literature connections and contrasts once we have presented our estimator so that the reader has a more concrete picture of what we are attempting here.

The rest of the paper is structured as follows: in Section 2 we develop the model and obtain the expression for the minimizer of the objective function. In Section 3 we derive the general form and properties of the HOLS estimator. This is a preparatory section to a degree, since specific results as regards efficiency depend on what the skedastic assumption will be. Section 4 examines the case of homoskedasticity and hosts the first efficiency battle of our estimator against OLS (HOLS wins). In Sections 5 and 6 we examine the cases of unconditional and conditional heteroskedasticity respectively, and the results remain favorable to our estimator. In Section 7 we determine how we can estimate consistently the HOLS variance-covariance matrix, and we wrap-up by summarizing the simple HOLS implementation routine. Section 8 concludes with a summary of the findings and some suggestions for future research from the many topics one could explore. The paper is accompanied by two supplementary files : Supplement I contains extensions and generalizations of the HOLS estimator while Supplement II has the various mathematical derivations and details on the Monte Carlo simulations.

2 Model set up and optimization of the objective function

We consider the model,

$$y = \mathrm{X}\beta + u, \ \text{ or } \ y_i = x_i'\beta + u_i, \ \ i = 1, ..., n$$
$$E(u) = 0, \ \ E(\mathrm{X}'u) = 0, \ \ E\left(|u_i|^{6+\delta}\right) < \infty, \ \ \delta > 0. \tag{1}$$

The column vectors y and u are $n \times 1$, X is a $n \times (K+1)$ matrix of regressors including a constant, traditionally in the first column, and β is a $(K+1) \times 1$ vector. Later we will consider also a centered regression setup, in which case the same symbols will denote variables centered on their sample means, the regressor matrix will not include a constant, and the coefficient vector will be a $K \times 1$ column vector. Lowercase symbols represent either a vector or a single variable, which should be clear from context. The regressors satisfy the Grenander conditions and are i.i.d. The error terms are independently distributed, and they may be either identically distributed or not, allowing for heteroskedasticity. The condition related to the 6th absolute error moment (needed to secure the existence of the asymptotic variance of the estimators we will consider), restricts somewhat the error distributions that can be accomodated since for example, if we want to allow for Student's-t errors we must assume integer degrees of freedom no less than seven under this assumption. Still, most of the distributions usually assumed for the error term satisfy this assumption, so we do not consider it as problematic.

We want to construct a general-purpose least-squares estimator that will take into account the *possible* existence of heteroskedasticity, in a direct way and not only as regards the estimator's variance-covariance matrix. We achieve that by augmenting the objective function that the LS estimator optimizes by a measure of heteroskedasticity. One way to reflect the "degree" or the "intensity" of heteroskedasticity is by the sample variance of

u_i^2:

$$H_u \equiv \frac{1}{n}\sum_{i=1}^{n}\left(u_i^2 - \frac{1}{n}\sum_{i=1}^{n}u_i^2\right)^2 = \frac{1}{n}\sum_{i=1}^{n}u_i^4 - \left(\frac{1}{n}\sum_{i=1}^{n}u_i^2\right)^2 .$$

This choice is inspired by the Breusch-Pagan (1979) test for conditional heteroskedasticity. The vector containing the elements $\{u_i^2 - n^{-1}\sum_{i=1}^{n}u_i^2,\ i = 1, ..., n\}$ (estimated from OLS residuals) is the component of the LM statistic of that test that measures the variability of the error term, and the higher it is (in the vector sense) the more likely it is that the null hypothesis of homoskedasticity will be rejected. Returning to H_u, if in reality the error term is homoskedastic, we have $H_u \xrightarrow{p} m_4 - \sigma_u^4$ (where m_4 is the 4th raw moment of the error term and σ_u^4 is the square of its variance), an expression that equals the variance of the limiting distribution of $\sqrt{n}\,(\hat{\sigma}_u^2 - \sigma_u^2)$, i.e. of the sample variance from an i.i.d. sample. So if homoskedasticity is what actually holds in the data, the quantity H_u reflects the uncertainty surrounding the common but unknown error variance.

If the error term is unconditionally heteroskedastic, we will assume that average raw moments have a finite limit. Then, we have $H_u \xrightarrow{p} \bar{m}_4 - \bar{\sigma}^4$,

$$\bar{m}_4 = \text{plim}\left(n^{-1}\sum_{i=1}^{n}u_i^4\right) < \infty,\ \ \bar{\sigma}^4 = \text{plim}\left(n^{-1}\sum_{i=1}^{n}u_i^2\right)^2 = \left(\bar{\sigma}^2\right)^2 < \infty,$$

Turning to our optimization problem, perhaps the first thought would be to formulate a constrained minimization task imposing some upper bound on the degree of heteroskedasticity, in an attempt to force the least-squares estimator to smooth the heteroskedasticity effect. This would move our approach near the universe of Tikhonov regularization and its various incarnations in statistics. But such an approach would ignore the fact that heteroskedasticity, if it exists, is a property of the data generating mechanism and imposing an upper bound would be artificial. To avoid this we just allow for its existence and at an unknown level (possibly zero) and we formulate a constrained minimization problem with an *equality* constraint. We construct therefore our LS estimator as the minimizer in the following problem:

$$\min_{\beta}\frac{1}{2n}\sum_{i=1}^{n}u_i^2 \quad s.t.\ \frac{1}{4n}\sum_{i=1}^{n}\left(u_i^2 - n^{-1}\sum_{i=1}^{n}u_i^2\right)^2 = \frac{1}{4}\bar{H}_u\,, \tag{2}$$

for some fixed unknown value $\bar{H}_u$. This leads to the Lagrangian, in canonical form,

$$\Lambda = \frac{1}{2n}\sum_{i=1}^{n} u_i^2 \ + \ \lambda\left[\frac{1}{4}\bar{H}_u \ - \ \frac{1}{4n}\sum_{i=1}^{n}\left(u_i^2 - n^{-1}\sum_{i=1}^{n} u_i^2\right)^2\right], \ \ \lambda \in \mathrm{R},$$

where the Lagrange multiplier λ is not constrained to be non-negative (and we will see that there are cases where its optimal value is negative). With this objective function we quantify the possibility of heteroskedasticity and of changes in it. This is why we call the estimator "heteroskedasticity-conscious": it estimates the coefficient vector by taking into account the *possible* existence of heteroskedasticity of an *unknown form and degree.* What is impotant to keep in mind is that our goal is statistical, not mathematical: we want to obtain a consistent estimator with as low variance as possible: therefore, even though we will go through the motions of function minimization, we will eventually choose the optimal value for λ so as to minimize the (asymptotic) variance of the estimator (given consistency). This means that the actual and unknown value of $\bar{H}_u$ plays no role here.

The above formulation of the objective function leads to an estimator that is in a category of its own: it is not “Restricted Least Squares” as the term is used in the literature because the constraint is not related to the unknown parameters directly. It is not “Regularized Least Squares” (Ridge regression, LASSO) because we neither tweak the inverse regressor moment matrix nor do we impose an inequality constraint on the coefficient vector. Also, it cannot be meaningfully formulated as Generalized Method of Moments, because the constraint we impose does not have a pre-assigned value: $\bar{H}_u$ remains an unknown quantity, and since it enters additively in a single constraint, it would not affect any “estimating equations” solution even if we treated it as an unkown to be optimally determined. Our approach can be generalized to take into account other possibe properties of the error term, and we provide such a generalization in Supplement I as a springboard for furhter research.

One could question whether such an estimator is conceptually more valid than OLS, or even if it is somehow misleading in a fundamental way. In what sense should we allow the point estimates to be affected because heteroskedasticity *may* exist? Why should we factor into the estimates the possible changes in the degree of heteroskedasticity? And, in the homoskedastic case, why should the point estimates be affected because the error variance is unknown and estimated with uncertainty?

The answer is that in this way we allow the estimator to take into account simultaneously the neighborhood of data-generating mechanisms surrounding the homoskedastic case. We don't really know whether heteroskedasticity exists, and if it exists, we don't really know its form or its intensity. The HOLS estimator takes these uncertainties into account, although without estimating heteroskedasticity per se. In the homoskedastic case, the estimator is forced to recognize the uncertainty in the estimation, something that makes it an *inherently statistical* tool rather than a mathematical approximation method that is coated with statistical properties.

Simplifying the Lagrangian from terms that do not affect the first-order conditions, this least squares estimator can be derived as

$$\hat{\beta} = \arg\min \left\{ \frac{1}{2} \sum_{i=1}^{n} u_i^2 \ - \ \frac{\lambda}{4} \sum_{i=1}^{n} \left(u_i^2 - n^{-1} \sum_{i=1}^{n} u_i^2 \right)^2 \right\}. \tag{3}$$

From a mathematical point of view, this minimization problem has a convex objective function but its equality constraint is not affine and so this is not a standard convex optimization problem. It appears, then, that we should also look at the general second-order conditions for a minimum, namely that the principal minors of the bordered Hessian are all negative. But, again given our statistical interests, it is really of no interest or consequence whether the stationary point of the above problem actually minimizes the objective function, as long as our goal of consistency and minimum variance is accomplished.

Calculating the first-order condition, we obtain (Supplement II-A.1),

$$\hat{\beta} \ = \ (\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{X}'y \ - \ \lambda \left[1 + \lambda \left(n^{-1} \sum_{i=1}^{n} u_i^2 \right) \right]^{-1} (\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{X}'u^{(3)},$$

where $u^{(3)}$ denotes the Hadamard matrix product (element-by-element multiplication), i.e. it is a vector that holds the 3d power of each element of u. Moreover, we set for compactness

$$\lambda\left[1+\lambda\left(n^{-1}\sum_{i=1}^{n}u_i^2\right)\right]^{-1}\equiv\alpha_n\xrightarrow{p}\alpha<\infty.$$

The limit α exists because if the error term is homokedastic it becomes i.i.d. and hence ergodic, while if it is unconditionally heteroskedastic we have already assumed that the limits of average moments exist and are finite. We will henceforth occupy ourselves with the efficiency parameter α_n and α rather than with λ, since this is more pertinent to our statistical aims here. Then,

$$\hat{\beta}\;=\;(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'y\;-\;\alpha_n(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'u^{(3)}. \tag{4}$$

The right-hand side expression cannot be computed since it includes the unknown error term. So eq.(4) describes an infeasible estimator, the Platonic ideal of HOLS, which plays the same role that the infeasible GLS estimator plays to the Feasible GLS. We call this ideal estimator PHLS and for completeness we study it in Supplement I, while for contrasting purposes we mention some results related to it in the main text.

One way to arrive at a feasible estimator is to treat the error components of the right-hand side as functions of the coefficient vector which, again under a mathematical point of view, would be the anticipated way to proceed. This estimator is also explored in Suppementary File I, but in the next section we present a simpler method that proved to perform better in statistical terms.

3 The HOLS Estimator and its properties

To arrive at the HOLS estimator we estimate an OLS regression and plug the OLS residuals in the right-hand-side of eq.(4) as an approximation to the true errors. We will also use the OLS residuals to estimate the (asymptotically) optimal efficiency parameter α, $\hat{\alpha}^*$, so the HOLS estimator is defined as

$$\hat{\beta}_{HOLS}\;=\;(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'y\;-\;\hat{\alpha}^*(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\,\hat{u}_{OLS}^{(3)}\,. \tag{5}$$

Note that $\mathrm{X}'\,\hat{u}^{(3)}_{OLS} \neq 0$, so the estimator is meaningful to compute. The optimal value of the efficiency parameter depends on the assumption made on the skedasticity of the error term and we will derive it in the relevant Sections. We turn to examine the main properties of the estimator.

Unbiasedness. This estimator is biased since for unbiasedness we would need, in addition to the main assumptions in (1),

$$E\left[\hat{\beta}_{HOLS}\right] - \beta \;=\; E\left[(\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{X}'u \;-\; \hat{\alpha}^*\,(\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{X}'\hat{u}^{(3)}_{OLS}\right] = 0$$

$$\Rightarrow (\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{X}'E\left(u - \hat{\alpha}^*\hat{u}^{(3)}_{OLS}\,|\mathrm{X}\right) = 0 \;\Rightarrow\; E\left(u\,|\mathrm{X}\right) \;=\; E\left(\hat{\alpha}^*\hat{u}^{(3)}_{OLS}\,|\mathrm{X}\right) = 0.$$

Even if we strengthen our initial assumptions to full independence of the error term from the regressors, and even if we assume a symmetric error term, the expected value $E\left(\hat{\alpha}^*\hat{u}^{(3)}_{OLS}\,|\mathrm{X}\right)$ won't generally be zero, due to its non-linearity. The above result means that we must assess the performance of the HOLS estimator in finite samples in terms of Mean Squared Error (MSE) also.

Consistency. Set $\left[E\left(x_i x_i'\right)\right]^{-1} \equiv \mathrm{Q}^{-1}$. Given the assumptions of model (1), for consistency of the estimator we require in addition,

$$\hat{\beta}_{HOLS} - \beta \;=\; \left(\tfrac{1}{\mathrm{n}}\mathrm{X}'\mathrm{X}\right)^{-1}\left(\tfrac{1}{\mathrm{n}}\mathrm{X}'u\right) \;-\; \hat{\alpha}^*\left(\tfrac{1}{\mathrm{n}}\mathrm{X}'\mathrm{X}\right)^{-1}\left(\tfrac{1}{n}\mathrm{X}'\hat{u}^{(3)}_{OLS}\right) \;\xrightarrow{p}\; 0$$

$$\Rightarrow\; -\alpha^*\mathrm{Q}^{-1}\mathrm{plim}\left(\tfrac{1}{n}\mathrm{X}'\hat{u}^{(3)}_{OLS}\right) = 0 \Rightarrow\; E\left(x_i u_i^3\right) = 0.$$

The last expression follows from the assumptions of the model that make the OLS residuals consistent estimators of the error term, and by the continuity of the power function. Apart from the trivial case where $\alpha^* = 0 \Rightarrow \lambda = 0$ (which would make the estimator asymptotically identical to OLS), we have the following two alternative ways to obtain consistency:

A. Assume $E\left(u_i^3\,|\mathrm{X}\right) = 0$, which implies that $E\left(x_i u_i^3\right) = 0$ and guarantees consistency. But $E\left(u_i^3\,|\mathrm{X}\right) = 0 \implies E\left(u_i^3\right) = 0$. While there is an infinite number of distributions that have zero third moment but are not symmetric (one can easily construct one using a mixture model), admitedly such a condition would almost reflexively signal a symmetric error distribution, and this is not a negligible restriction: skewed residual/error distributions are not an infrequent phenomenon in applied practice, even though a lot of theoretical work is based on symmetric distributions. The necessity of this assumption, especially under conditional heteroskedasticity, comes from the fact that if the 2nd conditional moment of the error is a function of the regressors, so will in general be the conditional 3d moment, except if it is zero. We note that under conditional heteroskedasticity, the skewness *coefficient* (standardized 3d moment) can be independent of the regressors. An example is given by the Skew-normal distribution, where the skewness coefficient depends solely on the "slant" parameter that regulates skewness. So if we specify that regressors affect only the scale parameter of the distribution, we obtain conditional heteroskedasticity together with a skewness coefficient that does not depend on the regressors (see Azzalini & Capitanio 2014, pp.30-31). But again, this does not make the 3d moment itself independent of the regressors. This restriction to a zero third moment for the case of conditional heteroskedasticity is also present in Gourieroux, Monfort & Renault (1996), and Romano & Wolf (2017) make a critical note about it, indicating that it is almost unavoidably interpreted as an error symmetry assumption. In case we require a model that allows for a skewed error term, there is another way to obtain consistency of the estimator, as follows:

B. Assume $E\left(\mathrm{X}|\, u\right) = 0$. This condition is not equivalent to (and so it does not necessarily imply) the usual “mean-independence” assumption $E\left(u\,|\mathrm{X}\right) = 0$, although it is one of those cases where non-equivalence is mostly a theoretical curiosity. By the Law of Iterated Expectations, this assumption implies $E\left(\mathrm{X}\right) = 0$. But this does not require that the regressors are “inherently” zero-mean: we can achieve that simply by centering the sample (the regressors *and* the dependent variable -we center also the latter in order to have available the results for centered regression of the Frisch-Waugh-Lovell theorem). Given centered regressors, the condition postulates that there is no information in the error term about deviations of the regressors from their mean values. In practice, the assumption leads to $E\left(x_i h(u_i)\right) = 0$ for any measurable function of the error term, providing thus the desired consistency property (where in our case we have $h(u_i) = u_i^3$). Given this assumption, consistency is achieved under homoskedasticity, unconditional heteroskedasticity, and heteroskedasticity conditional on the regressors, since we can have $E\left(u_i^2\,|\mathrm{X}\right) = v\left(\mathrm{X}\right)$ together with $E\left(x_i u_i^2\right) = 0$ which is implied by $E\left(\mathrm{X}|\, u\right) = 0$. In the same manner, consistency will here be preserved also under non-zero conditional skewness, and even under $E\left(u_i^3\right) \neq E\left(u_j^3\right)$, $i \neq j$, i.e. when we have an unconditionally *heteroclitic* error. In all, if we set up a *centered* regression and we make the assumption $E\left(\mathrm{X}|\, u\right) = 0$, the estimator will be consistent under any skedastic assumption and any skewness assumption for the error term, providing high flexibility in accomodating real-world data samples. Certainly, by centering the sample we cannot estimate a constant term. But we can always use the OLS estimate for it on the uncentered sample, so the only price we pay for consistency of the HOLS estimator will be the non-reduced variance of the estimate of the constant term, hardly a heavy one.

To cover all cases, for the theoretical results we henceforth make the assumption $E\left(\mathrm{X}|\, u\right) = 0$ and treat the sample $\{y, \mathrm{X}\}$ as being centered already. Related to simulations, we will need to center the sample only when the error term is skewed.

Asymptotic distribution and variance. Define the limiting "White matrices"

$$\mathrm{W}_2 \equiv \operatorname{plim}\left(\frac{1}{n}\sum_{i=1}^{n} u_i^2 x_i x_i'\right), \ \mathrm{W}_4 \equiv \operatorname{plim}\left(\frac{1}{n}\sum_{i=1}^{n} u_i^4 x_i x_i'\right), \ \mathrm{W}_6 \equiv \operatorname{plim}\left(\frac{1}{n}\sum_{i=1}^{n} u_i^6 x_i x_i'\right).$$

Next, note that although $\hat{u}_{OLS}^{(3)} \xrightarrow{p} u^{(3)}$, we show in Supplement II (A.2.2 and A.3) that the vector $n^{-1/2}\mathrm{X}'\,\hat{u}_{OLS}^{(3)}$ does not tend to $n^{-1/2}\mathrm{X}'\,u^{(3)}$. The asymptotic vector here is (for non-optimized efficiency parameter)

$$\sqrt{n}\left(\hat{\beta}_{HOLS} - \beta\right) \;\longrightarrow\; \mathrm{Q}^{-1}\left[\left(I_K + 3\alpha \mathrm{W}_2 \mathrm{Q}^{-1}\right)\left(n^{-1/2}\mathrm{X}'u\right) - \alpha\left(n^{-1/2}\mathrm{X}'u^{(3)}\right)\right] \qquad (6)$$

wher I_K is the $K \times K$ identity matrix. In Supplement II A.3 we obtain the asymptpotic variance of this vector as

$$\begin{aligned}\mathrm{Avar}\left(\hat{\beta}_{HOLS}\right) &= \mathrm{Q}^{-1}\left[\mathrm{W}_2 - 2\alpha\left(\mathrm{W}_4 - 3\mathrm{W}_2\mathrm{Q}^{-1}\mathrm{W}_2\right)\right. \\ &\qquad \left. + \alpha^2\left(\mathrm{W}_6 \ - \ 6\mathrm{W}_4\mathrm{Q}^{-1}\mathrm{W}_2 \ + \ 9\mathrm{W}_2\mathrm{Q}^{-1}\mathrm{W}_2\mathrm{Q}^{-1}\mathrm{W}_2\right)\right]\mathrm{Q}^{-1}\,. \end{aligned} \tag{7}$$

where we wrote for economy Avar $\left(\hat{\beta}_{HOLS}\right) \equiv$ Avar $\left[\sqrt{n}\left(\hat{\beta}_{HOLS} - \beta\right)\right]$, a notational convention that we will follow throughout. As with the efficiency paramater α, the whole variance expression will take more specific forms depending on the skedastic assumption. Note that the corresponding variance of the OLS estimator is Avar $\left(\hat{\beta}_{OLS}\right) = \mathrm{Q}^{-1}\mathrm{W}_2\mathrm{Q}^{-1}$. Therefore, the HOLS estimator nests OLS if the efficiency parameter α is zero, which is a desirable property for the cases where HOLS does not outperform OLS (and there is a single such case as we will see), and the best it can do is to become OLS. Then the optimal α is zero and we expect that given sufficient sample size, the OLS estimate of α that will be used will indeed be very close to zero.

The Grenander conditions guarantee the existence of the probability limit of the inverse regressor moment matrix. The assumptions for consistency imply that the limiting random vector has zero mean, while the assumption that the possibly non-identical error moments have a finite average at the limit allows for a limiting distribution in any case. Finally, the assumption that the 6th error moment exists guarantees that the asymptotic variance exists and is finite. The vector then satisfies a classical multivariate Central Limit Theorem, converging to a mutlivariate Normal distribution.

This concludes the general derivation and presentation of the HOLS estimator. In summary, the estimator satisfies an approximated stationary point of an augmented "least-sqaures" objective function,and it is then optimized with respect to an efficiency parameter in order to have minimum possibe variance. It is consistent and can accomodate skewness in the error distribution.

We turn now to examine different skedastic asumptions, and the performance of the HOLS estimator compared to OLS.

4 The case of homoskedasticity

Here we assume that the error term is conditionally homoskedastic, and so also unconditionally, and in general identically distributed symmetrically around zero: the classical model. We can also extend the model by allowing for non-zero error skewness, adjusting the estimation procedure to maintain consistency as described previously. Studying the homoskedasticity case is important from a practical perspective: we want to know whether we can use the HOLS estimator regardless of whether heteroskedasticity exists in the sample, obtaining efficiency gains even if heterokedasticity is in reality absent.

The asymptotic vector to which the HOLS estimator converges here becomes

$$\sqrt{n}\left(\hat{\beta}_{HOLS} - \beta\right) \quad \longrightarrow \quad \mathrm{Q}^{-1}\left[\left(1+3\alpha\sigma_u^2\right)\left(n^{-1/2}\mathrm{X}'u\right) - \alpha\left(n^{-1/2}\mathrm{X}'u^{(3)}\right)\right] .$$

Its variance is (Supplement II-A.4.2)

$$\mathrm{Avar}\left(\hat{\beta}_{HOLS}\right) \ = \left[\sigma_u^2 \ - \ 2\sigma_u^4\gamma_2\alpha \ + \ \left(m_6 - 3\sigma_u^6\left(2\gamma_2+3\right)\right)\alpha^2\right]\mathrm{Q}^{-1}. \tag{8}$$

where $\gamma_2 = m_4/\sigma_u^4 - 3$ is the excess kurtosis coefficient of the error term. Minimizing the expression with respect to α we obtain

$$\alpha^* \ = \ \frac{\sigma_u^4\gamma_2}{m_6 - 3\sigma_u^6\left(2\gamma_2+3\right)} \, . \tag{9}$$

We see immediately that if excess kurtosis of the error term is zero then the optimal efficiency factor is also zero and the best HOLS can do in such a case is to become OLS. But the central member of the system of distributions we use is the Normal, which has zero excess kurtosis (after all, "excess" kurtosis has been defined with respect to the kurtosis of the Normal distribution). So under Homoskedastic Normality the HOLS estimator does not outperform OLS in terms of efficiency, athough it has the wisdom to realize that and to become OLS. This is a result arrived at also by Gourieroux, Monfort & Renault (1996), see their section 4.3. Still, the case of zero excess kurtosis is the only exception to the rule of efficiency gains. Inserting eq.(9) into eq.(8), the optimized asymptotic variance is

(Supplement II-A.4.2)

$$\text{Avar}^{*}\left(\hat{\beta}_{HOLS}\right) = \left(1 \; - \; \frac{\sigma_u^6\gamma_2^2}{m_6 - 3\sigma_u^6\left(2\gamma_2+3\right)}\right)\sigma_u^2\mathrm{Q}^{-1}. \tag{10}$$

By the Cauchy-Schwarz inequality we have,

$$\left[E\left(uu^3\right)\right]^2 < E\left(u^2\right)E\left[\left(u^3\right)^2\right] \Rightarrow \left[E\left(u^4\right)\right]^2 < E\left(u^2\right)E\left(u^6\right) \Rightarrow m_4^2 \; < \; \sigma_u^2 m_6.$$

Strict inequality applies because u and u^3 are not related by an affine function. With this result and a little algebra one can verify that

$$m_6 - 3\sigma_u^6\left(2\gamma_2+3\right) \; > \; \sigma_u^6\gamma_2^2 > 0\,.$$

From this, we have that the parenthesis in eq.(10) is strictly positive, and so we have a proper variance expression. But also, that the denominator in eq. (9) is always positive and so that if excess kurtosis is negative (platykurtic distributions) the optimal efficiency parameter will be negative too. Looking back at the original minimization problem and its Lagrangian (eqs (2) and (3)), this means that with platykurtic distributions the statistically optimal Lagrange multiplier will be negative (in comparison, the optimal efficiency parameter for the infeasible PHLS estimator is always positive as we show in Supplementary File I). Examples of well-known platykurtic distributions in econometrics are the Uniform, and the Bernoulli/Binomial for a large interval of the values for the probability parameter, roughly for $p \in (0.21,\, 0.79)$. Although these distributions are not usually *assumed* in regression models, what matters is whether they do represent occurrences of error terms in the data. As regards implementation of the HOLS estimator, we can consistently estimate $\hat{\alpha}^*$ using the OLS residuals through the more direct expression that propagates less uncertainty,

$$\hat{\alpha}^* \;\; = \frac{\hat{m}_4 - 3\hat{\sigma}_u^4}{\hat{m}_6 + \; 9\hat{\sigma}_u^6 - 6\hat{\sigma}_u^2\hat{m}_4}\,. \tag{11}$$

4.1 Efficiency gains.

4.1.1 Asymptotic gains.

We present in Table 1 the asymptotic efficiency gains of the HOLS estimator for four symmetric error distributions, a zero-mean Uniform, the Normal, the Logistic and the Laplace (calculations are in Supplement II-A.4.3). We have included the Uniform as an example of a platykurtic distribution, and to also reveal more clearly the link between efficiency gains and excess kurtosis.

Table 1: Asymptotic variance of the HOLS estimator as a fraction of OLS variance, for homoskedastic error distributions.

Error Distribution	**Variance** σ_u^2	**Excess** **Kurtosis**	**HOLS**
Uniform $\mathrm{U}\left(-c, c\right)$	$c^2/3$	-1.2	0.30
Normal $\mathrm{N}\left(0, \sigma^2\right)$	σ^2	0	1
Logistic $\Lambda\left(0, \omega\right)$	$\pi^2\omega^2/3$	1.2	0.94
Laplace $\mathrm{L}\left(0, \omega\right)$	$2\omega^2$	3	0.86

The efficiency gains of the HOLS estimator increase with the absolute value of excess kurtosis, and are much higher for platykurtic error distributions than for leptokurtic ones that have the same excess kurtosis in absolute terms. While leptokurtic distributions are admittedly what interests us more, note that these efficiency gains exist *for the case of homoskedasticity*: heteroskedasticity is absent here, but the HOLS estimator still performs better than OLS in terms of the asymptotic variance, outside of Homoskedastic Normality (or more accurately, homoskedastic zero-excess kurtosis). As we show in Supplementary File I the efficiency gains of the infeasibe PHLS estimator are much higher than HOLS (a dejavu from the case of GLS and Feasible GLS estimators), and they greatly outperform OLS even in the case of Normality: the PHLS achieves a reduction of 60% in asymptotic variance compared to OLS if the error term is Normal. Since the HOLS estimator is an approximation to PHLS, the impressive efficiency gains of the latter provide strong motivation to search for other methods to approximate PHLS in order to realize a larger part of the efficiency gains it promises.

We stress again the fact that we do not have a GMM-like over-identified system here, where the additional restrictions are known to lead to efficiency gains. The constraint that we incorporated in the initial objective function does not impose any restriction on the estimator because the actual value of H_u remains free to vary. Then, since asymptotically the HOLS estimator is unbiased, an obvious question arises: how can an unbiased estimator outperform asymptotically OLS (Gauss-Markov) in terms of efficiency in this classical setting?

The answer is that the HOLS estimator in reality estimates the same unknown vector but *based on a sample from a different population.* The typical member of the initial population is $\{x_i, u_i\}$ and we are given a (possibly centered) sample $\{\mathrm{X}, y;\ y = \mathrm{X}\beta + u\}$ from it. This is what the OLS estimator works with. But for the HOLS estimator, the population has members $\{x_i, \tilde{u}_i;\ \tilde{u}_i \equiv u_i - \alpha u_i^3\}$ which we approximate using OLS residuals to obtain the sample $\left\{\mathrm{X}, y_{HOLS};\ y_{HOLS} = \mathrm{X}\beta + u - \hat{\alpha}^* \hat{u}_{OLS}^{(3)}\right\}$. From another angle, what we are doing here is to perturb the initial "dependent variable" by keeping the same conditional expectation w.r.t regressors while changing the error term. In contrast, approaches like Cragg (1983, 1992) and Gourieroux, Monfort & Renault (1996) consider the use of instruments and additional orthogonality conditions to increase efficiency in the presence of conditional heteroskedastcity. Viewed from the perspective of our method, they perturb the conditional expectation of the model instead.

This way of viewing the situation also helps to clarify that the HOLS estimator is not some variant of shrinkage estimators, which where introduced by Stein (1958) and James & Stein (1961), and were the first to successfully challenge OLS in terms of estimation Risk and MSE. Initially the idea was to "shrink" the estimated coefficient vector towards zero in order to improve efficiency, but they have been extended towards other shrinkage directions, and they also host estimators that are weighted averages of other estimators. In this variant they appear closer to our work here. One such example can be found in Judge & Mittelhammer (2004), where they examine an optimal convex combination of OLS and another biased estimator, the shrinkage being towards the latter. In that work too, there exists an efficiency factor that is to be optimized in order to obtain MSE lower than that of OLS. But their composite estimator is designed in such a way so as to become OLS at the limit, making it a useful tool only for rather small samples. In contrast, HOLS maintains efficiency gains also at the limit. Another example is Hansen (2016), who constructed shrinkage estimators as a convex combination of the maximum likelihood estimator and one among various restricted estimators, where the shrinkage is towards a restricted parameter space. The common theme is the imposition of some kind of direct or indirect *restriction* on the unknown coeffiicent vector, which perhaps unsurprsiginly provides efficiency gains since it reduces the search area for the estimator. In contrast, our approach *expands* the view of the estimator, not as regards the unknown coefficient vector, but by forcing it to consider alternative data generating mechanisms. Since the property we examine, heteroskedasticity, hurts efficiency, informally we force the estimator to "hedge" against its possible existence, obtaining efficiency gains in the process.

4.1.2 Simulated Mean-Squared Error performance.

Since the HOLS estimator is biased in finite samples, of importance is its performance in terms of finite-sample MSE that takes into account also the bias. We present here our first simulated results (more details are to be found in Supplement II-B). The set up is as follows: we consider the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i, \quad i = 1, ..., n\,.$$

The two stochastic regressors are correlated and independent of the error term, and the observations in each sample are i.i.d. For each sample size and each error specification we run 2.000 simulations. We report the MSE of the HOLS estimator as a fraction of the MSE of the OLS estimator. The MSE is obtained as the sum of variances plus the sum of squared biases of all the coefficient estimates. Note that the variances and means used are the sample moments of the empirical distributions of the estimates. Namely, from each simulation we use only the point estimates. The finite-sample estimation of the HOLS variance-covariance matrix is examined in a later section.

Table 2 contains results for the four symmetric distributions of Table 1. In all cases, the error term has zero mean and unitary variance.

Table 2: Sample MSE of HOLS estimator as a fraction of OLS-MSE for symmetric homoskedastic error distributions. Uncentered regressions.

n	U	N	Λ	L
50	0.55	1.08	0.98	0.84
100	0.43	1.04	0.96	0.80
200	0.36	1.02	0.93	0.83
500	0.31	1.01	0.93	0.80
1000	0.31	1.00	0.94	0.81
2500	0.32	1.00	0.94	0.84
5000	0.31	1.00	0.95	0.83

Since by design the error term here is symmetric, we do not need to center the sample in the second estimation stage. The table verifies the theoretical results as regards the MSE/variance improvements anticipated from Table 1. We also see that the HOLS estimator recognizes rather quickly that the error term has zero excess kurtosis when this is the case, and the loss of efficiency is small and virtually disappears for sample sizes higher than 200.

In Table 3 we present simulation results when the error term is homoskedastic but skewed. We examine two distributional choices, the Skew-normal and the Asymmetric Laplace (as defined in Kotz, Kozubowski & Podgorski 2012, Proposition 3.1.2, p.137). Here too the distributions are centered and scaled to have zero mean and unitary variance. In the Skew-normal case, the skewness coefficient is $\gamma_1 \approx 0.40$, while the excess kurtosis coefficient is $\gamma_2 \approx 0.26$. In the Asymmetric Laplace case, we have $\gamma_1 \approx 1.97$, $\gamma_2 \approx 5.9$.

Table 3: Sample MSE / Variance of HOLS estimator as a fraction of OLS for skewed homoskedastic error distributions.

	SKEW NORMAL	**Asymmetric LAPLACE**		
n	Uncentered regr.	Uncentered regr.	Uncentered regr.	Centered regr.
	MSE	**MSE**	**Variance only**	**MSE**
50	1.05	0.79	0.76	0.78
100	1.02	0.82	0.76	0.78
200	1.01	0.87	0.77	0.76
500	1.00	0.96	0.79	0.78
1000	1.00	1.13	0.83	0.80
2500	1.01	1.44	0.85	0.79
5000	1.03	1.98	0.88	0.83

For the Skew-normal specification, we run only uncentered regressions, and so the existence of skewness creates inconsistency and bias, even asymptotically. But we see that the behavior of MSE closely resembles the Normal case. The reason is that the excess kurtosis is small and so the distribution resembles the Normal, and this dominates the skewness which creates the bias. In the case of the Asymmetric Laplace distribution where both skewness and kurtosis are high, we run both centered and uncentered regressions, and in the second case we report also comparisons as regards the variance, where we ignore the bias. We see that in the uncentered regression eventually the bias dominates MSE, even though the HOLS estimator has lower variance than OLS. We note that the bias eventually comes from the constant term alone. This is anticipated from the theory, as long as the error term has constant skew, because the bias relates to the term $\alpha\left(\frac{1}{n}\mathrm{X}'\mathrm{X}\right)^{-1}\left(\frac{1}{n}\mathrm{X}'u^{(3)}\right)$. With non-zero constant error skew and independence betwen the error term and the regressors, as is the simulation scenario, we have $\alpha\left(\frac{1}{n}\mathrm{X}'\mathrm{X}\right)^{-1}\left(\frac{1}{n}\mathrm{X}'u^{(3)}\right)\xrightarrow{p}\alpha\left[E\left(x_ix_i'\right)\right]^{-1}E\left(x_i\right)\cdot m_3=\alpha m_3\left(1,\,0_K\right)'$, the last equality holding as long as we have a constant in the regressor matrix. Finally, note that the variance in the uncentered regression is higher than in the centered regression (where it is equal to the MSE). This is due to the fact that, with non-zero-mean regressors, the existence of skewness affects also the variance since the matrix $n^{-1}\mathrm{X}'u^{(3)}u^{(3)\prime}\mathrm{X}$ that is present in the asymptotic variance-covariance matrix will not converge to the White matrix W_6.

5 Unconditional heteroskedasticity

Under unconditional heteroskedasticity (i.e. independent of the regressors), the OLS asymptotic variance is

$$\mathrm{Avar}\left(\hat{\beta}_{OLS}\right) = \bar{\sigma}^2\mathrm{Q}^{-1} \, ,$$

where as before $\bar{\sigma}^2$ is the limit of the average of the variances, assumed finite. For the HOLS estimator we obtained (see Supplement II-A.5)

$$\mathrm{Avar}\left(\hat{\beta}_{HOLS}\right) = \left[\bar{\sigma}^2 - 2\left(\bar{m}_4 - 3\bar{\sigma}^4\right)\alpha + \left(\bar{m}_6 + 9\bar{\sigma}^6 - 6\bar{\sigma}^2\bar{m}_4\right)\alpha^2\right]\mathrm{Q}^{-1} \, ,$$

with $\bar{m}_j$ denoting the limit of average moment j. Optimizing for α we get

$$\alpha^*_{HOLS} = \frac{\bar{m}_4 - 3\bar{\sigma}^4}{\bar{m}_6 + 9\bar{\sigma}^6 - 6\bar{\sigma}^2\bar{m}_4} \, ,$$

and

$$\mathrm{Avar}^*\left(\hat{\beta}_{HOLS}\right) = \left[1 - \frac{\bar{\sigma}^6\left(\bar{m}_4/\bar{\sigma}^4 - 3\right)^2}{\bar{m}_6 + 9\bar{\sigma}^6 - 6\bar{\sigma}^2\bar{m}_4}\right]\bar{\sigma}^2\mathrm{Q}^{-1} \, .$$

Essentially, due to the assumption of identically distributed regressors, we obtain again a version of the White matrices asymptotically, with the average moments in place of the common moments of the homoskedastic case.

As long as $\bar{m}_4/\bar{\sigma}^4 - 3 \neq 0$ the HOLS estimator will have lower asymptotic variance than OLS. In fact, this will hold also for the case of a Normal heteroskedastic error term. We have

$$\bar{m}_4 - 3\bar{\sigma}^4 = \lim \frac{1}{n}\sum_{i=1}^{n} m_{4,i} - 3\lim\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\right)^2$$

$$= \lim \frac{1}{n}\sum_{i=1}^{n} m_{4,i} - 3\lim\left(\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^4 + \frac{1}{n^2}\sum\sum_{i\neq j}\sigma_i^2\sigma_j^2\right)$$

$$= \lim \frac{1}{n}\sum_{i=1}^{n} m_{4,i} - 3\lim\frac{1}{n^2}\sum\sum_{i\neq j}\sigma_i^2\sigma_j^2.$$

Even if the error term is Normal and so we have $m_{4,i} = 3\sigma_i^4 \;\; \forall i$, the above expression won't be equal to zero in general. So we conclude that under unconditional heteroskedasticity HOLS will outperform OLS in terms of asymptotic variance, even in the case of Normality.

From a practical point of view, computing the optimal efficiency factor α proceeds in exactly the same way as in the homoskedastic case (see eq. (11)), by using sample averages of the OLS residuals: whether there exists a single common value for each moment, or we have different values and the average moments converge, under the imposed regularity conditions a Law of Large Numbers applies.

This implies that as with the OLS case, the homoskedastic and unconditionally heteroskedastic cases cannot be distinguished. The upside is that with the same estimator we can perform valid inference on the unknown coefficient vector irrespective of which of the two situations holds in the data.

5.1 Simulated MSE performance.

We apply a group-wise heteroskedasticity scheme with six different standard deviations with unequal proportional representation in the data, having as weighted average standard deviation $0.9996 \simeq 1$. We present the results in Table 4, for the same four distributions of Table 2. The first row of the table shows the resulting theoretical "sample average" excess kurtosis coefficient for each case. Heteroskedasticity increases the collective variability of the sample of heterogeneous random variables, resulting in higher average excess kurtosis than in the i.i.d. case (sign included). In this setup the efficiency gains of the HOLS estimator reduce for the platykurtic distributions and are increased for the leptokyrtic ones, veryfing once more that the absolute distance from zero excess kurtosis is what matters for this estimator. We also obtain efficiency gains when the error term is Normal but heteroskedastic.

Table 4: Sample MSE of HOLS estimator as a fraction of OLS-MSE for symmetric unconditionally heteroskedastic error distributions.

$(\bar{m}_4/\bar{\sigma}^4)-3=$	-0.66	0.90	2.46	4.80
n	U	N	Λ	L
50	0.96	1.00	0.92	0.80
100	0.91	0.96	0.88	0.74
200	0.89	0.95	0.86	0.76
500	0.86	0.96	0.85	0.78
1000	0.85	0.95	0.86	0.80
2500	0.84	0.95	0.88	0.81
5000	0.83	0.94	0.88	0.81

6 Conditional heteroskedasticity

The final case we will consider is heteroskedasticity of the error term conditional on the regressors. Here, the form of the variance matrix of the HOLS estimator is the one derived initially in Section 3 (eq. 7) since no simplification occurs in the White matrices. To obtain direct comparability with the OLS variance, we can make the following transformations,

$$\mathrm{Q}^{-1}\mathrm{W}_2\mathrm{Q}^{-1} = \mathrm{Avar}\left(\hat{\beta}_{OLS}\right) \equiv \mathrm{V}_2\ , \quad \mathrm{Q}^{-1}\mathrm{W}_4\mathrm{Q}^{-1} \equiv \mathrm{V}_4, \quad \mathrm{Q}^{-1}\mathrm{W}_6\mathrm{Q}^{-1}\ \equiv \mathrm{V}_6\, ,$$

and obtain (see Supplement II-A.6)

$$\mathrm{Avar}\left(\hat{\beta}_{HOLS}\right) = \mathrm{V}_2\, -\, 2\alpha\left(\mathrm{V}_4 - 3\mathrm{V}_2\mathrm{Q}\mathrm{V}_2\right)\, +\, \alpha^2\left(\mathrm{V}_6\, +\, 9\mathrm{V}_2\mathrm{Q}\mathrm{V}_2\mathrm{Q}\mathrm{V}_2\, -\, 6\mathrm{V}_2\mathrm{Q}\mathrm{V}_4\right). \quad (12)$$

Since the variance-covariance matrix does not simplify to a matrix proportional to the OLS variance as in the previous cases, we will obtain the optimal efficiency factor using the matrix trace operator. We get

$$\alpha^* = \frac{\mathrm{tr}\left\{\mathrm{V}_4 - 3\mathrm{V}_2\,\mathrm{Q}\mathrm{V}_2\right\}}{\mathrm{tr}\left\{\mathrm{V}_6\ +\ 9\mathrm{V}_2\mathrm{Q}\mathrm{V}_2\mathrm{Q}\mathrm{V}_2\ -\ 6\mathrm{V}_2\mathrm{Q}\mathrm{V}_4\right\}}\,, \tag{13}$$

and

$$\mathrm{tr}\left\{\mathrm{Avar}^*\left(\hat{\beta}_{HOLS}\right)\right\} = \mathrm{tr}\left\{\mathrm{V}_2\right\}\ -\ \frac{\mathrm{tr}^2\left\{\mathrm{V}_4 - 3\,\mathrm{V}_2\mathrm{Q}\mathrm{V}_2\right\}}{\mathrm{tr}\left\{\mathrm{V}_6\ +\ 9\mathrm{V}_2\mathrm{Q}\mathrm{V}_2\mathrm{Q}\mathrm{V}_2\ -\ 6\mathrm{V}_2\mathrm{Q}\mathrm{V}_4\right\}}\,.$$

This is the matrix analogue of the expressions obtained in the homoskedastic and unconditional heteroskedastic cases, and it will revert to them under the respective skedastic situations. This trace is smaller than $\mathrm{tr}\left\{\mathrm{V}_2\right\}$, the trace of the OLS asymptotic variance-covariance matrix, and again we have efficiency gains, since comparing the trace of the variance matrices is an established and intuitive way to assess relative efficiency in the multivariate case.

Conditional heteroskedasticity was the focus in Gourieroux, Monfort & Renault (1996). But in order to apply the overidentified GMM logic, they assumed $E\left(u_i^3\,|\mathrm{X}\right) = 0$, from which they obtained the orthogonality condition $E\left(x_i u_i^3\right) = 0$, that they turned into an "approximate" relation $E\left(x_i \hat{u}_{i,OLS}^2 u_i\right) = 0$ (equivalent at the limit), and then implemented a two-stage IV-GMM approach using as instrument the series $x_i \hat{u}_{i,OLS}^2$.

Table 5 contains the simulation results for the case of conditional heteroskedasticity. The first four columns relate to symmetric error distributions, while the last one to a zero-mean Asymmetric Laplace. Following partly MacKinnon (2013), the conditional heteroskedasticity scheme is

$$E\left(u_i^2|x_i\right) = \gamma \cdot \left(x_i'\beta\right)^2\,,\ \ \gamma = 0.1$$

For the symmetric error cases, we used the same systematic part for the regression as before (consant term, two correlated non-zero mean regressors). For the asymmetric case, to mimic the conditions under which the HOLS estimator remains consistent, we generated independent and symmetric regressors, while using the location parameter of the

Asymmetric Laplace to obtain mean-independence alongside conditional heteroskedasticity (see Supplement II-B for more details).

Table 5: Sample MSE of HOLS estimator as a fraction of OLS-MSE for conditionally heteroskedastic error distributions.

n	U	N	Λ	L	AsymL
50	1.18	1.22	1.18	1.01	0.77
100	1.27	1.09	0.95	0.83	0.69
200	1.06	0.90	0.77	0.68	0.71
500	0.95	0.78	0.69	0.66	0.71
1000	0.92	0.76	0.73	0.65	0.75
2500	0.92	0.77	0.75	0.71	0.77
5000	0.92	0.82	0.77	0.73	0.79

We see that for the smallest samples the HOLS estimator performs worse in some cases than OLS, although this is quickly reversed. As in the case with unconditional group-wise heteroskedasticity, "on average" heteroskedasticity increases excess kurtosis (sign included), so for example the draws from the Uniform distribution end up having a small positive excess kurtosis. Compared with the unconditional heteroskedasticity simulation results, HOLS performs even better here, for the leptokurtic distributions and sample sizes $n = 200$ and higher. The reason for the bad relative performance of HOLS in small samples is likely due to the fact that we use the estimated asymptotic optimal efficiency factor α^*. This was true also in the previous setups, but here α^* is computed using traces of matrix expressions, not sample means of residual powers. For the asymmetric error case, the HOLS estimator outperforms OLS throughout.

This concludes the examination of the HOLS estimator in comparison to the OLS estimator regarding efficiency under different skedastic and skewness assumptions. The theoretical results and the simulation evidence suggest that the HOLS estimator should be preferred to the OLS in most of the cases encountered in applied studies.

7 The VCV matrix of the HOLS estimator, and its estimation routine.

To proceed with inference, we need to estimate the variance-covariance matrix of the HOLS estimator. We show in Supplement II-A.7 that under homoskedasticity,

$$\left(\text{plim}\frac{1}{n}\sum_{i=1}^{n}\hat{u}_{i,HOLS}^{2}\right)\left(n^{-1}\mathrm{X}'\mathrm{X}\right)^{-1} = \left(\sigma_u^2 \ - \ 2\alpha^* m_4 \ + \ (\alpha^*)^2 m_6\right)\mathrm{Q}^{-1}.$$

The expression has the form of the variance of the infeasible PHLS estimator (see Supplement I). This happens because the additional terms that increase the variance of HOLS compared to the PHLS estimator are $O_p\left(\sqrt{n}\right)$ but also $o_p\left(n\right)$, so they disappear asymptotically when we examine sample means. We show in Supplement II-A.7 that use of this expression overestimates the HOLS variance when $\gamma_2 < 0$ (platykurtic error term) while when $\gamma_2 > 0$ it is inconsistent in general (while simulations not reported here show that it tends to underestimate the HOLS variance). Therefore we turn once more to the OLS residuals, and use them to estimate the HOLS variance. Which expressions we will use depends on the skedastic assumption made, athough the results suggest that for sample sizes larger than $n = 200$ we can use the more general expressions, eqs (12) and (13), that cover conditional heterosedasticity but also the cases of unconditional heteroskedasticity and homoskedasticity. Nevertheless this is not obligatory and if the researcher is confident that the error term is homoskedastic for example, they can use the simpler expressions, eqs (8) and (11).

We conclude the presentation of the HOLS estimator by detailing its estimation routine:

1) Run OLS estimation on the uncentered sample. By the Frisch-Waugh-Lovell theorem the OLS residuals are identical whether we operate on a centered or on an

uncentered sample. Obtain the residual series and its powers. If the evidence points to a symmetric error distribution, the next estimation stage can be performed using the uncentered sample. Otherwise, the sample should be centered for the second phase, while we keep the OLS estimate for the constant term and its variance, as well as the estimate for the error variance, if it is of autonomous interest.

2) Estimate the optimal $\hat{\alpha}^*$ depending on the skedastic assumption (see previous sections).

3) Transform the dependent variable by $y_{HOLS} = y \; - \; \hat{\alpha}^* \hat{u}^{(3)}_{OLS}$, where y is uncentered or centered depending on the case.

4) Run OLS on the $\{\mathrm{X}, y_{HOLS}\}$ (centered or uncentered) sample,

$$\hat{\beta}_{HOLS} \;\; = \;\; \left(\mathrm{X}'\mathrm{X}\right)^{-1} \mathrm{X}' \, y_{HOLS} \, . \tag{14}$$

5) Use the regressors and the OLS residuals in order to compute consistently the variance of the HOLS etimator using the relevant formulas depending on the skedastic assumption, and proceed with statistical inference.

8 Conclusions and directions for further research

In this paper we have presented a new method to construct least squres estimators, augmenting the objective function that such estimators minimize by a component that represents a property of the unknown error term. We focused on heteroskedasticity. The resulting estimator, which we have called HOLS, takes into account the possibility of a heteroskedastic error term already in the coefficient estimation stage, and not only for the calculation of its variance-covariance matrix. The estimator is easily implemented in two stages and is based on OLS residuals. The HOLS estimator is consistent and, in terms of finite-sample MSE and asymptotic efficiency, it outperforms OLS under conditional and unconditional heteroskedasticity, but also under homoskedasticity, both asymptotically but also in finite samples, in almost all cases. Therefore it is a serious contender for all-purpose use in applied studies, as an alternative to the "heteroskedasticity-robust" OLS approach that is the predominant applied method.

This being a new endeavor, obviously a lot of work remains to be done: perhaps first in order would be an extended simulation study around the estimation of the VCV matrix of the HOLS estimator and its behavior in tests of statistical significance. The adjustments examined in the literature as regards the improved estimation in finite samples of White's heteroskedasticity-robust VCV matrix could and should be examined for the case of the HOLS estimator also. Comparing the OLS and HOLS performance related to statistical tests in general, is also important. Extensions to deal with regressor endogeneity and an "Instrumental Variables" variant of the HOLS estimator would also be useful and widely applicable.

Moreover, since the HOLS estimator is an imperfect incarnation of an ideal infeasible estimator, other approximations to the latter could prove to achieve even larger efficiency gains. Finally, the method is not confined to dealing with possible heteroskedasticity, but can be used to accomodate other statistical aspects of the error term, skewness and autocorrelation being the first two that come to mind.

References

Azzalini, A. and A. Capitanio (2014): The skew-normal and related families. Cambridge U.K.: Cambridge University Press.

Breusch, T. S., and A.R. Pagan (1979). A simple test for heteroscedasticity and random coefficient variation. Econometrica, 47(5), 1287-1294.

Cragg, J. G. (1983): More efficient estimation in the presence of heteroscedasticity of unknown form, Econometrica, 51(3), 751-763.

Cragg, J. G. (1992): Quasi-Aitken estimation for heteroskedasticity of unknown form, Journal of Econometrics, 54, 179-201.

Gourieroux, C., A. Monfort, and E. Renault (1996): Two-stage generalized moment method with applications to regressions with heteroscedasticity of unknown form, Journal of Statistical Planning and Inference, 50, 37-63.

James, W., and C. Stein (1961): Estimation with quadratic loss. In Proceedings of the fourth Berkeley symposium on mathematical statistics and probability vol. 1, 361-379.

Judge, G.G., and R.C. Mittelhammer (2004): A Semiparametric Basis for Combining Estimation Problems Under Quadratic Loss, Journal of the American Statistical Association, 99:466, 479-487.

Hansen, B. E. (2016): Efficient shrinkage in parametric models. Journal of Econometrics, 190, 115-132.

Hausman, J. A. (1978): Specification tests in econometrics, Econometrica, 46,1251-1271.

Kotz, S., T. Kozubowski, and K. Podgorski (2012): The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance. New York: Springer Science & Business Media.

MacKinnon, J. G. (2013): Thirty years of heteroskedasticity-robust inference. In Chen, X. and N.R. Swanson N.R., eds., Recent advances and future directions in causality, prediction, and specification analysis (437-461). New York: Springer.

Romano, J. P., & M. Wolf (2017): Resurrecting weighted least squares, Journal of Econometrics, 197, 1-19.

Spanos, A. (2010): Akaike-type criteria and the reliability of inference: Model selection versus statistical model specification, Journal of Econometrics, 158, 204-220.

Stein, C. (1956): Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Proceedings of the third Berkeley symposium on mathematical statistics and probability.

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

–