Romano, Joseph P.; Wolf, Michael

Working Paper

Resurrecting weighted least squares

Working Paper, No. 172

Provided in Cooperation with:
Department of Economics, University of Zurich

Suggested Citation: Romano, Joseph P.; Wolf, Michael (2014) : Resurrecting weighted least squares, Working Paper, No. 172, University of Zurich, Department of Economics, Zurich, https://doi.org/10.5167/uzh-98546

This Version is available at:
https://hdl.handle.net/10419/111232

University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 172

Resurrecting Weighted Least Squares

Joseph P. Romano and Michael Wolf

September 2014

Resurrecting Weighted Least Squares

Joseph P. Romano *
Departments of Statistics and Economics
Stanford University
romano@stanford.edu

Michael Wolf
Department of Economics
University of Zurich
michael.wolf@econ.uzh.ch

September 2014

Abstract

Linear regression models form the cornerstone of applied research in economics and other scientific disciplines. When conditional heteroskedasticity is present, or at least suspected, the practice of reweighting the data has long been abandoned in favor of estimating model parameters by ordinary least squares (OLS), in conjunction with using heteroskedasticity consistent (HC) standard errors. However, we argue for reintroducing the practice of reweighting the data, since doing so can lead to large efficiency gains of the resulting weighted least squares (WLS) estimator over OLS even when the model for reweighting the data is misspecified. Efficiency gains manifest in a first-order asymptotic sense and thus should be considered in current empirical practice. Crucially, we also derive how asymptotically valid inference based on the WLS estimator can be obtained even when the model for reweighting the data is misspecified. The idea is that, just like the OLS estimator, the WLS estimator can also be accompanied by HC standard errors without knowledge of the functional form of conditional heteroskedasticity. A Monte Carly study demonstrates attractive finite-sample properties of our proposals compared to the *status quo*, both in terms of estimation and making inference.

KEY WORDS: Conditional heteroskedasticity, HC standard errors, weighted least squares.

JEL classification codes: C12, C13, C21.

*Research supported by NSF Grant DMS-0707085.

1 Introduction

Despite constant additions to the toolbox of applied researchers, linear regression models remain the cornerstone of empirical work in economics and other scientific disciplines. Any introductory course in econometrics starts with an assumption of conditional homoskedasticity: the conditional variance of the error terms does not depend on the regressors. In such an idyllic situation, one should estimate the model parameters by *ordinary least squares* (OLS) and use the conventional inference produced by any of the multitude of software packages.

Unfortunately, in many applications, applied researchers are plagued by conditional heteroskedasticity: the conditional variance of the error term is a function of the regressors. A simple example is a wage regression where wages (or perhaps log wages) are regressed on experience plus a constant. In most professions, there is a larger variation in wages for workers with many years of experience compared to workers with few years of experience. Therefore, in such a case, the conditional variance of the error term is an increasing function of experience.

In the presence of conditional heteroskedasticity, the OLS estimator still has attractive properties, such as being unbiased and being consistent (under mild regularity conditions). However, it is no longer the best linear unbiased estimator (BLUE). Even more problematic, conventional inference generally is no longer valid: confidence intervals do not have the correct coverage probabilities and hypothesis tests do not have the correct null rejection probabilities, even asymptotically. In early days, econometricians prescribed the cure of *weighted least squares* (WLS). It consisted of modeling the functional form of conditional heteroskedasticity, reweighting the data (both the response variable and the regressors), and running OLS combined with conventional inference with the weighted data. The rationale was that 'correctly' weighting the data (based on the true conditional variance model) results in efficiency gains over the OLS estimator. Furthermore, conventional inference based on the 'correctly' weighted data is valid, at least asymptotically.

Then came White (1980) who changed the game with one of the most influential and widely-cited papers in the field. He promoted *heteroskedasticity consistent* (HC) standard errors for the OLS estimator. His alternative cure consists of retaining the OLS estimator (that is, not weighting the data) but using HC standard errors instead of the conventional standard errors. The resulting inference is (asymptotically) valid in the presence of conditional heteroskedasticity of unknown form, which has been a major selling point. Indeed, the earlier cure had the nasty side effect of invalid inference if the applied researcher did not model the conditional heteroskedasticity correctly (arguably, a common occurrence).

As the years have passed, weighting the data has become out of fashion and applied researchers have instead largely favored the cure prescribed by White (1980) and his followers. The bad publicity for WLS is still ongoing. As an example, consider Angrist and Pischke (2010, Section 3.4.1) who discourage applied researchers from weighting the data with the following arguments, among others.

1. "If the conditional variance model is a poor approximation or if the estimates of it are very noisy, WLS estimators may have worse finite-sample properties than unweighted estimators."
2. "The inferences you draw [...] may therefore be misleading, and the hoped-for efficiency gain may not materialize."
3. "Any efficiency gain from weighting is likely to be modest, and incorrectly or poorly estimated weights can do more harm than good."

Alas, not everyone has converted and a few lone warriors defending WLS remain. At the forefront is Leamer (2010, p.43) who calls the current practice "White-washing" and argues that "...we should be doing the hard work of modeling the heteroskedasticity [...] to determine if sensible reweighting of the observations materially changes the locations of the estimates of interest as well as the widths of the confidence intervals."

In this paper, we offer a new, third cure, which is a simple combination of the two previous cures: use WLS combined with HC standard errors. The aim of this cure is to offer the best of both worlds. First, sensibly weighting the data can lead to noticeable efficiency gains over OLS, even if the conditional variance model is misspecified. Second, combining WLS with HC standard errors allows for valid inference, even if the conditional variance model is misspecified. The cure we offer is a simple and natural one. But, to the best of our knowledge, it has not been offered before. For example, Hayashi (2000, Section 2.8) describes the approach of estimating a parametric specification of conditional heteroskedasticity, but the corresponding inference for the parameter vector assumes the parametric specification is correctly specified; otherwise, it may not be valid. Our approach is similar in that it also specifies a parametric specification for the skedastic function, but it is different in that it produces valid inference under general forms of conditional heteroskedasticity even when the parametric specification does not include the true skedastic function.

As a bonus, we also propose a new estimator: *adaptive least squares* (ALS). Our motivation is a follows. Under conditional homoskedasticity, OLS is the optimal estimator and one should not weight the data at all. Using WLS in such a setting will lead to an efficiency loss, at least in small and moderate samples, because of the noise in the estimated conditional variance model. As a remedy, we propose to first carry out a test of conditional heteroskedasticity, based on the very conditional variance model intended to be used in weighting the data. If the test rejects, use WLS; otherwise, stick with OLS. In this way, one will only use WLS when it is worthwhile doing so, that is, when there is sufficient evidence in the data supporting the conditional variance model. Crucially, independent of the outcome of the test, always use HC standard errors.[1]

The remainder of the paper is organized as follows. Section 2 introduces the model. Section 3

[1]Tests for conditional heteroskedasticity had come with a different prescription in the past. Namely, if the test rejects, use OLS with HC standard errors, otherwise, use OLS with the conventional standard errors; for example, see Hayashi (2000, p.132). But such a practice is not recommended, since it has poor finite-sample properties under conditional heteroskedasticity in small and moderate samples; for example, see Long and Ervin (2000, Section 4.3). The reason is that when the test has low power, an invalid inference method will be chosen with non-negligible

describes the various estimators and derives the asymptotic distribution of the WLS estimator when the weighting of the data is possibly incorrect. Section 4 establishes validity of our proposed inference based on the WLS estimator when the weighting of the data is possibly incorrect. Section 5 examines finite-sample performance via a Monte Carlo study. Section 6 briefly discusses possible variations and extensions. Finally, Section 7 concludes. An Appendix contains details on various inference methods and all mathematical proofs.

2 The Model

We maintain the following set of assumptions throughout the paper.

(A1) The linear model is of the form

$$y_i = x_i'\beta + \varepsilon_i \quad (i = 1, \ldots, n) \ , \tag{2.1}$$

where $x_i \in \mathbb{R}^K$ is a vector of explanatory variables (regressors), $\beta \in \mathbb{R}^K$ is a coefficient vector, and ε_i is the unobservable error term with certain properties to be specified below.

(A2) The sample $\big\{(y_i, x_i')\big\}_{i=1}^n$ is independent and identically distributed (i.i.d.).

(A3) All the regressors are predetermined in the sense that they are orthogonal to the contemporaneous error term:

$$\mathbb{E}(\varepsilon_i | x_i) = 0 \ . \tag{2.2}$$

Of course, under the i.i.d. assumption (A2) it then also holds that

$$\mathbb{E}(\varepsilon_i | x_1, \ldots, x_n) = 0 \ ,$$

that is, the regressors are strictly exogenous.

(A4) The $K \times K$ matrix $\Sigma_{xx} := \mathbb{E}(x_i x_i')$ is nonsingular (and hence finite). Furthermore, $\sum_{i=1}^n x_i x_i'$ is invertible with probability one.

(A5) The $K \times K$ matrix $\Omega := \mathbb{E}(\varepsilon_i^2 x_i x_i')$ is nonsingular (and hence) finite.

(A6) There exists a nonrandom function $v : \mathbb{R}^K \to \mathbb{R}_+$ such that

$$\mathbb{E}(\varepsilon_i^2 | x_i) = v(x_i) \ . \tag{2.3}$$

Therefore, the *skedastic function* $v(\cdot)$ determines the functional form of the conditional heteroskedasticity. Note that under (A6),

$$\Omega = \mathbb{E}\big[v(x_i) \cdot x_i x_i'\big] \ .$$

probability. Instead, we use tests for conditional heteroskedasticity for an honorable purpose and thereby restore some of their lost appeal.

It is useful to introduce the customary vector-matrix notations

$$y := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} , \quad \varepsilon := \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} , \quad X := \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1K} \\ \vdots & \dots & \vdots \\ x_{n1} & \dots & x_{nK} \end{bmatrix} ,$$

so that equation (2.1) can be written more compactly as

$$y = X\beta + \varepsilon \ . \tag{2.4}$$

Furthermore, assumptions (A2), (A3), and (A5) imply that

$$\mathrm{Var}(\varepsilon|X) = \begin{bmatrix} v(x_1) & & \\ & \ddots & \\ & & v(x_n) \end{bmatrix} .$$

Remark 2.1 (Justifying the I.I.D. Assumption)**.** The application of WLS relies upon $\mathrm{Var}(\varepsilon|X)$ being a diagonal matrix. For the sake of theory, it is possible to generalize the set of assumptions (A2)–(A5) such that this condition is still satisfied. For the sake of simplicity, however, we prefer to maintain the set of assumptions (A2)–(A5), which are based on the key assumption (A2) of observing a random sample. Our reasoning here is that virtually all applications of WLS are restricted to such a setting, a leading example being cross-sectional studies. Therefore, allowing for more general settings would mainly serve to impress theoreticians as opposed to keeping it simple for our target audience, namely applied researchers. ■

3 Estimators: OLS, WLS, and ALS

3.1 Description of the Estimators

The ubiquitous estimator of β is the *ordinary least squares* (OLS) estimator

$$\hat{\beta}_{\mathrm{OLS}} := (X'X)^{-1}X'y \ .$$

Under the maintained assumptions, the OLS is unbiased and consistent. This is the good news. The bad news is that it is not efficient under conditional heteroskedasticity (that is, when the skedastic function $v(\cdot)$ is not constant).

A more efficient estimator can be obtained by reweighting the data (y_i, x_i') and then applying OLS in the transformed model

$$\frac{y_i}{\sqrt{v(x_i)}} = \frac{x_i'}{\sqrt{v(x_i)}}\beta + \frac{\varepsilon_i}{\sqrt{v(x_i)}} \ . \tag{3.1}$$

Letting

$$V := \begin{bmatrix} v(x_1) & & \\ & \ddots & \\ & & v(x_n) \end{bmatrix},$$

the resulting estimator can be written as

$$\hat{\beta}_{\text{BLUE}} := (X'V^{-1}X)^{-1}X'V^{-1}y \,. \tag{3.2}$$

It is the best linear unbiased estimator (BLUE) and is consistent; in particular, it is more efficient than the OLS estimator. But outside of textbooks, this 'oracle' estimator mainly exists in utopia, since the skedastic function $v(\cdot)$ is typically unknown.

A feasible approach is to estimate the skedastic function $v(\cdot)$ from the data in some way and to then apply OLS in the model

$$\frac{y_i}{\sqrt{\hat{v}(x_i)}} = \frac{x_i'}{\sqrt{\hat{v}(x_i)}}\beta + \frac{\varepsilon_i}{\sqrt{\hat{v}(x_i)}} \,, \tag{3.3}$$

where $\hat{v}(\cdot)$ denotes the estimator of $v(\cdot)$. The resulting estimator is the *weighted least squares* (WLS) estimator.[2] Letting

$$\hat{V} := \begin{bmatrix} \hat{v}(x_1) & & \\ & \ddots & \\ & & \hat{v}(x_n) \end{bmatrix},$$

the WLS estimator can be written as

$$\hat{\beta}_{\text{WLS}} := (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y \,.$$

It is not necessarily unbiased. If $\hat{v}(\cdot)$ is a consistent estimator of $v(\cdot)$, than WLS is asymptotically more efficient than OLS. But even if $\hat{v}(\cdot)$ is an inconsistent estimator of $v(\cdot)$, WLS can result in large efficiency gains over OLS in the presence of noticeable conditional heteroskedasticity; see Section 5.

Using OLS is straightforward and has become the *status quo* in applied economic research. But foregoing potentially large efficiency gains 'on principle' would seem an approach to data analysis that is hard to justify.

Remark 3.1 (Adaptive Least Squares)**.** Under conditional homoskedasticity — that is, when the skedastic function $v(\cdot)$ is constant — OLS is generally more efficient than WLS in finite samples. But, under certain assumptions on the scheme to estimate the skedastic function, OLS and WLS are asymptotically equivalent in this case. On the other hand, under (noticeable) conditional heteroskedasticity, WLS is generally more efficient, both in finite samples and even in a first-order asymptotic sense. (Such claims will be justified mathematically later.)

[2] Another convention is to call *weighted least squares estimator* what we call best linear unbiased estimator and to call *feasible weighted least squares estimator* what we call weighted least squares estimator.

Therefore, it is tempting to decide based on the data which route to take: OLS or WLS. Specifically, we suggest applying a test for conditional heteroskedasticity. Several such tests exists, the most popular ones being the tests of Breusch and Pagan (1979) and White (1980); also see Koenker (1981) and Koenker and Bassett (1982). If the null hypothesis of conditional homoskedasticity it not rejected by such a test, use the OLS estimator; otherwise, use the WLS estimator. We call the resulting estimator the *adaptive least squares* (ALS) estimator. The motivation is as follows. Under conditional homoskedasticity, the ALS estimator will be equal to the WLS estimator with a small probability only (roughly equal to the nominal size of the test). Therefore, in this case, ALS is expected to be more efficient than WLS in finite samples, though still less efficient than OLS.

Under conditional heteroskedasticity, the ALS estimator will be equal to the WLS estimator with probability tending to one (assuming that the chosen test is consistent against the existing nature of conditional heteroskedasticity). So for large sample sizes, ALS should be almost as efficient as WLS. For small sample sizes, when the power of the test is not near one, the efficiency is expected to be somewhere between OLS and WLS. (In fact, one could apply the same strategy, but letting the significance level α_n of the "pretest" tend to zero as the sample size tends to infinity; one just needs to ensure α_n tends to zero slowly enough so that the test still has power tending to one.)

Consequently, ALS sacrifices some efficiency gains of WLS under conditional heteroskedasticity in favor of being closer to the performance of OLS under conditional homoskedasticity.

These heuristics are confirmed by Monte Carlo simulations in Section 5. ■

Remark 3.2 (Best Linear Predictor)**.** Consider a new observation (y, x'). It is well known that even in the absence of a linear model (that is, with Assumption (A1) not holding), the best linear predictor of y in the mean squared error sense is given by

$$x'\beta^* \ , \quad \text{with} \quad \beta^* := \left[\mathbb{E}(xx')\right]^{-1}\mathbb{E}(x \cdot y) \ ;$$

for example, see Hayashi (2000, Proposition 2.8). Under assumptions (A2) and (A4), the OLS estimator consistently estimates β^*, that is,

$$\hat{\beta}_{\text{OLS}} \xrightarrow{P} \beta^* \ , \tag{3.4}$$

where $\xrightarrow{P}$ denotes convergence in probability. Moreover, the condition $E(x_i \cdot \epsilon_i) = 0$ is ensured when $\beta = \beta^*$, rather than the stronger assumption (A3). Consistency is not necessarily shared by the WLS and ALS estimators. (Note, however, that the weighted least squares estimator can similarly be viewed as a best linear predictor based on a mean weighted squared error criterion.)

The consistency result (3.4) is sometimes viewed as an attractive robustness property of the OLS estimator. But we feel that is not of great practical importance. The best predictor in the mean squared sense is the conditional expectation $\mathbb{E}(y|x)$. If the linearity assumption (A1) does not hold, the conditional expectation $\mathbb{E}(y|x)$ can be arbitrarily far from the best linear predictor $x'\beta^*$. Therefore, if it is suspected that the linearity assumption (A1) may not hold, rather than settling for the

best linear predictor, it may be more fruitful to include more covariates or to use a nonparametric approach to estimate the conditional expectation.

Needless to say, if the goal is to interpret the estimator of β (in the sense of quantifying the 'effect' of the various entries of the vector of covariates x on the response variable y) or to make inference for β, then all three estimators — OLS, WLS, and ALS — rely on the validity of the linearity assumption (A1). So for such purposes, it is just as (un)safe to use WLS or ALS instead of OLS. ■

3.2 Parametric Model for Estimating the Skedastic Function

In order to estimate the skedastic function $v(\cdot)$, we suggest the use of a parametric model $v_\theta(\cdot)$, where $\theta \in \mathbb{R}^d$ is a finite-dimensional parameter. Such a model could be suggested by economic theory, by exploratory data analysis (that is, residual plots from an OLS regression), or by convenience. In any case, the model used should nest the case of conditional homoskedasticity. In particular, for every $\sigma^2 > 0$, we assume the existence of a unique $\theta := \theta(\sigma^2)$ such that

$$v_\theta(x) \equiv \sigma^2 \ .$$

A flexible parametric model we suggest is

$$v_\theta(x_i) := \exp\bigl(\nu + \gamma_2 \log|x_{i,2}| + \ldots + \gamma_K \log|x_{i,K}|\bigr) \ , \quad \text{with} \quad \theta := (\nu, \gamma_2, \ldots, \gamma_K)' \ , \tag{3.5}$$

assuming that $x_{i,1} \equiv 1$ (that is, the original regression contains a constant). Otherwise, the model should be

$$v_\theta(x_i) := \exp\bigl(\nu + \gamma_1 \log|x_{i,1}| + \gamma_2 \log|x_{i,2}| + \ldots + \gamma_K \log|x_{i,K}|\bigr) \ , \quad \text{with} \quad \theta := (\nu, \gamma_1, \ldots, \gamma_K)' \ .$$

Such a model is a special case of the form of multiplicative conditional heteroskedasticity previously proposed by Harvey (1976) and Judge et al. (1988, Section 9.3), among others.

Another possibility is to not take exponents and use

$$v_\theta(x_i) := \nu + \gamma_2|x_{i,2}| + \ldots + \gamma_K|x_{i,K}| \ , \quad \text{with} \quad \theta := (\nu, \gamma_2, \ldots, \gamma_K)' \ , \tag{3.6}$$

The advantage of (3.5) over (3.6) is that it ensures variances are nonnegative, though the parameters in (3.6) can be restricted such that nonnegativity is satisfied. In all cases, the models obviously nest the case of conditional homoskedasticity.

Furthermore, we recommend to base the test for conditional heteroskedasticity used in computing the ALS estimator of Remark 3.1 on the very parametric model of the skedastic function used in computing the WLS estimator. The motivation is that in this fashion, the ASL estimator is set to the WLS estimator (as opposed to the OLS estimator) only if there is significant evidence for the type of conditional heteroskedasticity that forms the basis of the WLS estimator. In particular, we

do not recommend to use a 'generic' test of conditional heteroskedasticity, such as the test of White (1980), unless the parametric specification $v_\theta(\cdot)$ used by the test is also the parametric specification used by the WLS estimator.[3]

Having chosen a parametric specification $v_\theta(\cdot)$ the test for conditional heteroskedasticity is then carried out by regressing the squared OLS residuals on the parametric specification, possibly after a suitable transformation to ensure linearity on the right-hand side, and by then comparing n times the R^2-statistic of this regression against the quantile of a chi-squared distribution.

For example, if the parametric model is given by (3.5), the test specifies

$$H_0 : \gamma_2 = \ldots = \gamma_K = 0 \quad \text{vs.} \quad H_1 : \text{ at least one } \gamma_k \neq 0 \ \ (k = 2, \ldots, K) \ .$$

To carry out the test, fix a small constant $\delta > 0$, estimate the following regression by OLS:

$$\log\big[\max(\delta^2, \hat{\varepsilon}_i^2)\big] = \nu + \gamma_2 \log|x_{i,2}| + \ldots + \gamma_K \log|x_{i,K}| + u_i \ , \quad \text{with} \quad \hat{\varepsilon}_i := y_i - x_i'\hat{\beta}_{\text{OLS}} \ , \tag{3.7}$$

and denote the resulting R^2-statistic by R^2. Furthermore, denote by $\chi^2_{K-1,1-\alpha}$ the $1-\alpha$ quantile of the chi-squared distribution with $K-1$ degrees of freedom. Then the test for conditional heteroskedasticity rejects at nominal level α if $n \cdot R^2 > \chi^2_{K-1,1-\alpha}$. (The reason for introducing the constant δ here is that, because we are taking logs, we need to avoid a residual of zero, or even very near zero. If instead, we considered the specification (3.6), we would simply run a regression of ε_i^2 on the right-hand side of (3.7) and no constant δ needs to be introduced.)

Finally, the estimate of the skedastic function is given by

$$\hat{v}(\cdot) := v_{\hat{\theta}}(\cdot) \ ,$$

where $\hat{\theta}$ is an estimator of θ obtained by on OLS regression of the type (3.7).

3.3 Limiting Distribution of the WLS Estimator

The first goal is to consider the behavior of the weighted least squares estimator under a perhaps incorrectly specified skedastic function. The estimator $\hat{\beta}_{\text{BLUE}}$ assumes knowledge of the true skedastic function $v(\cdot)$. Instead, consider a generic WLS estimator that is based on the skedastic function $w(\cdot)$; this estimator is given by

$$\hat{\beta}_W := (X'W^{-1}X)^{-1}X'W^{-1}y \ , \tag{3.8}$$

where W is the diagonal matrix with (i,i) entry $w(x_i)$. Given two real-valued functions $a(\cdot)$ and $b(\cdot)$ defined on $\mathbb{R}^K$ (the space where x_i lives), define $\Omega_{a/b}$ to be the matrix given by

$$\Omega_{a/b} := \mathbb{E}\left[\frac{a(x_i)}{b(x_i)} \cdot x_i x_i'\right] \ .$$

[3]For example, we would not recommend the parametric specification of White's (1980) test, as it is of order K^2 and thus involves too many free parameters (unless the number of regressors, K, is very small compared to the sample size, n).

The first result deals with the case of a fixed employed choice of skedastic function $w(\cdot)$, though this choice may be misspecified, since the true skedastic function is $v(\cdot)$.

Lemma 3.1. *Assume (A1)–(A3) and (A6). Given a possibly mispecified skedastic function $w(\cdot)$ and the true skedastic function $v(\cdot)$, assume the matrices $\Omega_{1/w}$ and Ω_{v/w^2} are well-defined (in the sense of the corresponding expectations to exist and being finite). Also, assume $\Omega_{1/w}$ is invertible. (These assumptions reduce to the usual assumptions (A4) and (A5) in case $w(\cdot)$ is constant.) Then, as $n \to \infty$,*

$$\sqrt{n}(\hat{\beta}_W - \beta) \xrightarrow{d} N\big(0, \Omega_{1/w}^{-1}\Omega_{v/w^2}\Omega_{1/w}^{-1}\big) \ .$$

Corollary 3.1. *Assume the assumptions of Lemma 3.1 and in addition that both $w(\cdot)$ and $v(\cdot)$ are constant (so that, in particular, conditional homoskedasticity holds true). Then*

$$\sqrt{n}(\hat{\beta}_W - \beta) \xrightarrow{d} N\big(0, \Omega_{1/v}^{-1}\big) \ .$$

It is well known that, under conditional homoskedasticity, $\Omega_{1/v}^{-1}$ is the limiting variance of the OLS estimator. So as long as the skedastic function $w(\cdot)$ is constant, the limiting distribution of $\hat{\beta}_W$ is identical to the limiting distribution of $\hat{\beta}_{\text{OLS}}$ under conditional homoskedasticity.

Next, we consider the behavior of the WLS estimator based on an estimated skedastic function. Assume the parametric family of skedastic functions used to estimate $v(\cdot)$ is given by $v_\theta(\cdot)$, where $\theta = (\theta_1, \ldots, \theta_d)'$ varies in an open subset of $\mathbb{R}^d$. Note the true $v(\cdot)$ need not be specified by any $v_\theta(\cdot)$. However, we always specify a family $v_\theta(\cdot)$ that includes constant values σ^2, so as to always allow for conditional homoskedasticity. It is further tacitly assumed that $v_\theta(x) > 0$ on the support of x, so that $1/v_\theta(x)$ is well-defined with probability one. Assume that $1/v_\theta(\cdot)$ is differentiable at some fixed θ_0 in the following sense: there exists a vector-valued function of dimension $1 \times d$

$$r_{\theta_0}(x) = \big(r_{\theta_0,1}(x), \ldots, r_{\theta_0,d}(x)\big)$$

and a real-valued function $s_{\theta_0}(\cdot)$ such that

$$\left|\frac{1}{v_\theta(x)} - \frac{1}{v_{\theta_0}(x)} - r_{\theta_0}(x)(\theta - \theta_0)\right| \le \frac{1}{2}|\theta - \theta_0|^2 s_{\theta_0}(x) \ , \tag{3.9}$$

for all θ in some small open ball around θ_0 and all x in the support of the covariates. Evidently $r_{\theta_0}(x)$ is the gradient with respect to θ of $1/v_\theta(x)$. Next, we assume we have a consistent estimator $\hat{\theta}$ of θ_0 in the sense that

$$n^{1/4}|\hat{\theta} - \theta_0| \xrightarrow{P} 0 \ . \tag{3.10}$$

Of course, (3.10) holds if $\hat{\theta}$ is a $\sqrt{n}$-consistent estimator of θ_0. (The weaker condition may be useful if one lets the dimension d of the model increase with the sample size n.)

Theorem 3.1. *Assume conditions (3.9) and (3.10). Further assume*

$$\mathbb{E}\big[|x_i|^2 v(x_i)|r_{\theta_0}(x_i)|^2\big] < \infty \tag{3.11}$$

and

$$E\big[|x_i| \cdot |\varepsilon_i s_{\theta_0}(x_i)|\big] < \infty \ . \tag{3.12}$$

(Note that in the case the functions $r_{\theta_0}(\cdot)$ and $s_{\theta_0}(\cdot)$ can be taken to be uniformly bounded over the support of the covariates, then these two added assumptions (3.11) and (3.12) already follow from (A5) and (A6).)

Consider the estimator $\hat{\beta}_{WLS} := \hat{\beta}_{\hat{V}}$ given by (3.8) with W replaced by $\hat{W}$, and $\hat{W}$ is the diagonal matrix with (i,i) entry $v_{\hat{\theta}}(x_i)$. Then,

$$\sqrt{n}(\hat{\beta}_{WLS} - \beta) \xrightarrow{d} N\big(0, \Omega_{1/w}^{-1}\Omega_{v/w^2}\Omega_{1/w}^{-1}\big) \ , \tag{3.13}$$

where $v(\cdot)$ is the true skedastic function and $w(\cdot) := v_{\theta_0}(\cdot)$ corresponds to the limiting estimated skedastic function.

Remark 3.3. Actually, the proof shows that

$$\sqrt{n}(\hat{\beta}_{\text{WLS}} - \hat{\beta}_W) \xrightarrow{P} 0 \ , \tag{3.14}$$

where $\hat{\beta}_W$ is the WLS based on the known skedastic function $w(\cdot) = v_{\theta_0}(\cdot)$. ■

Corollary 3.2. *Assume the assumptions of Theorem 3.1 and in addition that both $v_{\theta_0}(\cdot)$ and $v(\cdot)$ are constant (so that, in particular, conditional homoskedasticity holds true). Then*

$$\sqrt{n}(\hat{\beta}_{WLS} - \beta) \xrightarrow{d} N\big(0, \Omega_{1/v}^{-1}\big) \ .$$

Remark 3.4 (Assumptions on the Parametric Specification $v_\theta(\cdot)$)**.** We need to argue that the estimation scheme based on a parametric specification $v_\theta(\cdot)$, as described in Subsection 3.2, satisfies the assumptions of Theorem 3.1. The specifications we apply in the numerical work, such as given in Subsection 3.2 are clearly smooth, but it needs to be argued that (3.10) holds for some θ_0, even under conditional heteroskedasticity. The technical arguments are given in Appendix B.2 in the Appendix. In particular, both under conditional homoskedasticity and under conditional heteroskedasticity, our proposed estimation scheme of the skedastic function leads to a nonrandom estimate $v_{\theta_0}(\cdot)$ in the limit, as assumed by Theorem 3.1.

Remark 3.5 (Efficiency of WLS under Homoskedasticity and Limiting Value θ_0.)**.** It is well known that, under conditional homoskedasticity, $\Omega_{1/v}^{-1}$ is the limiting variance of the OLS estimator. So as long as the skedastic function $w(\cdot) := v_{\theta_0}(\cdot)$ is constant, the limiting distribution of $\hat{\beta}_{\text{WLS}}$ is identical to the limiting distribution of $\hat{\beta}_{\text{OLS}}$ in this case.

In Appendix B.2, it is argued that the estimator $\hat{\theta}$ tends in probability to some θ_0. However, $v_{\theta_0}(\cdot)$ need not correspond to the true skedastic function $v(\cdot)$. Furthermore, even when $v(\cdot)$ is constant and the specification for $v_\theta(\cdot)$ nests conditional homoskedasticity, it may or may not be the case that $v_{\theta_0}(\cdot)$ is constant.

On the one hand, consider the specification (3.6). Then, using OLS when regressing ε^2 (or $\hat{\varepsilon}^2$) on the right-hand-side of (3.6) gives a limiting value of θ_0 that corresponds to the best linear predictor of $\mathbb{E}(\varepsilon_i^2|x_i)$. Hence, if $\mathbb{E}(\varepsilon_i^2|x_i)$ is constant, then so is $v_{\theta_0}(\cdot)$.

On the other hand, consider the specification (3.5), where $\log(\varepsilon_i^2)$ is modeled by a linear function of covariates. In such a case, OLS is consistent for θ_0, which corresponds to the best linear predictor of $\mathbb{E}\big[\log(\varepsilon_i^2)|x_i\big]$. In the homoskedastic case where $\mathbb{E}(\varepsilon_i^2|x_i)$ is constant, one does not necessarily have that

$$\mathbb{E}\Big\{\log\big[\max(\delta^2, \varepsilon_i^2)\big]\big|x_i\Big\} \quad \text{is constant.} \tag{3.15}$$

Of course, (3.15) would hold in the more structured case where ε_i and x_i are independent under conditional homoskedasticity. For example, this would be the case if (A6) is strengthened to

(A6') $\{x_i\}_{i=1}^n$ is a K-variate i.i.d. sample and ε_i is given by

$$\varepsilon_i = \sqrt{v(x_i)} \cdot z_i \ ,$$

where $v(\cdot) : \mathbb{R}^K \to \mathbb{R}_+$ is a nonrandom skedastic function and $\{z_i\}_{i=1}^n$ is a univariate i.i.d. sample with mean zero and variance one, independent of $\{x_i\}_{i=1}^n$.

But in general (3.15) may fail. Therefore, to ensure in general that there is asymptotic efficiency loss of using WLS instead of OLS under conditional homoskedasticity, one needs to use a specification of the form (3.6); otherwise, one must assume that when conditional homoskedasticity holds, so does (3.15).

Finally, since whenever $v_{\theta_0}(\cdot)$ is constant, OLS and WLS are asymptotically equivalent, then in such a case, OLS and ALS are asymptotically equivalent, as well. ■

4 Inference: OLS, WLS, and ALS

4.1 Description of the Inference Methods

In most applications, it is of additional interest to conduct inference for β, by computing confidence intervals for (linear combinations of) β or by carrying out hypothesis tests for (linear combinations of) β. Unfortunately, when $\hat{v}(\cdot)$ is not a consistent estimator of the skedastic function $v(\cdot)$, then the textbook inference based on the WLS estimator can be misleading, in the sense that confidence intervals do not have the correct coverage probabilities and hypothesis tests do not have the correct null rejection probabilities, even asymptotically. This is an additional reason why applied researchers have shied away from WLS estimation. The contribution of this section is to propose a method by which consistent inference for β based on the WLS estimator can be obtained even if $\hat{v}(\cdot)$ is an inconsistent estimator. Our proposal is simple and straightforward. The idea is rooted in inference for β based on the OLS estimator.

It is well known that under conditional heteroskedasticity (A6), the OLS standard errors are not consistent and the resulting inference is misleading (in the sense specified in the previous paragraph). As a remedy, theoreticians have proposed *heteroskedasticity consistent* (HC) standard errors. Such research dates back to Eicker (1963, 1967), Huber (1967), and White (1980). Further refinements have been provided by MacKinnon and White (1985) and Cribari-Neto (2004).

As is well known (e.g., Hayashi 2000, Proposition 2.1), under assumptions (A1)–(A5),

$$\sqrt{n}(\hat{\beta}_{\text{OLS}} - \beta) \xrightarrow{d} N\big(0, \text{Avar}(\hat{\beta}_{\text{OLS}})\big) \quad \text{with} \quad \text{Avar}(\hat{\beta}_{\text{OLS}}) = \Sigma_{xx}^{-1}\Omega\Sigma_{xx}^{-1} \ , \tag{4.1}$$

where the symbol $\xrightarrow{d}$ denotes convergence in distribution. By assumptions (A2) and (A4) and the continuous mapping theorem, $n(X'X)^{-1}$ is a consistent estimator of Σ_{xx}^{-1}. Therefore, the problem of consistently estimating $\text{Avar}(\hat{\beta}_{\text{OLS}})$ is reduced to finding a consistent estimator $\hat{\Omega}$ of Ω. Inference for β can then be based in the standard fashion on

$$\widehat{\text{Avar}}_{\text{HC}}(\hat{\beta}_{\text{OLS}}) := n^2(X'X)^{-1}\hat{\Omega}(X'X)^{-1} \ . \tag{4.2}$$

For now, we focus on the case where the parameter of interest is β_k, for some $1 \leq k \leq K$. The OLS estimator of β_k is $\hat{\beta}_{k,\text{OLS}}$ and its HC standard error[4] implied by (4.2) is

$$\text{SE}_{\text{HC}}(\hat{\beta}_{k,\text{OLS}}) := \sqrt{\frac{1}{n}\big[\widehat{\text{Avar}}_{\text{HC}}(\hat{\beta}_{\text{OLS}})\big]_{k,k}} \ . \tag{4.3}$$

Then, for example, a two-sided confidence interval for β_k with nominal level $1-\alpha$ is given by

$$\hat{\beta}_{k,\text{OLS}} \pm t_{n-K,1-\alpha/2} \cdot \text{SE}_{\text{HC}}(\hat{\beta}_{k,\text{OLS}}) \ , \tag{4.4}$$

where $t_{n-K,1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the t distribution with $n-K$ degrees of freedom.[5] Alternatively, hypothesis tests of the form $H_0 : \beta_k = \beta_{k,0}$ can be based on the test statistic

$$\frac{\hat{\beta}_{k,\text{OLS}} - \beta_{k,0}}{\text{SE}_{\text{HC}}(\hat{\beta}_{k,\text{OLS}})}$$

in conjunction with suitable quantiles of the t_{n-K} distribution as critical values.

As stated before, finding a consistent estimator of $\text{Avar}(\hat{\beta}_{\text{OLS}})$ reduces to finding a consistent estimator of Ω in (4.2). There exist five widely used such estimators in the literature, named HC0–HC4. They are all of the 'sandwich' form

$$\hat{\Omega} := \frac{1}{n}X'\hat{\Psi}X \quad \text{with} \quad \hat{\Psi} := \text{diag}\{\hat{\psi}_1, \ldots, \hat{\psi}_n\} \ . \tag{4.5}$$

[4] In our terminology, a standard error is an estimate of the standard deviation of an estimator rather than the actual standard deviation of the estimator itself.

[5] On asymptotic grounds, one could also use the $1-\alpha/2$ quantile of the standard normal distribution instead. Taking the quantile from the t_{n-K} distribution results in somewhat more conservative inference in finite samples and is the standard practice in statistical software packages.

Therefore, to completely define one of the HC estimators, it is sufficient to specify a typical element, $\hat{\psi}_i$, of the diagonal matrix $\hat{\Psi}$. In doing so, let $\hat{\varepsilon}_i$ denote the ith OLS residual given by

$$\hat{\varepsilon}_i := y_i - x_i'\hat{\beta}_{\text{OLS}} \ ,$$

let h_i denote the ith diagonal element of the 'hat' matrix $H := X(X'X)^{-1}X'$, and let $\bar{h}$ denote the grand mean of the $\{h_i\}_{i=1}^n$. The various HC estimators use the following specifications.

$$\begin{aligned}
&\text{HC0}: \ \hat{\psi}_i := \hat{\varepsilon}_i^2 \ , \\
&\text{HC1}: \ \hat{\psi}_i := \frac{n}{n-K} \cdot \hat{\varepsilon}_i^2 \ , \\
&\text{HC2}: \ \hat{\psi}_i := \frac{\hat{\varepsilon}_i^2}{(1-h_i)} \ , \\
&\text{HC3}: \ \hat{\psi}_i := \frac{\hat{\varepsilon}_i^2}{(1-h_i)^2} \ , \ \text{and} \\
&\text{HC4}: \ \hat{\psi}_i := \frac{\hat{\varepsilon}_i^2}{(1-h_i)^{\delta_i}} \quad \text{with} \quad \delta_i := \min\left\{4 \, , \frac{h_i}{\bar{h}}\right\} .
\end{aligned} \tag{4.6}$$

HC0 dates back to White (1980) but results in inference that is generally liberal in small to moderate samples. HC1–HC3 are various improvements suggested by MacKinnon and White (1985): HC1 uses a global degrees-of-freedom adjustment, HC2 is based on influential analysis, and HC3 approximates a jackknife estimator. HC4 is the most recent proposal by Cribari-Neto (2004) designed to also handle observations x_i with strong leverage.

Of the estimators HC0–HC3, the one that delivers the most reliable finite-sample inference is HC3; for example, see MacKinnon and White (1985), Long and Ervin (2000), and Angrist and Pischke (2009, Section 8.1). It is also the default option in several statistical software packages to carry out HC estimation, such as in the R function `vcov()`; for example, see Zeileis (2004). On the other hand, we are not aware of any simulation studies evaluating the performance of the HC4 estimator outside of Cribari-Neto (2004).[6]

It is a characteristic feature of a HC standard error of the form (4.2)–(4.3) that its variance is larger than the variance of the conventional standard error based on the assumption of conditional homoskedasticity:

$$\text{SE}_{\text{CO}}(\hat{\beta}_{k,\text{OLS}}) := \sqrt{s^2\left[(X'X)^{-1}\right]_{k,k}} \quad \text{with} \quad s^2 := \frac{1}{n-K}\sum_{i=1}^n \hat{\varepsilon}_i^2 \ . \tag{4.7}$$

(A HC standard error as well as the conventional standard error are functions of the data. They are therefore random variables and, in particular, have a variance.) As a result, inference based on a HC standard error tends to be liberal[7] in small samples, especially when there is no or only little

[6] His Monte Carlo study only considers a single parametric specification of the skedastic function $v(\cdot)$.

[7] This means that confidence intervals tend to undercover and that hypothesis tests tend to overreject under the null.

conditional heteroskedasticity. These facts have been demonstrated by Kauermann and Carroll (2001) analytically and by Long and Ervin (2000), Kauermann and Carroll (2001), Cribari-Neto (2004), and Angrist and Pischke (2009, Section 8.1), among others, via Monte Carlo studies.

As a rule-of-thumb remedy, Angrist and Pischke (2009, Section 8.1) propose to take the maximum of a HC standard error and the conventional standard error. Letting

$$\mathrm{SE}_{\max}(\hat{\beta}_{k,\mathrm{OLS}}) := \max\bigl\{\mathrm{SE}_{\mathrm{HC}}(\hat{\beta}_{k,\mathrm{OLS}}), \mathrm{SE}_{\mathrm{CO}}(\hat{\beta}_{k,\mathrm{OLS}})\bigr\} \ ,$$

a more conservative confidence interval for β_k is then given by

$$\hat{\beta}_{k,\mathrm{OLS}} \pm t_{n-K,1-\alpha/2} \cdot \mathrm{SE}_{\max}(\hat{\beta}_{k,\mathrm{OLS}}) \ . \tag{4.8}$$

(In particular, they recommend the use of the HC3 standard error.)

We next turn to inference on β_k based on the WLS estimator. The textbook solution is to assume that $\hat{v}(\cdot)$ is a consistent estimator for the skedastic function $v(\cdot)$ and to then compute a conventional standard error from the transformed data

$$\tilde{y}_i := \frac{y_i}{\sqrt{\hat{v}(x_i)}} \quad \text{and} \quad \tilde{x}_i := \frac{x_i}{\sqrt{\hat{v}(x_i)}} \ . \tag{4.9}$$

More specifically,

$$\mathrm{SE}_{\mathrm{CO}}(\hat{\beta}_{k,\mathrm{WLS}}) := \sqrt{\tilde{s}^2\bigl[(\tilde{X}'\tilde{X})^{-1}\bigr]_{k,k}} \quad \text{with} \quad \tilde{s}^2 := \frac{1}{n-K}\sum_{i=1}^{n}\tilde{\varepsilon}_i^2 \quad \text{and} \quad \tilde{\varepsilon}_i := \tilde{y}_i - \tilde{x}_i'\hat{\beta}_{\mathrm{WLS}} \ . \tag{4.10}$$

The problem is that this standard error is incorrect when $\hat{v}(\cdot)$ is not a consistent estimator and, as a result, a confidence interval for β_k based on the WLS estimator combined with this standard error generally does not have correct coverage probability, even asymptotically. In the absence of some supernal information on the skedastic function $v(\cdot)$, applied researchers cannot be confident about having a consistent estimator $\hat{v}(\cdot)$. Therefore, they have rightfully shied away from the textbook inference based on the WLS estimator. The safe 'solution' is to simply use the OLS estimator combined with a HC standard error. This *status quo* in applied economic research is succinctly summarized by Angrist and Pischke (2010, p.10):

> Robust standard errors, automated clustering, and larger samples have also taken the steam out of issues like heteroskedasticity and serial correlation. A legacy of White's (1980[a]) paper on robust standard errors, one of the most highly cited from the period, is the near death of generalized least squares in cross-sectional applied work.[8] In the interests of replicability, and to reduce the scope for errors, modern applied researchers often prefer simpler estimators though they might be giving up asymptotic efficiency.

[8] For cross-sectional data, generalized least squares equates to weighted least squares.

In contrast, we side with Leamer (2010) who views conditional heteroskedasticity as an opportunity, namely an opportunity to construct more efficient estimators and to obtain shorter confidence intervals by sensibly weighting the data. But such benefits should not come at the expense of valid inference when the model for the skedastic function is misspecified. To this end, ironically, the same tool that killed off the WLS estimator can be used to resurrect it.

Our proposal is simple: applied researchers should use the WLS estimator combined with a HC standard error.[9] Doing so allows for valid inference, under weak regularity conditions, even if the employed $\hat{v}(\cdot)$ is not a consistent estimator of the skedastic function $v(\cdot)$. Specifically, the WLS estimator is the OLS estimator applied to the transformed data (4.9). And, analogously, a corresponding HC standard error is also obtained from these transformed data. In practice, the applied researcher only has to transform the data and then do as he would have done with the original data instead: run OLS and compute a HC standard error.

Denote the HC standard error computed from the transformed data by $\mathrm{SE}_{\mathrm{HC}}(\hat{\beta}_{k,\mathrm{WLS}})$.

Then a confidence interval for β_k based on the WLS estimator is given by

$$\hat{\beta}_{k,\mathrm{WLS}} \pm t_{n-K,1-\alpha/2} \cdot \mathrm{SE}_{\mathrm{HC}}(\hat{\beta}_{k,\mathrm{WLS}}) \ . \tag{4.11}$$

As before, one might prefer a more conservative approach for small sample sizes using the maximum of a HC standard error and the conventional standard error:

$$\mathrm{SE}_{\max}(\hat{\beta}_{k,\mathrm{WLS}}) := \max\big\{\mathrm{SE}_{\mathrm{HC}}(\hat{\beta}_{k,\mathrm{WLS}}), \mathrm{SE}_{\mathrm{CO}}(\hat{\beta}_{k,\mathrm{WLS}})\big\} \ ,$$

resulting in the confidence interval

$$\hat{\beta}_{k,\mathrm{WLS}} \pm t_{n-K,1-\alpha/2} \cdot \mathrm{SE}_{\max}(\hat{\beta}_{k,\mathrm{WLS}}) \ . \tag{4.12}$$

(In particular, we recommend the use of the HC3 standard error, or perhaps even the HC4 standard error.)

Remark 4.1 (Adaptive Least Squares; Remark 3.1 continued)**.** Should a researcher prefer ALS for the estimation of β, he generally also needs a corresponding method for making inference on β.

The method then is straightforward. If the ALS estimator is equal to the OLS estimator, use the confidence interval (4.4) or even the confidence interval (4.8). If the ALS estimator is equal to the WLS estimator, use the confidence interval (4.11) or even the confidence interval (4.12).

Note that in this setting, the test for conditional heteroskedasticity 'determines' the inference method but not in the way it has been generally promoted in the literature to date: namely, always use the OLS estimator and then base inference on a HC standard error (4.3) if the test rejects and on the conventional standard error (4.7) otherwise. This practice is *not* recommended since, under conditional heteroskedasticity, an invalid inference method (based on the conventional standard

[9] In spite of its simplicity, we have not seen this proposal anywhere else so far.

error) will be chosen with non-negligible probability in small to moderate samples because the power of the test is not near one. As a result, the finite-sample properties of this practice, under conditional heteroskedasticity, are poor in small to moderate samples; for example, see Long and Ervin (2000, Section 4.3).

In contrast, our proposal does not incur such a problem, since the pretest instead decides between two inference methods that are *both* valid under conditional heteroskedasticity. ■

So far, we have only discussed inference for a generic component, β_k, of β. The extension to more general inference problems is straightforward and detailed in Appendix A.

4.2 Consistent Estimation of the Limiting Covariance Matrix

We now consider estimating the unknown limiting covariance matrix of the WLS estimator, which recalling (3.13) is given by

$$\Omega_{1/w}^{-1}\Omega_{v/w^2}\Omega_{1/w}^{-1} ,$$

where, again, $w(\cdot) \coloneqq v_{\theta_0}(\cdot)$ and $v(\cdot)$ is the true skedastic function. First, $\Omega_{1/w}$ is estimated by

$$\hat{\Omega}_{1/w} \coloneqq \frac{X'\hat{W}^{-1}X}{n} = \frac{X'V_{\hat{\theta}}^{-1}X}{n} . \tag{4.13}$$

Second, we are left to consistently estimate Ω_{v/w^2}, which we recall is just

$$\Omega_{v/w^2} = \mathbb{E}\left(\frac{v(x_i)}{v_{\theta_0}^2(x_i)} \cdot x_i x_i'\right) = \mathbb{E}\left(\frac{\varepsilon_i^2}{v_{\theta_0}^2(x_i)} \cdot x_i x_i'\right) . \tag{4.14}$$

Of course, by the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\varepsilon_i^2}{v_{\theta_0}^2(x_i)} \cdot x_i x_i'\right) \xrightarrow{P} \Omega_{v/w^2} .$$

We do not know $v_{\theta_0}(x_i)$, but it can be estimated by $v_{\hat{\theta}}(x_i)$. In addition, we do not observe the true errors, but they can be estimated by the residuals after some consistent model fit. So given some consistent estimator $\hat{\beta}$, such as the ordinary least squares estimator, define the ith residual by

$$\hat{\varepsilon}_i \coloneqq y_i - x_i\hat{\beta} = \varepsilon_i - x_i'(\hat{\beta} - \beta) . \tag{4.15}$$

The resulting estimator of (4.14) is then

$$\hat{\Omega}_{v/w^2} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{\varepsilon}_i^2}{v_{\hat{\theta}}^2(x_i)} \cdot x_i x_i'\right) . \tag{4.16}$$

Furthermore, note that (3.9) implies that there exists a real-valued function $R_{\theta_0}(\cdot)$ such that

$$\left|\frac{1}{v_{\theta}^2(x)} - \frac{1}{v_{\theta_0}^2(x)}\right| \leq R_{\theta_0}(x)|\theta - \theta_0| \tag{4.17}$$

for all θ in some small open ball around θ_0 and all x in the domain of the covariates.

Theorem 4.1. *Assume the conditions of Theorem 3.1. Consider the estimator* $\hat{\Omega}_{1/w}^{-1}\hat{\Omega}_{v/w^2}\hat{\Omega}_{1/w}^{-1}$, *where* $\hat{\Omega}_{1/w}$ *is given in (4.13) and* $\hat{\Omega}_{v/w^2}$ *is given in (4.16). Then,*

$$\hat{\Omega}_{1/w}^{-1}\hat{\Omega}_{v/w^2}\hat{\Omega}_{1/w}^{-1} \xrightarrow{P} \Omega_{1/w}^{-1}\Omega_{v/w^2}\Omega_{1/w}^{-1} \ , \tag{4.18}$$

provided the following moment conditions are satisfied:

$$\mathbb{E}\big[|x_{ij}x_{ik}x_{il}x_{im}/v_{\theta_0}^2(x_i)|\big] < \infty \ , \tag{4.19}$$

$$\mathbb{E}\big[|x_{ij}x_{ik}x_{il}\varepsilon_i/v_{\theta_0}^2(x_i)|\big] < \infty \ , \tag{4.20}$$

and

$$\mathbb{E}\big[|x_i|^2\varepsilon_i^2 R_{\theta_0}(x_i)|\big] = \mathbb{E}\big[|x_i|^2 v(x_i)R_{\theta_0}(x_i)|\big] < \infty \ . \tag{4.21}$$

4.3 Asymptotic Validity of the Inference Methods

It is easy to see that the estimator $\hat{\Omega}_{1/w}^{-1}\hat{\Omega}_{v/w^2}\hat{\Omega}_{1/w}^{-1}$ is none other than the HC0 described in (4.5) and (4.6). Of course, having proven consistency of the HC0 estimator, consistency of the HC1 estimator follows immediately. For motivations to use, alternatively, the estimators HC2–HC4, see MacKinnon and White (1985) and Cribari-Neto (2004).

Being able to consistently estimate the limiting covariance matrix of the WLS estimator results in validity of the corresponding inference methods detailed in Subsection 4.1.

As far as inference based on the ALS estimator is concerned, there are two cases to consider. In the first case, the limiting skedastic function $v_{\theta_0(\cdot)}$ is constant. In this case (see Remark 3.3),

$$\sqrt{n}\big(\hat{\beta}_{\text{OLS}} - \hat{\beta}_{\text{WLS}}\big) \xrightarrow{P} 0 \ .$$

Therefore, there is a common limiting distribution for all three estimators — OLS, WLS, and ALS — and validity of the inference based on the ALS estimator follows from the validity of the inference of the other two estimators. In the second case, the limiting skedastic function $v_{\theta_0}(\cdot)$ is not constant. In this case, the ALS estimator will be equal to the WLS estimator with probability tending to one and the validity of the inference based on the ALS estimator follows from the validity of the inference based on the WLS estimator.

5 Monte Carlo Study

5.1 Basic Set-Up

We consider the simple regression model

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \ , \tag{5.1}$$

based on an i.i.d. sample $\{(y_i, x_i)\}_{i=1}^n$. In our design, $x_i \sim U[1,4]$ and

$$\varepsilon_i := \sqrt{v(x_i)} \cdot z_i \ , \tag{5.2}$$

where $z_i \sim N(0,1)$, and z_i is independent of x_i. The sample size is $n \in \{20, 50, 100\}$. The parameter of interest is β_2.

When generating the data, we consider four parametric specifications for the skedastic function $v(\cdot)$. First, $v(\cdot)$ is a power function:

$$v(x) = x^\gamma \ , \quad \text{with} \quad \gamma \in \{0, 1, 2, 4\} \ . \tag{5.3}$$

This specification includes conditional homoskedasticity for the choice $\gamma = 0$. Second, $v(\cdot)$ is a power of the log function:

$$v(x) = \big[\log(x)\big]^\gamma \ , \quad \text{with} \quad \gamma \in \{2, 4\} \ . \tag{5.4}$$

Third, $v(\cdot)$ is the exponential of a second-degree polynomial:

$$v(x) = \exp(\gamma x + \gamma x^2) \ , \quad \text{with} \quad \gamma \in \{0.1, 0.15\} \ . \tag{5.5}$$

Fourth, $v(\cdot)$ is a power of a step function:

$$v(x) = \begin{cases} 1^\gamma \ , & 1 \le x < 2 \\ 2^\gamma \ , & 2 \le x < 3 \\ 3^\gamma \ , & 3 \le x \le 4 \end{cases} \ , \quad \text{with} \quad \gamma \in \{1, 2\} \ . \tag{5.6}$$

The four specifications are graphically displayed in Figures 1–4. Note that for ease of interpretation, we actually plot $\sqrt{v(x)}$ as a function, since $\sqrt{v(x)}$ corresponds to the conditional standard deviation and thus lives on the same scale as x.

The parametric model used for estimating the skedastic function is

$$v_\theta(x) = \exp\big(\nu + \gamma \log|x|\big) \ , \quad \text{with} \quad \theta := (\nu, \gamma)' \ . \tag{5.7}$$

The model assumed for the skedastic function is correctly specified in (5.3) (with $\nu = 0$) and it is misspecified in (5.4)–(5.6). We estimate ν and γ from the data by the OLS regression

$$\log\big[\max(\delta^2, \hat\varepsilon_i^2)\big] = \nu + \gamma \log|x_i| + u_i \ , \tag{5.8}$$

where the $\hat\varepsilon_i$ are the OLS residuals of (5.1) and δ is chosen as $\delta = 0.1$ throughout. The resulting estimator of (ν, γ) is denoted by $(\hat\nu, \hat\gamma)$. WLS is then based on

$$\hat v(x) := \exp(\hat\nu + \hat\gamma \log x) \ . \tag{5.9}$$

Remark 5.1 (Choice of the Parametric Specification $v_\theta(\cdot)$)**.** As explained in Remark 3.5, under conditional homoskedasticity, WLS is asymptotically as efficient as OLS when using a specification of the form (3.6), but not necessarily using the form (3.5) as chosen in this Monte Carlo study. The reasoning for preferring (3.5) in empirical work is two-fold.

First, the specification (5.7) is equivalent to

$$v_\theta(x) = \sigma^2|x|^\gamma \ , \quad \text{with } \nu = \log \sigma^2 \ .$$

Therefore, estimating such a specification, implicitly, estimates the 'best' power of $|x|$ for modeling the skedastic function.[10] On the other hand, a specification of the form (3.6) would boil down to

$$v_\theta(x) = \nu + \gamma|x| \ .$$

This specification sets the power of $|x|$ equal to one, which may be far from optimal. One approach would be to choose another power, but then which one? A reasonable solution here would be to choose the power based on some residual plots (where the residuals are obtained from a first-stage OLS fit); but, clearly, such a method is difficult to implement in a Monte Carlo study. Another approach would be to include more than one power on the right-hand side, such as both $|x|$ and $|x|^2$. Again, it is not be clear which powers are 'best', and including many powers will result in inflated estimation uncertainty.

Second, a specification of the form (3.6) does not guarantee positivity of the weights $v_{\hat\theta}(x_i)$ to be used for WLS. Of course, there are several solutions to this problem, such as restricting the estimate $\hat\theta$ in a suitable fashion. But, again, it is not necessarily clear which such solution should be used in practice.

As an alternative, we also experimented with the specification

$$v_\theta(x) = \nu + \gamma_1|x| + \gamma_2|x|^2 \ ,$$

which guarantees that WLS is asymptotically equivalent to OLS; see Remark 3.5.[11] Compared to the specification (5.7), the results for WLS and ALS were similar under conditional homoskedasticity but worse under conditional heteroskedasticity. ■

5.2 Estimation

We consider the following three estimators of β_2.

- **OLS:** The OLS estimator of β_2.

[10] In particular, this specification is the one proposed by Judge et al. (1988, Section 9.3) for the case of a single covariate (in addition to the constant potentially).

[11] This specification might result in some non-positive weights $v_{\hat\theta}(x_i)$; such weights were then all set equal to a small, positive number.

- **WLS:** The WLS estimator of β_2 based on $\hat{v}(\cdot)$ given by (5.9).
- **ALS:** The ALS estimator of β_2 of Remark 3.1. The test for conditional heteroskedasticity used rejects the null of conditional homoskedasticity if $nR^2 > \chi^2_{1,0.9}$, where R^2 is the R^2-statistic from the OLS regression (5.8) and $\chi^2_{1,0.9}$ is the 0.9 quantile of the chi-squared distribution with one degree of freedom.

The performance measure is the empirical mean squared error (eMSE). For a generic estimator $\tilde{\beta}_2$, it is defined as

$$\text{eMSE}(\tilde{\beta}_2) := \frac{1}{B}\sum_{b=1}^{B}(\tilde{\beta}_{2,b} - \beta_2)^2 \ ,$$

where B denotes the number of Monte Carlo repetitions and $\tilde{\beta}_{2,b}$ denotes the outcome of $\tilde{\beta}_2$ in the bth repetition. The simulations are based on $B = 50,000$ Monte Carlo repetitions. Without loss of generality, we set $(\beta_1, \beta_2) = (0, 0)$ when generating the data.

The results are presented in Tables 1–2 and can be summarized as follows.

- As expected, in the case of conditional homoskedasticity (that is, in specification (5.3) with $\gamma = 0$), OLS is more efficient than WLS. But the differences are rather small and decreasing in n. In the worst case, $n = 20$, the ratio of the two eMSE's (WLS/OLS) is only 1.12.
- When there is conditional heteroskedasticity, WLS is generally more efficient than OLS. Only when the degree of conditional heteroskedasticity is low and the sample is small ($n = 20$) can OLS be more efficient, though the differences are always small.
- When the degree of conditional heteroskedasticity is high and the sample size is large, the differences between OLS and WLS can be vast, namely, the ratio (WLS/OLS) can be in the neighborhood of 0.3.
- ALS sacrifices some of the efficiency gains of WLS under conditional heteroskedasticity, especially when the sample size is small. On the other hand, it is closer to the performance of OLS under conditional homoskedasticity
- The previous statements hold true even when the model used to estimate the skedastic function is misspecified.

In sum, using WLS offers possibilities of vast improvements over OLS in terms of mean squared error while incurring only modest downside risk. ASL constitutes an attractive compromise between WLS and OLS.

Remark 5.2 (Nonnormal Error Terms)**.** To save space, we only report results when the distribution of the z_i in (5.2) is standard normal. However, we carried out additional simulations changing this distribution to a t-distribution with five degrees of freedom (scaled to have variance one) and a chi-squared distribution with five degrees of freedom (centered and scaled to have variance one). In both cases, the numbers for the two eMSE's increase compared to the normal distribution but their

ratios remain virtually unchanged. Therefore, the preceding summary statements appear 'robust' to nonnormality of the error terms. ■

Remark 5.3 (Failure of Assumption (A.6'))**.** The scheme (5.2) to generate the error terms ε_i satisfies assumption (A6') of Remark 3.5. It is therefore mathematically guaranteed that even the specification (5.7) guarantees that WLS and ALS are asymptotically as efficient as OLS under conditional homoskedasticity.

To study the impact of the failure of (A6') on the finite-sample performance under conditional homoskedasticity, we also consider error terms of the following form in specification (5.4) with $\gamma = 0$:

$$\varepsilon_i := \begin{cases} z_{i,1} & \text{if } x_i < 2\,, \quad \text{where } z_{i,1} \sim N(0,1)\,, \\ z_{i,2} & \text{if } 2 \leq x_i < 3\,, \quad \text{where } z_{i,2} \sim t_5^*\,, \text{ and} \\ z_{i,3} & \text{if } 3 \leq x_i < 4\,, \quad \text{where } z_{i,3} \sim \chi_5^{2,*}\,. \end{cases} \tag{5.10}$$

Here, t_5^* denotes a t-distribution with five degrees of freedom (scaled to have variance one) and $\chi_5^{2,*}$ denotes a chi-squared distribution with five degrees of freedom (centered and scaled to have variance one). The results are presented at the bottom of Table 1. It can be seen that even if assumption (A.6') does not hold, the efficiency loss of WLS and ALS compared to OLS under conditional homoskedasticity may still tend to zero as the sample size tends to infinity. ■

5.3 Inference

We next study the finite-sample performance of the following six confidence intervals for β_2.

- **OLS-HC:** The interval (4.4).
- **OLS-Max:** The interval (4.8).
- **WLS-HC:** The interval (4.11).
- **WLS-Max:** The interval (4.12).
- **ALS-HC:** The ALS inference of Remark 4.1 based on either interval (4.4) or interval (4.11).
- **ALS-HC:** The ALS inference of Remark 4.1 based on either interval (4.8) or interval (4.12).

There are two performance measures: first, the empirical coverage probability of a confidence interval with nominal confidence level $1-\alpha = 95\%$; and second, the ratio of the average length of a confidence interval over the average length of OLS-HC. (By construction, this ratio is independent of the nominal level.) Again, the simulations are based on $B = 50,000$ Monte Carlo replications. Again, without loss of generality, we set $(\beta_1, \beta_2) = (0, 0)$ when generating the data.

The results are presented in Tables 3–4 and can be summarized as follows.

- The coverage properties of all six intervals are at least satisfactory. Nevertheless, the HC intervals can undercover somewhat for small sample sizes. This problem is mitigated by using the Max intervals instead, at the expense of increasing the average lengths somewhat.

Overall, the three HC intervals have comparable coverage to each other and three Max intervals have comparable coverage to teach other, as well.

- Although there are only minor differences in terms of coverage (within the HC type and the Max type, respectively), there can be major differences in terms of average length. On average, WLS-HC and ALS-CH are never longer than OLS-HC but they can be dramatically shorter in the presence of strong conditional heteroskedasticity and in extreme cases only about half as long. The findings are the same when comparing WLS-Max and ALS-Max to OLS-Max.
- The previous statements hold true even when the model used to estimate the skedastic function is misspecified.

In sum, confidence intervals based on WLS or ALS offer possibilities of vast improvements over OLS in terms of expected length. This benefit does not come at any noticeable expense in terms of noticeably inferior coverage properties.

Remark 5.4 (Nonnormal Error Terms)**.** To save space, we only report results when the distribution of the z_i in (5.2) is standard normal. However, we carried out additional simulations changing this distribution to a t-distribution with five degrees of freedom (scaled to have variance one) and a chi-squared distribution with five degrees of freedom (centered and scaled to have variance one).

For the case of the t-distribution, empirical coverage probabilities generally slightly increase; for the case of the chi-squared distribution, they decrease and can fall below 92% for the HC confidence intervals and below 93% for the Max confidence intervals. Nevertheless, for both cases, OLS-HC continues to have comparable coverage performance to WLS-HC and OLS-Max continues to have comparable coverage performance to WLS-Max.

Furthermore, in both cases, the ratios of average lengths remain virtually unchanged compared to the normal distribution.

Therefore, the preceding summary statements appear 'robust' to nonnormality of the error terms. ■

Remark 5.5 (Hypothesis Tests)**.** By the well-understood duality between confidence intervals and hypothesis tests, we can gain the following insights. Hypothesis tests on β_k based on WLS or ALS offer possibilities of vast improvements over hypothesis tests based on OLS in terms of power while incurring basically no downside risk. This benefit does not come at any noticeable expense in terms of elevated null rejection probabilities. ■

6 Variations and Extensions

We briefly discuss a few natural variations and extensions to the proposed methodology.

- In this paper, we have focused on standard inference based on asymptotic normality of an estimator coupled with an estimate of the limiting covariance matrix. An anticipated criticism is that, by trying to estimate the true skedastic function, increased error in finite samples may result. But, increased efficiency results in shorter confidence intervals. If coverage error were too high in finite samples (though our simulations indicate adequate performance), the conclusion should not be to abandon weighted least squares, but to consider alternative inference methods that offer improved higher-order asymptotic accuracy (and thus translate to improved finite-sample performance). For example, one can consider bootstrap methods. In our setting, such inference would correspond to using the WLS estimator in combination with either the pairs bootstrap (e.g., see Efron and Tibshirani, 1993, Section 9.5) or the wild bootstrap (e.g., see Davison and Hinkley, 1997, Section 6.2), since these two bootstrap methods are appropriate for regression models that allow for conditional heteroskedasticity; a recent comparison for OLS estimation is provided in Flachaire (2005). As an alternative to bootstrapping, one can consider higher-order accuracy by using Edgeworth expansions, as studied in Hausman and Palmer (2012). It is beyond the scope of this paper to establish the asymptotic validity of such schemes applied to WLS and to study their finite-sample performance. Consequently, we leave such topics for future research.

- In this paper, we have focused on the case of a stochastic design matrix X, which is the relevant case for economic applications. Alternatively, it would be possible handle the case of a nonstochastic design matrix X, assuming certain regularity conditions on the asymptotic behavior of X, such as in Amemiya (1985, Section 3.5).

- Our goal in the present work is to primarily offer enough evidence to change the current practice by showing that improvements offered by weighted least squares are nontrivial. A more ambitious goal would be to estimate the skedastic function $v(\cdot)$ in a nonparametric fashion. For example, one could use a sieve of parametric models by allowing the number of covariates used in the modeling of $v(\cdot)$ to increase with n. Of course, nonparametric smoothing techniques could be used as well. The hope would be further gains in efficiency, which ought to be possible.

- Finally, it would be of interest to extend the proposed methodology to the context of *instrumental variables* (IV) regression. HC inference of the HC0–HC1 type based on two-stage least squares (2SLS) estimation is already standard; for example, see Hayashi (2000, Section 3.5). On the other hand, improved HC inference of the HC2–HC3 type is still in its infancy; for example, see Steinhauer and Würgler (2010). To the best of our knowledge, weighted two-stage least squares (W2SLS) estimation has not been considered at all yet in the context of IV regressions. Therefore, also this topic is beyond the scope of the paper.

7 Conclusion

As the amount of data collected is ever growing, the statistical toolbox of applied researchers is ever expanding. Nevertheless, it can be safely assumed that linear models will remain a part of our toolbox for quite some time to come.

A textbook analysis of linear models always starts with an assumption of conditional homoskedasticity, that is, an assumption that the conditional variance of the error term is constant. Under such an assumption, one should estimate model parameters by *ordinary least squares* (OLS), as doing so is efficient. Unfortunately, the real world is plagued by conditional heteroskedasticity, since the conditional variance often depends on the explanatory variables. In such a setting, OLS is no longer efficient. If the true functional form of the conditional variance (that is, the *skedastic function*) were known, efficient estimators of model parameters could be constructed by properly weighting the data (using the inverse of square root of the skedastic function) and running OLS on the weighted data set. Of course, the true skedastic function is rarely known. In the olden days, applied researchers resorted to weighting the data based on an estimate of the skedastic function, resulting in the *weighted least squares* (WLS) estimator.

Under conditional heteroskedasticity, textbook inference based the OLS estimator can be misleading. But the same is true for textbook inference based on the WLS estimator, unless the model for estimating the skedastic function is correctly specified. These shortcomings have motivated the development of heteroskedasticity consistent (HC) standard errors for the OLS estimator. Such standard errors ensure the (asymptotic) validity of inference based on the OLS estimator in the presence of conditional heteroskedasticity of unknown form. Over time, applied researchers have by and large adopted this practice, causing WLS to become extinct for all practical purposes.

In this paper, we promote the use of HC standard errors for the WLS estimator instead. This practice ensures (asymptotic) validity of inference based on the WLS estimator even when the model for estimating the skedastic function is misspecified. The benefits of our proposal in the presence of noticeable conditional heteroskedasticity are two-fold. First, using WLS generally results in more efficient estimation. Second, HC inference based on WLS has more attractive properties in the sense that confidence intervals for model parameters tend to be shorter and hypothesis tests tend to be more powerful. The price to pay is some efficiency loss compared to OLS results in small samples in the textbook setting of conditional homoskedasticity.

As a bonus, we propose a new *adaptive least squares* (ALS) estimator, where a pretest on conditional homoskedasticity is used in order to decide whether to weight the data (that is, whether to use WLS) or not (that is, to use OLS). Crucially, in either case, one uses HC standard errors so that (asymptotically) valid inference is ensured.

Having no longer to live in fear of invalid inference, applied researchers should rediscover their long-lost friend the WLS estimator, or its new companion the ALS, estimator: the benefits over their current company, the OLS estimator, can be substantial.

References

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press, Cambridge, MA.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton University Press, Princeton, New Yersey.

Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. Discussion Paper Series No. 4800, IZA.

Breusch, T. and Pagan, A. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47:1287–1294.

Cribari-Neto, F. (2004). Asymptotic inference under heterskedasticty of unknown form. *Computational Statistics & Data Analysis*, 45:215–233.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

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

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In LeCam, L. M. and Neyman, J., editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 59–82, Berkeley, CA. University of California Press.

Flachaire, E. (2005). Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap. *Computational Statistics & Data Analysis*, 49:361–377.

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

Hausman, J. and Palmer, C. (2012). Heteroskedasticity-robust inference in finite samples. *Economic Letters*, 116:232–235.

Hayashi, F. (2000). *Econometrics*. Princeton University Press, Princeton, New Jersey.

Huber, P. (1967). The behavior of maximum likelihood estimation under nonstandard conditions. In LeCam, L. M. and Neyman, J., editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233, Berkeley, CA. University of California Press.

Judge, G. G., Hill, R. C., Griffiths, W. E., Lütkepohl, H., and Lee, T.-C. (1988). *Introduction To The Theory And Practice Of Econometrics*. John Wiley & Sons, New York, second edition.

Kauermann, G. and Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456):1387–1396.

Koenker, R. (1981). A note on studentizing a test for heteroscedasticity. *Journal of Econometrics*, 17:107–112.

Koenker, R. and Bassett, G. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, 50:43–61.

Leamer, E. E. (2010). Tantalus on the road to asymptotia. *Journal of Economic Perspectives*, 24(2):31–46.

Long, J. S. and Ervin, L. H. (2000). Using heteroskedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54:217–224.

MacKinnon, J. G. and White, H. L. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite-sample properties. *Journal of Econometrics*, 29:53–57.

Steinhauer, A. and Würgler, T. (2010). Leverage and covariance matrix estimation in finite-sample IV regressions. Working Paper 521, IEW, University of Zurich.

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

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, 11(10):1–17.

A More General Inference Problems

A.1 Inference for a Linear Combination

Generalize the parameter of interest from a component β_k to a linear combination $a'\beta$, where $a \in \mathbb{R}^K$ is vector specifying the linear combination of interest. The OLS estimator of $a'\beta$ is $a'\hat{\beta}_{\text{OLS}}$. A HC standard error is given by

$$\text{SE}_{\text{HC}}(a'\hat{\beta}_{\text{OLS}}) := \sqrt{\frac{1}{n} a' \big[\widehat{\text{Avar}}_{\text{HC}}(\hat{\beta}_{\text{OLS}})\big] a} \ ,$$

where $\widehat{\text{Avar}}_{\text{HC}}(\hat{\beta}_{\text{OLS}})$ is as described in Subsection (4.1). The conventional standard error is given by

$$\text{SE}_{\text{CO}}(a'\hat{\beta}_{\text{OLS}}) := \sqrt{s^2 a' \big[(X'X)^{-1}\big] a} \quad \text{with} \quad s^2 := \frac{1}{n-K} \sum_{i=1}^{2} \hat{\varepsilon}_i^2 \quad \text{and} \quad \hat{\varepsilon}_i := y_i - x_i' \hat{\beta}_{\text{OLS}} \ .$$

The WLS estimator of $a'\beta$ is $a'\hat{\beta}_{\text{WLS}}$. A HC standard error is given by

$$\text{SE}_{\text{HC}}(a'\hat{\beta}_{\text{WLS}}) := \sqrt{\frac{1}{n} a' \big[\widehat{\text{Avar}}_{\text{HC}}(\hat{\beta}_{\text{WLS}})\big] a} \ ,$$

where $\widehat{\text{Avar}}_{\text{HC}}(\hat{\beta}_{\text{WLS}})$ is as described in Subsection (4.1). The conventional standard error is given by

$$\text{SE}_{\text{CO}}(a'\hat{\beta}_{\text{WLS}}) := \sqrt{\tilde{s}^2 a' \big[(\tilde{X}'\tilde{X})^{-1}\big] a} \quad \text{with} \quad \tilde{s}^2 := \frac{1}{n-K} \sum_{i=1}^{2} \tilde{\varepsilon}_i^2 \quad \text{and} \quad \tilde{\varepsilon}_i := \tilde{y}_i - \tilde{x}_i' \hat{\beta}_{\text{WLS}} \ .$$

From here on, the extension of the inference methods for β_k discussed in Subsection 4.1 to inference methods for $a'\beta$ is clear.

A.2 Testing a Set of Linear Restrictions

Consider testing a set of linear restrictions on β of the form

$$H_0 : R\beta = r \ ,$$

where $R \in \mathbb{R}^{p \times K}$ is matrix of full row rank specifying $p \leq K$ linear combinations of interest and $r \in \mathbb{R}^p$ is a vector specifying their respective values under the null.

A HC Wald statistic based on the OLS estimator is given by

$$W_{\text{HC}}(\hat{\beta}_{\text{OLS}}) := \frac{n}{p} \cdot (R\hat{\beta}_{\text{OLS}} - r)' \big[R\, \widehat{\text{Avar}}_{\text{HC}}(\hat{\beta}_{\text{OLS}}) R'\big]^{-1} (R\hat{\beta}_{\text{OLS}} - r)$$

and its conventional counterpart is given by

$$W_{\text{CO}}(\hat{\beta}_{\text{OLS}}) := \frac{n}{ps^2} \cdot (R\hat{\beta}_{\text{OLS}} - r)' \big[R(X'X)^{-1}R'\big]^{-1} (R\hat{\beta}_{\text{OLS}} - r) \ .$$

A HC Wald statistic based on the WLS estimator is given by

$$W_{\text{HC}}(\hat{\beta}_{\text{WLS}}) := \frac{n}{p} \cdot (R\hat{\beta}_{\text{WLS}} - r)' \big[R\, \widehat{\text{Avar}}_{\text{HC}}(\hat{\beta}_{\text{WLS}}) R'\big]^{-1} (R\hat{\beta}_{\text{WLS}} - r)$$

and its conventional counterpart is given by

$$W_{\text{CO}}(\hat{\beta}_{\text{WLS}}) := \frac{n}{p\tilde{s}^2} \cdot (R\hat{\beta}_{\text{WLS}} - r)' \big[R(\tilde{X}'\tilde{X})^{-1}R'\big]^{-1}(R\hat{\beta}_{\text{WLS}} - r) \ .$$

For a generic Wald statistic W, the corresponding p-value is obtained as

$$PV(W) := \text{Prob}\{F \geq \tilde{W}\} \ , \quad \text{where } F \sim F_{p,n} \ .$$

Here, $F_{p,n}$ denotes the F distribution with p and n degrees of freedom.

HC inference based on the OLS estimator reports $PV(W_{\text{HC}}(\hat{\beta}_{\text{OLS}}))$, while more conservative inference based on the OLS estimator reports

$$PV_{\max}(\hat{\beta}_{\text{OLS}}) := \max\big\{PV(W_{\text{HC}}(\hat{\beta}_{\text{OLS}})), PV(W_{\text{CO}}(\hat{\beta}_{\text{OLS}}))\big\} \ .$$

HC inference based on the WLS estimator reports $PV(W_{\text{HC}}(\hat{\beta}_{\text{WLS}}))$, while more conservative inference based on the WLS estimator reports

$$PV_{\max}(\hat{\beta}_{\text{WLS}}) := \max\big\{PV(W_{\text{HC}}(\hat{\beta}_{\text{WLS}})), PV(W_{\text{CO}}(\hat{\beta}_{\text{WLS}}))\big\} \ .$$

B Mathematical Results

B.1 Proofs

Proof of Lemma 3.1. Replacing y by $X\beta + \varepsilon$ in the definition of $\hat{\beta}_W$ in (3.8) yields

$$\sqrt{n}(\hat{\beta}_W - \beta) = \left(\frac{X'W^{-1}X}{n}\right)^{-1} \frac{X'W^{-1}\varepsilon}{\sqrt{n}} \ . \tag{B.1}$$

By Slutsky's Theorem, the proof consists in showing

$$\frac{X'W^{-1}X}{n} \xrightarrow{P} \Omega_{1/w} \tag{B.2}$$

and

$$\frac{X'W^{-1}\varepsilon}{n^{1/2}} \xrightarrow{d} N(0, \Omega_{v/w^2}) \ . \tag{B.3}$$

To show (B.2), its left side has (j,k) element given by

$$\frac{1}{n}\sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{w(x_i)} \xrightarrow{P} \mathbb{E}\left(\frac{x_{1j}x_{1k}}{w(x_1)}\right) \ ,$$

by the law of large numbers. To show (B.3), first note that

$$\mathbb{E}(X'W^{-1}\varepsilon) = \mathbb{E}\big[X'W^{-1}\mathbb{E}(\varepsilon|X)\big] = 0$$

by Assumption (A3). Furthermore, $X'W^{-1}\varepsilon$ is a sum of i.i.d. random vectors $x_i \cdot \varepsilon_i / w(x_i)$ with common covariance matrix having (j,k) element

$$\text{Cov}\left(\frac{x_{1j}\varepsilon_1}{w(x_1)}, \frac{x_{1k}\varepsilon_1}{w(x_1)}\right) = \mathbb{E}\left[\frac{x_{1j}x_{1k}\varepsilon_1^2}{w^2(x_1)}\right] = \mathbb{E}\left[\frac{x_{1j}x_{1k}}{w^2(x_1)}E(\varepsilon_1^2|x_1)\right] = \mathbb{E}\left[\frac{x_{1,j}x_{1k}v(x_1)}{w^2(x_1)}\right] \ .$$

Thus, each vector $x_i \cdot \varepsilon_i / w(x_i)$ has covariance matrix Ω_{v/w^2}. Therefore, by the multivariate Central Limit Theorem, (B.3) holds. ■

PROOF OF THEOREM 3.1. Let W be the diagonal matrix with (i,i) element $v_{\theta_0}(x_i)$. Similarly to (B.1), we have

$$\sqrt{n}(\hat{\beta}_{\text{WLS}} - \beta) = \left(\frac{X'\hat{W}^{-1}X}{n} \right)^{-1} \frac{X'\hat{W}^{-1}\varepsilon}{\sqrt{n}} \ . \tag{B.4}$$

First, we show that

$$\frac{X'\hat{W}^{-1}\varepsilon}{\sqrt{n}} - \frac{X'W^{-1}\varepsilon}{\sqrt{n}} \xrightarrow{P} 0 \ . \tag{B.5}$$

Even though the assumptions imply that $\hat{W}$ and W are close, one needs to exercise some care, as the dimension of these matrices increases with the sample size n. The left-hand side of (B.5) is

$$\frac{X'(\hat{W}^{-1} - W^{-1})\varepsilon}{\sqrt{n}} = n^{-1/2} \sum_{i=1}^{n} x_i \cdot \varepsilon_i \left(\frac{1}{v_{\hat{\theta}}(x_i)} - \frac{1}{v_{\theta_0}(x_i)} \right) = A + B \ ,$$

where

$$A := n^{-1/2} \sum_{i=1}^{n} x_i \varepsilon_i r_{\theta_0}(x_i)(\hat{\theta} - \theta_0) \ , \tag{B.6}$$

and, with probability tending to one, B is a vector with jth component satisfying

$$|B_j| \leq \frac{1}{2} n^{-1/2} |\hat{\theta} - \theta_0|^2 \sum_{i=1}^{n} |x_{ij} \varepsilon_i s_{\theta_0}(x_i)| \ . \tag{B.7}$$

The jth component of A is

$$n^{-1/2} \sum_{i=1}^{n} x_{ij} \varepsilon_i \sum_{l=1}^{K} r_{\theta_0, l}(x_i)(\hat{\theta}_l - \theta_l) \ .$$

So to show $A = o_P(1)$, it suffices to show that, for each j and l,

$$(\hat{\theta}_l - \theta_l) n^{-1/2} \sum_{i=1}^{n} x_{ij} \varepsilon_i r_{\theta_0, l}(x_i) \xrightarrow{P} 0 \ .$$

The first factor $(\hat{\theta} - \theta_0) = o_P(1)$, and so it suffices to show that

$$n^{-1/2} \sum_{i=1}^{n} x_{i,j} \varepsilon_i r_{\theta_0}(x_i) = O_P(1) \ .$$

The terms in this sum are i.i.d. random variables with mean zero and finite second moments, where finite second moments follow from (3.11), and so this normalized sum converges in distribution to a multivariate normal distribution. Therefore, $A = o_P(1)$. To show $|B| = o_P(1)$, write the right-hand side of (B.7) as

$$\frac{1}{2} \sqrt{n} |\hat{\theta} - \theta_0|^2 \frac{1}{n} \sum_{i=1}^{n} |x_{ij} \varepsilon_i s_{\theta_0}(x_i)| \ . \tag{B.8}$$

The first factor $\sqrt{n}|\hat{\theta} - \theta_0|^2 = o_P(1)$ by assumption while the average of the i.i.d. variables $|x_{ij}\varepsilon_i s_{\theta_0}(x_i)|$ obeys the law of large numbers by the moment assumption (3.12). Thus, $|B| = o_P(1)$ also and (B.5) holds.

Next, we show that

$$\frac{X'\hat{W}^{-1}X}{n} - \frac{X'W^{-1}X}{n} \xrightarrow{P} 0 \ . \tag{B.9}$$

To this end simply write (B.9) as

$$\frac{X'(\hat{W}^{-1} - W^{-1})X}{n} = \frac{1}{n}\sum_i x_i x_i' \left(\frac{1}{v_{\hat{\theta}}(x_i)} - \frac{1}{v_{\theta_0}(x_i)} \right) \ ,$$

and then use the differentiability assumption as above (which is even easier now because one only needs to invoke the law of large numbers and not the central limit theorem). It now also follows by the limit (B.2) and the fact that the limiting matrix there is positive definite that

$$\left(\frac{X'\hat{W}^{-1}X}{n} \right)^{-1} - \left(\frac{X'W^{-1}X}{n} \right)^{-1} \xrightarrow{P} 0 \ . \tag{B.10}$$

Then, the convergences (B.5) and (B.10) are enough to show that the right-hand side of (B.4) satisfies

$$\left(\frac{X'\hat{W}^{-1}X}{n} \right)^{-1} \frac{X'\hat{W}^{-1}\varepsilon}{\sqrt{n}} - \left(\frac{X'W^{-1}X}{n} \right)^{-1} \frac{X'W^{-1}\varepsilon}{\sqrt{n}} \xrightarrow{P} 0 \ ;$$

just by making simple use of the equality

$$\hat{a}\hat{b} - ab = \hat{a}(\hat{b} - b) + (\hat{a} - a)b \ .$$

Finally, Slutsky's theorem yields the result. ■

Proof of Theorem 4.1. First, the estimator (4.13) is consistent because of (B.2) and (B.9). To analyze (4.16) , we first consider the behavior of this estimator with $v_{\hat{\theta}}(\cdot)$ replaced by the fixed $v_{\theta_0}(\cdot)$, but retaining the residuals (instead of the true error terms). From (4.15) it follows that

$$\hat{\varepsilon}_i^2 = \varepsilon_i^2 - 2(\hat{\beta} - \beta)' x_i \cdot \varepsilon_i + (\hat{\beta} - \beta)' x_i \cdot x_i'(\hat{\beta} - \beta) \ .$$

Then, multiplying the last expression by $x_i x_i'/v_{\theta_0}^2(x_i)$ and averaging over i yields

$$\frac{1}{n}\sum_{i=1}^n \left(\frac{\hat{\varepsilon}_i^2}{v_{\theta_0}^2(x_i)} \cdot x_i x_i' \right) - \frac{1}{n}\sum_{i=1}^n \left(\frac{\varepsilon_i^2}{v_{\theta_0}^2(x_i)} \cdot x_i x_i' \right) = C_n + D_n \ , \tag{B.11}$$

where

$$C_n := -\frac{2}{n}\sum_{i=1}^n x_i x_i' \cdot (\hat{\beta} - \beta)' x_i \cdot \varepsilon_i / v_{\theta_0}^2(x_i)$$

and

$$D_n := \frac{1}{n}\sum_{i=1}^n x_i x_i' \cdot (\hat{\beta} - \beta)' x_i x_i' (\hat{\beta} - \beta)/v_{\theta_0}^2(x_i) \ .$$

The first goal is to show both C_n and D_n tend to zero in probability. The (j,k) term in the matrix D_n is given by

$$D_n(j,k) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik}\sum_{l=1}^{K}(\hat{\beta}_l - \beta_l)x_{il}\sum_{m=1}^{K}(\hat{\beta}_m - \beta_m)x_{im}/v^2_{\theta_0}(x_i) \ .$$

Thus, it suffices to show that, for each j, k, l, and m,

$$(\hat{\beta}_l - \beta_l)(\hat{\beta}_m - \beta_m)\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik}x_{il}x_{im}/v^2_{\theta_i}(x_i) \xrightarrow{P} 0 \ . \tag{B.12}$$

But $(\hat{\beta}_l - \beta_l)(\hat{\beta}_m - \beta_m) \xrightarrow{P} 0$ and the average on the right-hand side of (B.12) satisfies the law of large numbers under the assumption of the "fourth moment condition" (4.19) and thus tends to something finite in probability. Therefore, (B.12) holds and so $D_n \xrightarrow{P} 0$.

Next, we show $C_n \xrightarrow{P} 0$. But, $(-1/2)$ times the (j,k) term of C_n is given by

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik}\sum_{l=1}^{K}(\hat{\beta}_l - \beta_l)x_{il}\varepsilon_i/v^2_{\theta_0}(x_i) \ .$$

So, it suffices to show that, for each j, k, and l,

$$(\hat{\beta}_l - \beta_l)\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik}x_{il}\varepsilon_i/v^2_{\theta_0}(x_i) \xrightarrow{P} 0 \ . \tag{B.13}$$

But $(\hat{\beta}_l - \beta_l) \xrightarrow{P} 0$ and the average on the right-hand side of (B.13) satisfies the law of large numbers under the assumption of the "fourth moment condition" (4.20) and thus tends to something finite in probability. Therefore, $C_n \xrightarrow{P} 0$.

In summary, what we have shown so far is that (B.11) tends to zero in probability. Thus, the proof of consistency will be complete if we can show that also

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\varepsilon_i^2}{v^2_{\hat{\theta}}(x_i)}\cdot x_i x_i'\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\varepsilon_i^2}{v^2_{\theta_0}(x_i)}\cdot x_i x_i'\right) \xrightarrow{P} 0 \ . \tag{B.14}$$

By property (4.17) of the function $R_{\theta_0}(\cdot)$, the left-hand-side of (B.14) has (j,k) component that can be bounded by the absolute value of

$$|\hat{\theta} - \theta_0|\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik}\varepsilon_i^2 R_{\theta_0}(x_i) \ . \tag{B.15}$$

But $(\hat{\theta} - \theta_0) \xrightarrow{P} 0$ and the average in (B.15) obeys the law of large numbers under the moment condition (4.21) and thus tends to something finite in probability. Therefore, (B.14) holds. ■

B.2 Verification of Assumptions for the Parametric Specification $v_\theta(\cdot)$

The main theorems assume the family $v_\theta(\cdot)$ leads to a $\hat{\theta}$ satisfying (3.10). Assume the family $v_\theta(\cdot)$ is of the form (which is slightly more general than (3.5))

$$v_\theta(x) := \exp\left[\sum_{j=1}^{d} \theta_j g_j(x)\right] , \tag{B.16}$$

where $\theta = (\theta_1, \ldots, \theta_d)'$ and $g(x) = (g_1(x), \ldots, g_d(x))'$. It is tacitly assumed that $g_1(x) = 1$ to ensure that this specification nests the case of conditional homoskedasticity. Fix $\delta > 0$ and let $h_\delta(\varepsilon) := \log\big[\max(\delta^2, \varepsilon^2)\big]$. The estimator $\hat{\theta}$ is obtained by regressing the residuals $\hat{\varepsilon}_i$, or more precisely $h_\delta(\hat{\varepsilon}_i)$ on $g(x_i)$. Before analyzing the behavior of $\hat{\theta}$, we first consider $\tilde{\theta}$, which is obtained by regressing $h_\delta(\varepsilon_i)$ on $g(x_i)$. (Needless to say, we do not know the ε_i, but we can view $\tilde{\theta}$ as an oracle 'estimator'.) As argued in Hayashi (2000, Section 2.9), $\tilde{\theta}$ is a consistent estimator of

$$\theta_0 := \big[\mathbb{E}(g(x_i)g(x_i)')\big]^{-1}\mathbb{E}\big[g(x_i) \cdot h_\delta(\varepsilon_i)\big] .$$

To show that $\tilde{\theta}$ is moreover $\sqrt{n}$-consistent, note that $\tilde{\theta} = L_n^{-1} m_n$, where L_n is the $d \times d$ matrix

$$L_n := \frac{1}{n}\sum_{i=1}^{n} g(x_i)g(x_i)'$$

and m_n is the $d \times 1$ vector

$$m_n := \frac{1}{n}\sum_{i=1}^{n} g(x_i) \cdot h_\delta(\varepsilon_i) .$$

Since L_n is an average of i.i.d. random matrices, it is a $\sqrt{n}$-consistent estimator of

$$L := \mathbb{E}\big[g(x_i)g(x_i)'\big]$$

(under the assumption of finite second moments of products and invertibility of L), and in fact is asymptotically multivariate normal as well.[12] Similarly, $\sqrt{n}(m_n - m)$ is asymptotically multivariate normal under moment conditions, where

$$m := \mathbb{E}\big[g(x_i) \cdot h_\delta(\varepsilon_i)\big] .$$

But, if L_n and m_n are each $\sqrt{n}$-consistent estimators of L and m, respectively, it is easy to see that $\tilde{\theta} = L_n \cdot m_n$ is a $\sqrt{n}$-consistent estimator of $L \cdot m = \theta_0$.[13]

However, our algorithm uses the residuals $\hat{\varepsilon}_i$ after an OLS fit of y_i on x_i, rather than the true errors ε_i. So, we must argue that the difference between $\tilde{\theta}$ above and $\hat{\theta}$ obtained when using the residuals is of order $o_P(n^{-1/4})$, which would then verify (3.10). Note that $\hat{\theta} = L_n \cdot \hat{m}_n$, where

$$\hat{m}_n := \frac{1}{n}\sum_{i=1}^{n} g(x_i) \cdot h_\delta(\hat{\varepsilon}_i) .$$

[12] Note that L is clearly invertible in the case $g(x) := (1, \log(x))'$ as used in the Monte Carlo study of Section 5.

[13] Alternatively, by the usual arguments that show asymptotic normality of OLS, under moment assumptions, $\sqrt{n}(\tilde{\theta} - \theta_0)$ is asymptotically normal, and hence $\sqrt{n}$-consistent.

Hence, it suffices to show that

$$\hat{m}_n - m = o_P\big(n^{-1/4}\big) \ . \tag{B.17}$$

To do this, first note that

$$\big|\max(\delta, |\hat{\varepsilon}_i|) - \max(\delta, |\varepsilon_i|)\big| \leq |\hat{\varepsilon}_i - \varepsilon_i| \ .$$

Then,

$$\begin{aligned}
\big|h_\delta(\hat{\varepsilon}_i) - h_\delta(\varepsilon_i)\big| &= \big|\log[\max(\delta^2, \hat{\varepsilon}_i^2)] - \log[\max(\delta^2, \varepsilon_i^2)]\big| \\
&= 2\big|\log[\max(\delta, |\hat{\varepsilon}_i|)] - \log[\max(\delta, |\varepsilon_i|)]\big| \\
&\leq \frac{2}{\delta}\big|\max(\delta, |\hat{\varepsilon}_i|)] - \max(\delta, |\varepsilon_i|)\big| \\
&\leq \frac{2}{\delta}|\hat{\varepsilon}_i - \varepsilon_i| \\
&= \frac{2}{\delta}|x_i'(\hat{\beta} - \beta)| \ ,
\end{aligned}$$

where the first inequality follows from the mean value theorem of calculus.

Therefore,

$$|\hat{m}_n - m| \leq \frac{2}{n\delta}\sum_{i=1}^n |g(x_i)| \cdot \big|x_i'(\hat{\beta} - \beta)\big| \ .$$

But assuming $\mathbb{E}\big|g_k(x_i) \cdot x_j\big| < \infty$ for any i, j, one can apply the law of large numbers to conclude that

$$|\hat{m}_n - m| = O_P\big(|\hat{\beta} - \beta|/\delta\big) = O_P\big(n^{-1/2}\big) \ ,$$

which certainly implies (B.17). As an added bonus, the argument shows that one can let $\delta := \delta_n \to 0$ as long as δ_n goes to zero slowly enough; in particular, as long as $\delta_n n^{1/4} \to \infty$.

The argument for the linear specification

$$v_\theta(x) := \sum_{j=1}^d \theta_j g_j(x)$$

is similar. Here, the estimator $\hat{\theta}$ is obtained by regressing the residuals $\hat{\varepsilon}_i^2$ on $g(x_i)$. As above, first consider $\tilde{\theta}$ obtained by regressing the actual errors ε_i^2 on $g(x_i)$. Then, $\tilde{\theta}$ is a consistent estimator of

$$\theta_0 := \big[\mathbb{E}(g(x_i)g(x_i)')\big]^{-1}\mathbb{E}\big[g(x_i) \cdot \varepsilon_i^2\big] \ .$$

As before, it is $\sqrt{n}$-consistent, as $\tilde{\theta} = L_n^{-1}m_n$, with L_n defined exactly as before, but with m_n now defined as the $d \times 1$ vector

$$m_n := \frac{1}{n}\sum_{i=1}^n g(x_i) \cdot \varepsilon_i^2 \ .$$

Again, we must argue that the difference between $\tilde{\theta}$ and $\hat{\theta}$ is of order $o_P\big(n^{-1/4}\big)$, and it suffices to show (B.17) where now

$$\hat{m}_n := \frac{1}{n}\sum_{i=1}^n g(x_i) \cdot \hat{\varepsilon}_i^2 \ .$$

But,

$$|\hat{m}_n - m_n| = \Big|\frac{1}{n}\sum_{i=1}^{n} g(x_i)\cdot(\hat{\varepsilon}_i^2 - \varepsilon_i^2)\Big| = \Big|\frac{1}{n}\sum_{i=1}^{n} g(x_i)\cdot\big[-2(\hat{\beta}-\beta)'x_i\cdot\varepsilon_i + (\hat{\beta}-\beta)'x_i\cdot x_i'(\hat{\beta}-\beta)\big]\Big|$$
$$\leq 2\Big|\frac{1}{n}\sum_{i=1}^{n} g(x_i)\cdot(\hat{\beta}-\beta)'x_i\cdot\varepsilon_i\Big| + \Big|\frac{1}{n}\sum_{i=1}^{n} g(x_i)\cdot(\hat{\beta}-\beta)'x_i\cdot x_i'(\hat{\beta}-\beta)\Big| \,.$$

Under moment assumptions, the sum in the first term is an average of mean-zero random vectors and is of order $O_P\big(n^{-1}\big)$ because $\hat{\beta}-\beta$ is of order $O_P\big(n^{-1/2}\big)$ and an average of zero-mean i.i.d. random variables with finite variance is also of order $O_P\big(n^{-1/2}\big)$. The second term does not have mean zero, but under moment assumptions, is of order $|\hat{\beta}-\beta|^2$, which is $O_P\big(n^{-1}\big)$. Therefore, $|\hat{m}_n - m_n|$ is actually of order $O_P\big(n^{-1/2}\big)$, which is clearly way more than needed.

C Figures and Tables

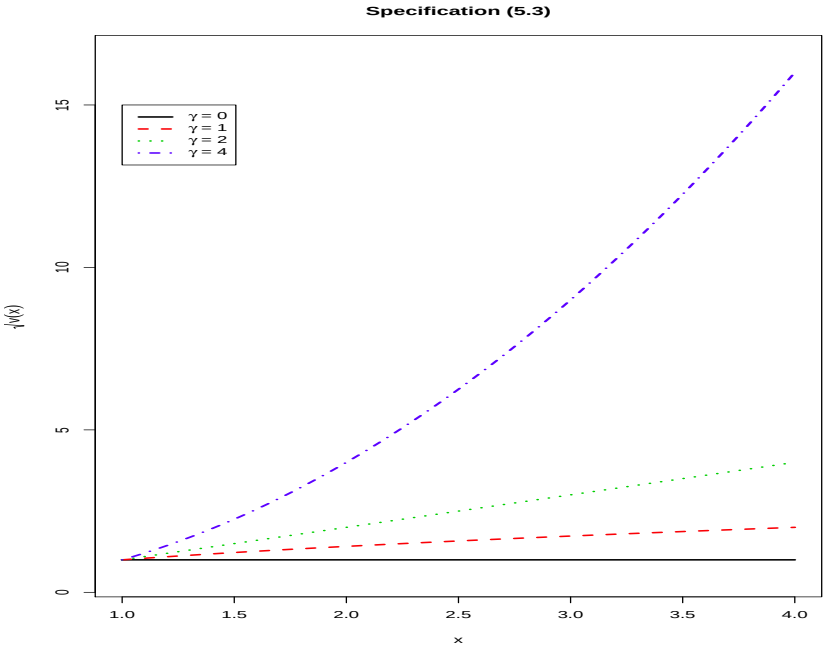

Figure 1: Graphical display of the parametric specification (5.3) for the skedastic function $v(\cdot)$. Note that for ease of interpretation, we actually plot $\sqrt{v(x)}$ as a function of x.

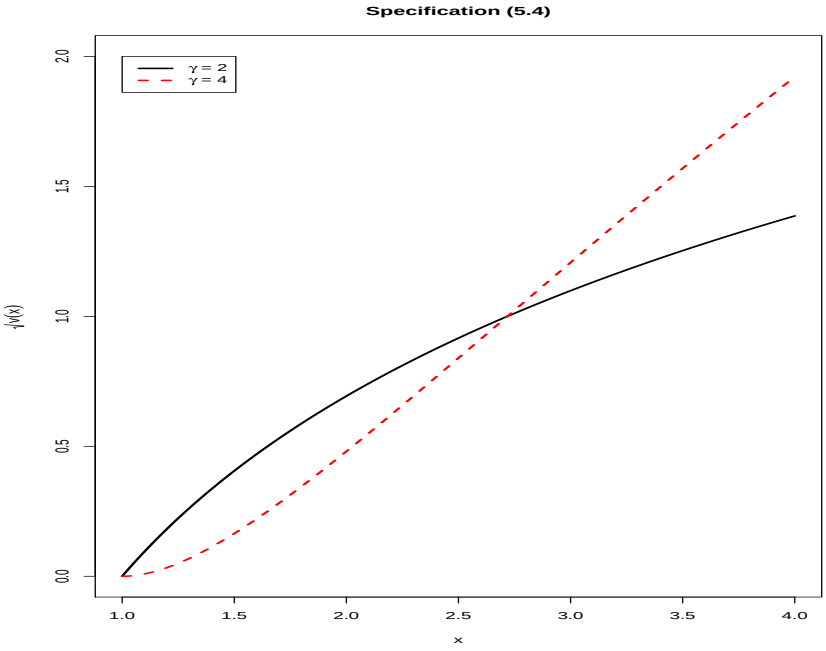

Figure 2: Graphical display of the parametric specification (5.4) for the skedastic function $v(\cdot)$. Note that for ease of interpretation, we actually plot $\sqrt{v(x)}$ as a function of x.

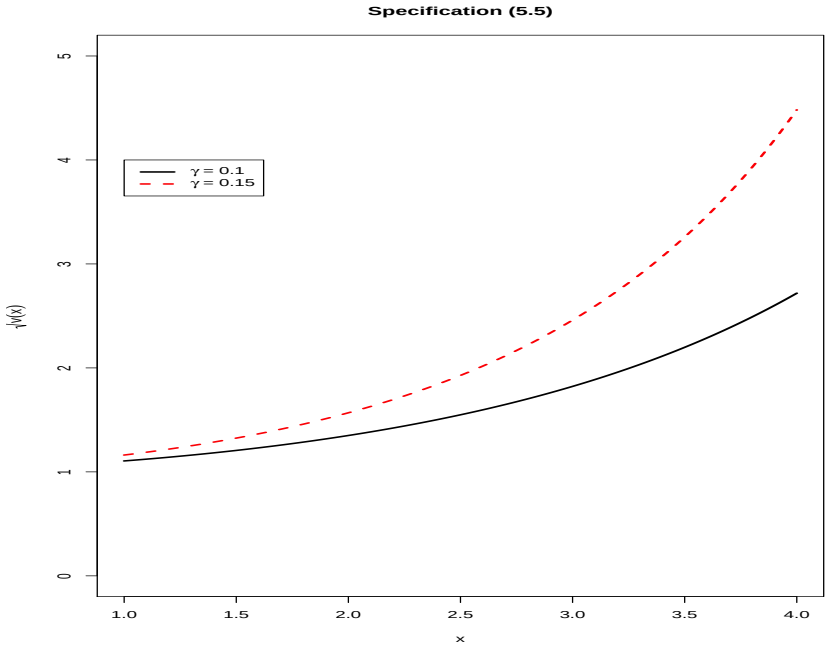

Figure 3: Graphical display of the parametric specification (5.5) for the skedastic function $v(\cdot)$. Note that for ease of interpretation, we actually plot $\sqrt{v(x)}$ as a function of x.

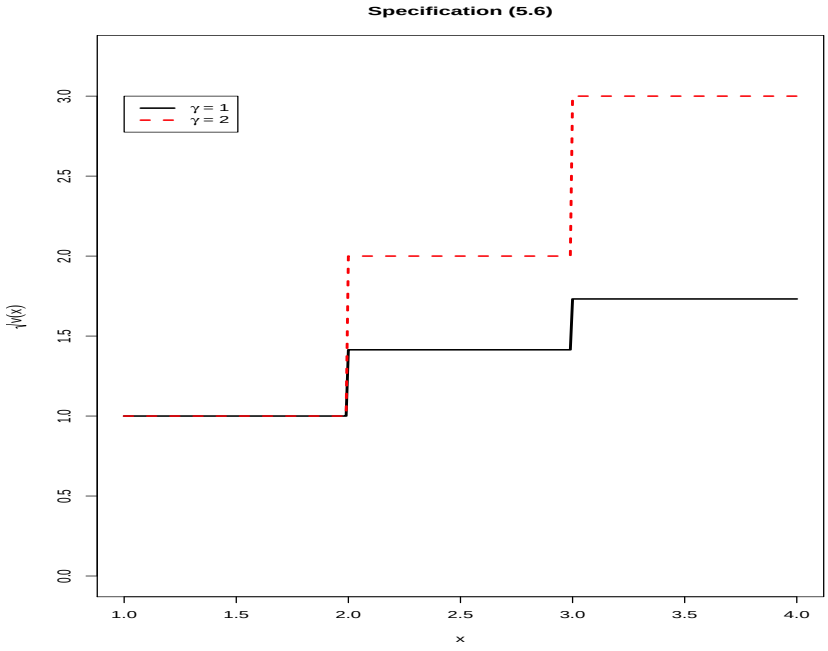

Figure 4: Graphical display of the parametric specification (5.6) for the skedastic function $v(\cdot)$. Note that for ease of interpretation, we actually plot $\sqrt{v(x)}$ as a function of x.

	OLS	WLS	ALS
$v(x) = x^\gamma$			
$\gamma = 0$			
$n = 20$	0.073	0.082 (1.12)	0.077 (1.04)
$n = 50$	0.028	0.029 (1.05)	0.028 (1.02)
$n = 100$	0.014	0.014 (1.02)	0.014 (1.01)
$\gamma = 1$			
$n = 20$	0.185	0.189 (1.03)	0.188 (1.02)
$n = 50$	0.070	0.066 (0.95)	0.069 (0.99)
$n = 100$	0.034	0.031 (0.92)	0.032 (0.95)
$\gamma = 2$			
$n = 20$	0.555	0.461 (0.83)	0.513 (0.93)
$n = 50$	0.211	0.157 (0.74)	0.171 (0.81)
$n = 100$	0.103	0.072 (0.70)	0.073 (0.71)
$\gamma = 4$			
$n = 20$	6.517	3.307 (0.51)	4.348 (0.67)
$n = 50$	2.534	0.957 (0.38)	0.994 (0.39)
$n = 100$	1.242	0.418 (0.34)	0.418 (0.34)
$\gamma = 0$, error terms ε_i of form (5.10)			
$n = 20$	0.074	0.082 (1.12)	0.077 (1.04)
$n = 50$	0.028	0.029 (1.06)	0.028 (1.02)
$n = 100$	0.014	0.014 (1.03)	0.014 (1.01)

Table 1: Empirical mean squared errors (eMSE's) of estimators of β_2 when the parametric model used to estimate the skedastic function $v(\cdot)$ is correctly specified. (In parentheses are the ratios of the eMSE of a given estimator over the eMSE of OLS.) All numbers are based on 50,000 Monte Carlo repetitions.

	OLS	WLS	ALS
$v(x) = \big[\log(x)\big]^{\gamma}$			
$\gamma = 2$			
$n = 20$	0.066	0.045 (0.69)	0.053 (0.81)
$n = 50$	0.025	0.014 (0.55)	0.015 (0.60)
$n = 100$	0.012	0.006 (0.50)	0.006 (0.50)
$\gamma = 4$			
$n = 20$	0.101	0.047 (0.46)	0.058 (0.58)
$n = 50$	0.039	0.013 (0.33)	0.013 (0.33)
$n = 100$	0.019	0.005 (0.25)	0.005 (0.25)
$v(x) = \exp(\gamma x + \gamma x^2)$			
$\gamma = 0.1$			
$n = 20$	0.250	0.236 (0.94)	0.246 (0.98)
$n = 50$	0.096	0.083 (0.87)	0.089 (0.93)
$n = 100$	0.047	0.039 (0.83)	0.041 (0.86)
$\gamma = 0.15$			
$n = 20$	0.530	0.413 (0.78)	0.473 (0.89)
$n = 50$	0.206	0.143 (0.70)	0.155 (0.75)
$n = 100$	0.101	0.067 (0.67)	0.068 (0.67)
$v(x)$ of form (5.6)			
$\gamma = 1$			
$n = 20$	0.148	0.151 (1.02)	0.150 (1.02)
$n = 50$	0.056	0.054 (0.96)	0.056 (1.00)
$n = 100$	0.027	0.025 (0.93)	0.026 (0.96)
$\gamma = 2$			
$n = 20$	0.365	0.303 (0.83)	0.337 (0.93)
$n = 50$	0.138	0.108 (0.77)	0.112 (0.81)
$n = 100$	0.067	0.051 (0.75)	0.051 (0.75)

Table 2: Empirical mean squared errors (eMSE's) of estimators of β_2 when the parametric model used to estimate the skedastic function $v(\cdot)$ is misspecified. (In parentheses are the ratios of the eMSE of a given estimator over the eMSE of OLS.) All numbers are based on 50,000 Monte Carlo repetitions.

	OLS-HC	OLS-Max	WLS-HC	WLS-Max	ASL-HC	ASL-Max
			$v(x) = x^\gamma$			
			$\gamma = 0$			
$n = 20$	95.4	96.5 (1.03)	93.5 (0.99)	95.0 (1.04)	94.5 (0.98)	95.8 (1.02)
$n = 50$	95.1	95.9 (1.03)	94.3 (0.99)	95.2 (1.02)	94.7 (1.00)	95.5 (1.02)
$n = 100$	95.1	95.7 (1.02)	94.8 (1.00)	95.4 (1.02)	95.0 (1.00)	95.5 (1.02)
			$\gamma = 1$			
$n = 20$	95.3	96.5 (1.04)	93.8 (0.94)	95.2 (0.98)	94.4 (0.96)	95.8 (1.01)
$n = 50$	95.1	95.9 (1.02)	94.5 (0.95)	95.4 (0.97)	94.5 (0.96)	95.3 (0.98)
$n = 100$	95.0	95.7 (1.02)	94.8 (0.95)	95.5 (0.97)	94.7 (0.97)	95.4 (0.99)
			$\gamma = 2$			
$n = 20$	94.8	96.0 (1.04)	94.0 (0.86)	95.2 (0.89)	93.9 (0.90)	95.1 (0.94)
$n = 50$	94.8	95.5 (1.02)	94.5 (0.84)	95.3 (0.86)	94.2 (0.85)	95.1 (0.88)
$n = 100$	94.8	95.3 (1.02)	94.8 (0.83)	95.4 (0.85)	94.8 (0.83)	95.3 (0.85)
			$\gamma = 4$			
$n = 20$	93.9	94.8 (1.02)	94.0 (0.66)	94.6 (0.67)	93.1 (0.70)	93.9 (0.71)
$n = 50$	94.4	94.7 (1.01)	94.3 (0.59)	94.7 (0.60)	94.2 (0.59)	94.6 (0.60)
$n = 100$	94.6	94.7 (1.00)	94.6 (0.57)	95.0 (0.58)	94.6 (0.57)	95.0 (0.58)

Table 3: Empirical coverage probabilities in percent of nominal 95% confidence intervals for β_2 when the parametric model used to estimate the skedastic function $v(\cdot)$ is correctly specified. (In parentheses are the ratios of the average length of a given confidence interval over the average length of OLS-HC.) All numbers are based on 50,000 Monte Carlo repetitions.

	OLS-HC	OLS-Max	WLS-HC	WLS-Max	ALS-HC	ALS-Max
$v(x) = \left[\log(x)\right]^{\gamma}$						
$\gamma = 2$						
$n = 20$	94.8	96.3 (1.05)	94.6 (0.77)	95.9 (0.81)	94.1 (0.80)	95.6 (0.85)
$n = 50$	94.6	95.7 (1.04)	94.6 (0.72)	96.2 (0.78)	94.5 (0.72)	96.0 (0.78)
$n = 100$	94.8	95.5 (1.03)	94.8 (0.70)	96.6 (0.77)	94.8 (0.70)	96.5 (0.77)
$\gamma = 4$						
$n = 20$	93.8	94.9 (1.02)	94.9 (0.61)	94.2 (0.62)	93.5 (0.63)	94.2 (0.64)
$n = 50$	94.3	94.7 (1.01)	94.3 (0.54)	94.6 (0.55)	94.2 (0.54)	94.6 (0.55)
$n = 100$	94.5	94.7 (1.00)	94.5 (0.52)	94.8 (0.53)	94.5 (0.52)	94.8 (0.53)
$v(x) = \exp(\gamma x + \gamma x^2)$						
$\gamma = 0.1$						
$n = 20$	94.9	95.8 (1.03)	93.3 (0.90)	94.5 (0.93)	93.7 (0.94)	94.8 (0.97)
$n = 50$	94.8	95.2 (1.01)	94.3 (0.91)	94.7 (0.92)	94.1 (0.93)	94.5 (0.94)
$n = 100$	94.9	95.0 (1.01)	94.7 (0.91)	94.9 (0.91)	94.6 (0.92)	94.8 (0.92)
$\gamma = 0.15$						
$n = 20$	94.5	95.2 (1.02)	93.3 (0.83)	94.2 (0.85)	93.2 (0.88)	94.0 (0.90)
$n = 50$	94.6	94.9 (1.01)	94.1 (0.82)	94.4 (0.81)	93.9 (0.83)	94.2 (0.84)
$n = 100$	94.7	95.8 (1.00)	94.6 (0.81)	94.7 (0.81)	94.6 (0.81)	94.7 (0.81)
$v(x)$ of form (5.6)						
$\gamma = 1$						
$n = 20$	95.2	96.4 (1.04)	93.7 (0.94)	95.1 (0.98)	94.2 (0.96)	95.6 (1.00)
$n = 50$	95.0	95.8 (1.03)	94.4 (0.95)	95.2 (0.98)	95.2 (0.97)	96.1 (1.00)
$n = 100$	95.0	95.7 (1.02)	94.7 (0.96)	95.2 (0.97)	94.6 (0.96)	95.2 (0.98)
$\gamma = 2$						
$n = 20$	94.7	96.0 (1.04)	93.9 (0.86)	94.9 (0.89)	93.5 (0.89)	94.7 (0.93)
$n = 50$	94.8	95.5 (1.03)	94.4 (0.86)	94.9 (0.87)	94.1 (0.87)	94.7 (0.88)
$n = 100$	94.8	95.3 (1.02)	94.8 (0.86)	95.0 (0.87)	94.8 (0.86)	95.0 (0.87)

Table 4: Empirical coverage probabilities in percent of nominal 95% confidence intervals for β_2 when the parametric model used to estimate the skedastic function $v(\cdot)$ is misspecified. (In parentheses are the ratios of the average length of a given confidence interval over the average length of OLS-HC.) All numbers are based on 50,000 Monte Carlo repetitions.