Feasible Generalized Least Squares Using Machine Learning

Steve Miller*
Department of Applied Economics, University of Minnesota
Richard Startz
Department of Economics, University of California, Santa Barbara

February 6, 2018

Abstract

In the presence of heteroskedastic errors, regression using Feasible Generalized Least Squares (FGLS) offers potential efficiency gains over Ordinary Least Squares (OLS). However, FGLS adoption remains limited, in part because the form of heteroskedasticity may be misspecified. We investigate machine learning methods to address this concern, focusing on Support Vector Regression. Monte Carlo results indicate the resulting estimator and an accompanying standard error correction offer substantially improved precision, nominal coverage rates, and shorter confidence intervals than OLS with heteroskedasticity-consistent (HC3) standard errors. Reductions in root mean squared error are over 90% of those achievable when the form of heteroskedasticity is known.

Keywords: Keywords: Heteroskedasticity, Support Vector Regression, Weighted Regression

*We thank Marc Bellemare, Clément de Chaisemartin, and Doug Steigerwald for helpful comments. We are also grateful for computational resources provided by the Minnesota Supercomputing Institute.

1 Introduction

Heteroskedastic errors render Ordinary Least Squares (OLS) estimators inefficient and induce bias in the corresponding standard errors. Two broad methods exist for dealing with heteroskedasticity: use another estimator that exploits heteroskedasticity in pursuit of efficiency, or construct estimates using OLS but adjust the standard errors. The most common forms of alternate estimators are Weighted Least Squares (WLS) if the form of heteroskedasticity is known and Feasible Generalized Least Squares (FGLS) if the form of heteroskedasticity must be estimated. If a researcher chooses to use OLS coefficient estimates, the most popular corrections to standard errors are the heteroskedasticity-consistent variants (Eicker 1963, Huber 1967, White 1980, MacKinnon & White 1985, Cribari-Neto 2004).

More recent empirical work tends to favor simply correcting OLS standard errors, largely because doing so imposes less stringent assumptions. The first set of assumptions concern identification. The OLS estimator is consistent under slightly weaker exogeneity assumptions than the FGLS estimator.[1] In addition, in finite samples, OLS is unbiased under strict exogeneity, while FGLS is biased unless further assumptions are made. While these identification concerns are clearly important, they are not the focus of the present paper.

This paper is focused on a second set of assumptions limiting the adoption of WLS and FGLS; specifically, those concerning the form of heteroskedasticity. WLS requires a rarely satisfied assumption that the exact form of heteroskedasticity is known. Similarly, estimation of the form of heteroskedasticity in FGLS requires either assumptions about parametric functional form or selection of a nonparametric approach and the associated details (e.g. kernel choice). If the functional form used for parametric estimation or bandwidth choices for nonparametric estimation lead to a poor approximation of the true skedastic function, the estimated weights from this first stage may cause FGLS to be *less* efficient than OLS. Further, if the transformed model produced by FGLS does not eliminate heteroskedasticity, standard errors and the resulting inference (if using standard errors estimated under the assumption of conditional homoskedasticity) will be incorrect. While the latter problem

[1] Specifically, the OLS estimator is consistent under the assumption $E[ux] = 0$ while the FGLS estimator may not be.

can in principle be dealt with by estimating heteroskedasticity-consistent standard errors even when estimating parameters via FGLS (Wooldridge 2010, Romano & Wolf 2017), the potential for misspecification of the form of heteroskedasticity has helped limit the use of FGLS in applied research.

In this paper, we consider whether support vector regression (SVR), an approach from the field of machine learning, may help address these concerns and make FGLS a more viable tool. SVR and many other machine learning tools couple nonparametric estimation with controls on overfitting, yielding excellent out-of-sample predictive performance in finite samples. These properties are likely to prove useful in modeling the form of heteroskedasticity. Since, for the purposes of FGLS, any causal relationship between regressors and the error variance is not of direct interest, the purely predictive nature of these algorithms is suitable and their performance attractive.

Since SVR can be considered a nonparametric estimator, it is worth providing some intuition as to why we might expect FGLS using SVR to outperform nonparametric approaches proposed before, such as nearest-neighbor (Carroll 1982), kernel (Robinson 1987), and series (Newey 1994) estimation of the skedastic function.[2] The advantages of SVR in this context come from its three defining characteristics: the use of a kernel, a penalty on complexity, and an ϵ-insensitive loss function rather than squared error loss. The use of kernels affords SVR the same flexibility in function approximation as many nonparametric methods, including the kernel regression approach in (Robinson 1987). The flexibility in all of these methods comes at the cost of potential overfitting, and high variance in the estimated weights translates to variance in the final FGLS estimates. The first guard against overfitting in SVR is squared ℓ_2 norm penalization of the coefficient vector, giving SVR the same advantages of kernel ridge regression. What makes SVR unique is its ϵ-insensitive loss function, which ignores errors smaller than ϵ and penalizes larger errors linearly beyond that threshold. This linearity renders SVR less susceptible than kernel ridge regression to outliers, much as in the absolute value loss function used in least absolute deviation (LAD) estimation. Altogether, these features suggest that SVR strikes a balance between flexibility and complexity, together with less susceptibility to outliers. While all of these

[2] For an example application of a kernel estimator see, e.g., O'Hara & Parmeter (2013).

properties offer promise for estimating the skedastic function using SVR in the context of FGLS, to our knowledge this approach has yet to be examined.

Our investigations offer two contributions. First, we present Monte Carlo evidence suggesting that the use of SVR to estimate the skedastic function (the relationship between the error variance and regressors) can offer substantial gains in precision without requiring strong functional form assumptions. As with other FGLS methods, these gains predictably come at the cost of some precision under homoskedasticity. However, when heteroskedasticity is present, the estimator provides dramatic reductions in root mean squared error (RMSE) compared to OLS with heteroskedasticity-consistent standard errors. On simulated datasets we construct, the reductions in RMSE are 87-98% of those achieved if the form of heteroskedasticity is known and an appropriate parametric estimation approach is used. Moreover, these benefits accrue not only for purely simulated data, but also for more realistic data generated via wild bootstrap. For comparison with prior work investigating parametric FGLS (Romano & Wolf 2017), we apply our estimator to the well-known Boston housing dataset (Harrison & Rubinfeld 1978, Drucker et al. 1997, Wooldridge 2015). We find that FGLS using support vector regression offers lower RMSE than either parametric approach and similar RMSE to nonparametric alternatives but better coverage properties.

Our second contribution is a finite sample correction to heteroskedasticity-consistent (HC) standard errors estimated after conducting FGLS. Flexible estimation of the skedastic function via support vector regression or nonparametric approaches results in weights that can themselves be highly variable. In finite samples, failing to account for that extra source of variability can result in incorrect test sizes (Rothenberg 1988),[3] and we find this to be the case even for commonly-used HC estimators. In response, we propose a correction to standard errors that brings coverage rates near nominal levels for FGLS. The correction accounts for the higher degrees of freedom inherent in more flexible modeling of the skedastic function. This correction may be of independent interest for researchers modeling the skedastic function in other ways, as it offers improved coverage properties for existing nonparametric methods.

Our investigation of this approach is organized as follows. Section 2 sets out the problem

[3] A similar concern was addressed in the context of generalized method of moments estimation by Windmeijer (2005).

of estimation under heteroskedasticity and the general FGLS approach, then motivates the use of support vector regression and an associated correction to standard errors. Section 3 details options for estimating the skedastic function, with a focus on our proposal to use support vector regression for that purpose. Section 4 describes the Monte Carlo studies we use to investigate the performance of the proposed estimator. Section 5 presents results, and the final section concludes.

2 Problem setup and motivation

We are interested in estimation of and inference about parameters in linear models where error terms may be conditionally heteroskedastic. We restrict our attention to linear data generating processes of the form

$$Y = X\beta + u, \tag{1}$$

where Y is an n by 1 vector of outcomes of interest, X is an n by k matrix of predictors, β is a k by 1 parameter vector, and u is an n by 1 vector of unobserved error terms. We make several standard, simplifying assumptions to focus our attention on heteroskedasticity. In particular, we assume regressors are exogenous ($E[u|X] = 0$), the observed data represents a random sample, and there is variability in but no perfect collinearity among regressors, such that $X'X$ is invertible. We assume throughout that errors are uncorrelated across observations and u exhibits heteroskedasticity of unknown form, where $v(X) = Var(u|X)$ is non-constant and we may not have theoretical or intuitive guidance as to the process giving rise to $Var(u|X)$.

A well-known approach to estimating (1) is Feasible Generalized Least Squares (FGLS), which, as its name indicates, is a feasible form of Generalized Least Squares (GLS). FGLS consists of three primary steps:

1. Estimate (1) via OLS. Denote the residuals for observation i by $\hat{u}_i$.
2. Estimate a model $\hat{u}_i{}^2 = g(z_i)$ explaining the squared residuals, where z_i may include some or all elements of x_i or transformations of them. Denote the predicted values from this model by $\tilde{u}_i^2$.

3. Estimate (1) again, this time minimizing a weighted sum of squared residuals, with weight for squared residual i given by $\frac{1}{\tilde{u}_i^2}$.

The potential benefit of this procedure is that the estimator remains consistent but is asymptotically more efficient than OLS because it attaches less weight to observations considered to be noisier. The resulting estimator can be written

$$\hat{\beta}_{FGLS} = (X'\hat{W}^{-1}X)^{-1}X'\hat{W}^{-1}Y \tag{2}$$

where $\hat{W}$ is a diagonal matrix with entries $\hat{w}_{ii} = \tilde{u}_i^2$ from step 3 in the procedure above. Denote the residuals from this estimator by $\hat{u}_{i,FGLS}$.

Several different approaches have been proposed for estimation of the skedastic function in the second step. The simplest approach is to estimate a linear model explaining ${\hat{u}_i}^2$. To ensure positive predictions, often $\log({\hat{u}_i}^2)$ is treated as the response variable and predictions are exponentiated. Suggestions differ in terms of what form the right hand side should take in such models: Wooldridge (2010) suggests using regressors as they appear in the main model, while Romano & Wolf (2017) suggest the log of the absolute value of regressors. Other transformations or specifications are of course possible while restricting attention to linear models. Past proposals include nonparametric approaches to estimating $v(X)$, including kernel regression (Carroll 1982), nearest neighbors (Robinson 1987), and series estimation (Newey 1994).

Here we propose to use support vector regression as another potential approach for estimating the form of heteroskedasticity. We motivate our proposal with two observations. First, the potential efficiency gains of FGLS are largest if the method used for skedastic function estimation can consistently estimate $v(X)$. Support vector regression, through its use of kernels, is flexible enough to consistently estimate a wide class of functions without functional form restrictions in parametric estimation. The cost of the flexibility in this and other nonparametric methods is in increased variance, which translates to slower convergence rates.

To illustrate the effect of variance in skedastic function estimation, as well as what advantages support vector regression brings, we use an approximation to $\hat{\beta}_{FGLS}$. Provided that the chosen method is sufficiently flexible so that the estimated variances $\tilde{u}_i^2$ are sufficiently close to the true variances $v(x_i)$, then $\hat{\beta}_{FGLS}$ may be approximated via a Taylor expansion around $\hat{\beta}_{GLS}$. As in Rothenberg (1988), we can use this Taylor expansion to approximate the additional variance in a particular element of $\hat{\beta}_{FGLS}$ that is introduced by the estimation of weights.[4] Denoting the cth element of the FGLS and GLS estimators as $\hat{\beta}_{FGLS,c}$ and $\hat{\beta}_{GLS,c}$, respectively, we have:

$$
\begin{aligned}
&Var(\sqrt{n}(\hat{\beta}_{FGLS,c} - \hat{\beta}_{GLS,c})) \approx \\
&\qquad \sum_i \sum_j C'AXW^{-1}[W^{-1}(I_n - H_{GLS})]_{ij} cov(\tilde{u}_i^2, \tilde{u}_j^2) W^{-1}XAC. \qquad (3)
\end{aligned}
$$

Here $A = (X'W^{-1}X)^{-1}$, and C is a vector containing a one in position c and zeros elsewhere, used to select the coefficient of interest. The matrices I_n and H_{GLS} are an identity matrix and the GLS hat matrix, respectively, while the subscript ij denotes the ijth element of the corresponding matrix. The ijth element of the term in square brackets can be interpreted as the influence of the jth outcome on the weighted GLS residual for observation i. Note the entire approximation takes a familiar sandwich form.

The approximation (3) hints at the potential advantages of support vector regression, which accrue through controls on variance. In particular (3) indicates that when $i = j$, the variance of the method used to estimate $\tilde{u}_i^2$ contributes to the variance of $\hat{\beta}_{FGLS,c}$ in proportion to i's influence on its own weighted GLS residual. Support vector regression controls guards against variance of $\tilde{u}_i^2$ in two key ways. First, SVR penalizes model complexity in a manner akin to (kernel) ridge regression, thereby reducing the variance of its predictions. Second, SVR minimizes a loss function that is insensitive to errors below a threshold and only increases linearly beyond that point, making it more robust to outliers. Based on (3), these features should reduce $Var(\tilde{u}_i^2)$ and the resulting contribution to $Var(\hat{\beta}_{FGLS,c})$. These variance controls should be especially important for observations with high leverage or low true structural variance. Because high leverage observations are those with atypical and/or extreme covariate vectors, those are also precisely the observations where nonparametric approaches without such controls may be more variable.

The approximation (3) also reminds us that inference using FGLS estimates should

[4] In his discussion of FGLS estimators, Rothenberg (1988) assumes a low-dimensional parametric form and studies deviations in those parameters. Our approximation uses deviations in each observation's structural variance. Estimation of n parameters is of course not our goal: we use this only as a device to illustrate tradeoffs in skedastic function estimation choices.

account for the difference between $\hat{\beta}_{FGLS,c}$ and $\hat{\beta}_{GLS,c}$. Recent work by Romano & Wolf (2017) partly addresses this issue, illustrating that heteroskedasticity-robust standard errors will provide asymptotically valid inference even if the estimated weights used in step 3 do not correspond to the inverse of the conditional variance $v(x_i)$. In particular, Romano & Wolf (2017) suggest estimators of the form

$$\widehat{Var}(\hat{\beta}_{FGLS}|X) = (X'\hat{W}^{-1}X)^{-1}\hat{\Omega}(X'\hat{W}^{-1}X)^{-1} \tag{4}$$

where $\hat{\Omega}$ is an estimator of $E\left[\frac{u_i^2}{(\hat{v}_{lim}(x_i))^2}x_i x_i'\right]$, where $\hat{v}_{lim}(x_i)$ is the probability limit of the estimator of $v(x_i)$. Their proposal takes the form of well-known sandwich estimators used for robust inference after OLS. Most common heteroskedasticity-robust standard errors use a form of $\hat{\Omega}$ which can be written $\hat{\Omega} = X'\hat{\Sigma}X$ where $\hat{\Sigma}$ is a diagonal matrix. For example, we consider the HC3 variant suggested by MacKinnon & White (1985), which, adapted to the weighted least squares estimator, has ith diagonal entry

$$\hat{\Sigma}_{ii} = \frac{\hat{u}_{i,FGLS}^2}{(\tilde{u}_i^2)^2(1 - h_{i,FGLS})^2}$$

where

$$h_{i,FGLS} = [X(X'\hat{W}^{-1}X)^{-1}X'\hat{W}^{-1}]_{ii}.$$

Note that the squared residuals ($\hat{u}_i^2$) and hat matrix values $h_{i,OLS} \equiv [X(X'X)^{-1}X']_{ii}$ from the first stage OLS regression could also be used, but efficiency gains in FGLS suggest that using the second stage residuals and hat matrix values should provide a lower variance estimate of $E[u_i^2]$.

While the approach proposed by Romano & Wolf (2017) offers consistent estimation of standard errors, inference in finite samples may still suffer. Specifically, the additional variability of $\hat{\beta}_{FGLS}$ as approximated in (3) may be non-negligible in finite samples, and estimated standard errors which ignore that variability may be too small and result in over-rejection of null hypotheses. Rothenberg (1988) noted this issue for parametric FGLS and provided a correction to t-statistics based on Edgeworth expansions under correct parametric specification of the skedastic function. We seek an alternate standard error correction to apply when using support vector regression or other flexible nonparametric approaches to estimation of weights.

To this end, we propose a standard error correction that blends the proposal from Romano & Wolf (2017) with insights from the approximation (3). Our correction requires only an estimate of the degrees of freedom used in skedastic function estimation, together with a conventional OLS hat matrix and the number of covariates. Aside from its simplicity, the key advantage of our correction is that estimates of degrees of freedom are available for nonparametric estimators and some popular machine learning algorithms, most notably SVR, while analytical covariance estimates are not always available.

Specifically, we propose the following revised estimator $\hat{\Sigma}_{ii}^{FGLS}$ as a part of $\widehat{Var}(\hat{\beta}_{FGLS}|X)$ (see Appendix for details):

$$\hat{\Sigma}_{ii}^{FGLS} = \frac{\hat{u}_{i,FGLS}^2}{(\tilde{u}_i^2)^2} \left(\frac{1}{(1 - h_{i,FGLS})^2} + 4 \frac{h_{i,ols}}{k} \hat{df} \right). \tag{5}$$

Here, $\hat{df}$ is an estimate of the degrees of freedom used by the second stage weight estimation, and $h_{i,OLS}$ are hat values from the first stage OLS estimation. We refer to the resulting estimator $\widehat{Var}(\hat{\beta}_{FGLS}|X)$ which uses $\hat{\Sigma}_{ii}^{FGLS}$ as $HCFGLS$. The first term in parentheses is the standard multiplier used in HC3 standard errors. The second term in parentheses is the new correction we introduce, which addresses the use of estimated weights in FGLS and has intuitive interpretation. First, the correction is larger for observations with higher relative leverage (those with larger values of $\frac{h_{i,ols}}{k}$), since those points have greater potential to influence the OLS regression line, resulting in more variable squared residuals feeding into estimation of the skedastic function. Similarly, the correction is larger for more flexible skedastic models with higher $\hat{df}$, since more flexible models may result in higher variance estimates of the weights. Note that in the special case of $\hat{df} = 0$, which corresponds to using OLS without any reweighting, our correction term reduces to zero and HCFGLS is equivalent to HC3. Further, the correction does not affect consistency of the variance covariance matrix estimates, since the second term in parentheses converges in probability to zero as the sample grows in size (and $h_{i,ols}$ approaches zero).

With this broadly applicable standard error correction in place, we turn next to the specific methods we investigate for estimation of the skedastic function.

3 Estimating the skedastic function

Before describing our simulations, we briefly explain the methods we use to estimate the skedastic function $g(x_i)$ as part of FGLS. We begin with a short explanation of previously proposed parametric and nonparametric estimators before focusing our discussion on support vector regression. For all of the methods we consider, we first estimate a model of the log squared residuals and then exponentiate the predictions to arrive at estimated weights. We do this not only for the guarantee of positive weights described earlier, but also for potential reductions in variance. As seen in (3), $Var(\tilde{u}_i^2)$ has larger (approximate) influence on the variance of the FGLS estimator when $v(x_i)$ is small (through W^{-1}). By estimating the skedastic function in the log space, prediction errors for these low true variance points thus contribute more to whatever loss function is used in the relevant method.

3.1 Previously proposed parametric estimators

As a baseline, we estimate the conditional variance function via OLS using four specifications. First, we estimate a model using the correct – but generally unknown in practice – functional form of the conditional variance function. That is, if $log(u^2) = Z\alpha$, where entries in Z are some transformations of X, we estimate $log(\hat{u}^2) = Z\alpha + e$ via OLS. We refer to this specification as parametric FGLS with the correct functional form. The second model uses the same regressors as the main model. That is, we estimate the model $log(\hat{u}^2) = X\alpha + e$ via OLS, which ignores any differences between Z and X. We refer to this specification as parametric FGLS with the main model functional form. Finally, for comparison with prior work, we estimate the proposed WLS-S1 and WLS-S2 specifications from Romano & Wolf (2017). WLS-1 entails a regression:

$$log(\max(\hat{u}^2, \delta^2)) = log(|X|)\alpha + e,$$

where the constant δ bounds squared residuals away from zero to limit the possibility of extreme weights. WLS-S2 is identical to our main model specification except for the use of δ to bound weights:

$$log(\max(\hat{u}^2, \delta^2)) = X\alpha + e.$$

For all of these linear models of the skedastic function, our proposed standard error correction uses $\hat{df}$ equal to the number of parameters.

3.2 Previously proposed nonparametric estimators

Prior researchers have suggested the use of nonparametric estimation of the form of heteroskedasticity as part of FGLS. We implement two such estimators to compare the performance of support vector regression to more established nonparametric approaches. First, Carroll (1982) suggested an FGLS estimator using kernel regression to estimate the skedastic function. Under relatively mild symmetry and smoothness assumptions on the kernel used, Carroll (1982) established consistency in a simple linear regression case and used Monte-Carlo evidence to evaluate performance. We evaluate his suggestion with a Gaussian kernel. Second, we also employ the k nearest neighbor approach suggested by Robinson (1987), which estimates the outcome for a focal observation using the average of observed outcomes for the k observations that are in a specified sense closest to the focal observation. Distance between observations can be specified in a number of ways, e.g. Mahalanobis distance based on covariate values.

The finite sample correction to standard errors we proposed earlier is easily computed for both estimators. For kernel regression, $\hat{df}$ is equal to the trace of the smoother (weight) matrix, which captures the sum of the influence of individual observations on their own fitted values (Friedman et al. 2001). For k nearest neighbor estimation, $\hat{df}$ is simply n/k: the number of observations divided by the number of neighbors (Friedman et al. 2001). The intuition for $\hat{df}$ in this case is clear if observations fall into n/k distinct neighborhoods of k neighbors each, in which case a single degree of freedom (estimate) would be needed for each of the n/k neighborhoods.

3.3 Support Vector Regression

Support vector regression, like OLS, seeks to minimize a function of residuals, but it penalizes residuals in a different way, penalizes coefficient vectors that are large in magnitude, and allows for nonlinearities through the use of kernels (Vapnik 1995, Drucker et al. 1997). In the form of SVR we consider, the loss function for residuals is the ϵ-insensitive loss function, which imposes no penalty on residuals smaller than $|\epsilon|$ and penalizes residuals linearly in the degree to which they exceed $|\epsilon|$. Coefficient size is penalized using the squared ℓ_2 norm, which reduces the variance of the model and hence overfitting.

In the our use of SVR to model the log squared residuals from OLS, the method can be seen to solve the following problem:

$$\min_{\gamma=[\gamma_1,\ldots,\gamma_m,\ldots\gamma_M]} \sum_{i=1}^{n} L_\epsilon(log(\hat{u}_i^2) - g(x_i,\gamma)) + \lambda \sum_{m=1}^{M} \gamma_m^2 \tag{6}$$

where

$$g(x_i,\gamma) = \sum_{m=1}^{M} \gamma_m h_m(x_i) = \gamma^T h(x_i)$$

$$L_\epsilon(r) = \begin{cases} 0 & \text{if } |r| < \epsilon, \\ |r| - \epsilon & \text{otherwise.} \end{cases}$$

The formulation (6) contains the ϵ-insensitive loss function $L_\epsilon(\cdot)$ and the squared ℓ_2 norm penalization of the coefficient vector γ, both of which were described above. The two goals of minimizing loss and coefficient size are traded off according to the tuning parameter λ. The flexibility of the method comes through the use of basis functions $h_m(\cdot)$, which map the input data into M dimensions (the function $h(\cdot)$ simply collects these basis functions into a vector). In particular, by letting M grow large ($M >> k$), and through appropriate choices of $h_m(\cdot)$, the method gains many of the advantages of series estimators. As with all nonparametric methods, the predictive performance of SVR will be influenced by the values of its tuning parameters, ϵ and λ. Preferred approaches to tuning parameter selection use either theoretical guidance or cross-validation; we select tuning parameters via cross-validation for the simulations we describe later.

The solution to (6) makes use of a dual formulation that connects SVR to kernel regression. In particular, the solution depends only on the inner product of $h(x_i)$ and $h(x_j)$. For some choices of the individual basis functions, this inner product is equivalent to a kernel function $K(x_i, x_j)$ applied to the original data vectors x_i and x_j. As such, SVR can also be seen to share properties of kernel estimators. Importantly, however, the ϵ-insensitive loss function L_ϵ results in different weighting of these kernel terms in the resulting prediction. In particular, only observations that are at least ϵ away from the prediction surface contribute to a new prediction: the solution is sparse. Among the observations in that sparse solution, linear penalization of residuals beyond ϵ caps the contribution an individual observation can contribute to the prediction, limiting prediction variance due to outliers.

In this study, we use the commonly-applied Gaussian or radial basis function (RBF) kernel:

$$K(x_i, x_j) = e^{\phi||x_i - x_j||^2}. \tag{7}$$

The RBF kernel corresponds to an infinite series of basis functions $h_m(\cdot)$, making SVR with the RBF kernel it a remarkably flexible approximation tool. See Steinwart & Christmann (2008) for a formalization of the universal consistency of SVR with the RBF kernel. Note that another tuning parameter ϕ, has been introduced in the kernel function, which influences the width of the kernel. As with ϵ and λ, we choose ϕ using cross-validation in our simulations.

Before proceeding, it is worth noting when our choice of kernel is likely to be appropriate, and when it is not. The RBF kernel can, in principle, approximate any continuous, bounded function with arbitrary precision (Micchelli et al. 2006). As a result, if the skedastic function is believed to be both continuous and bounded, the RBF kernel is an attractive choice. Still, its flexibility is best leveraged with large amounts of data; with smaller numbers of observations a less flexible kernel (e.g. linear or polynomial) may be more appropriate. Similarly, if the skedastic function is known to be discontinuous, an estimation technique producing discontinous approximations, such as a regression tree, may be more appropriate.

Finally, in light of our proposed correction to standard errors based on (5), we also require an estimate of the degrees of freedom used in estimation of the skedastic function. An unbiased estimator for the degrees of freedom used in SVR is (Gunter & Zhu 2007):

$$\hat{df} = |\varepsilon_+| + |\varepsilon_-|, \tag{8}$$

where ε_+ and ε_- are the sets of observations with residuals equal to ϵ and $-\epsilon$, and $|\cdot|$ denotes the number of elements in a set. Note that due to the (implicit) mapping of the data into a higher dimensional space via basis functions, $\hat{df}$ may be much larger than k for SVR. This is precisely when the proposed standard error correction is likely to matter.[5]

[5] It is possible to empirically estimate degrees of freedom for the other methods using resampling and

4 Monte Carlo simulations

4.1 Simple data generating process

To examine the potential for estimation of the error variance support vector regression, we evaluate the performance of several data generating processes and conduct Monte Carlo simulations using various estimators. Because all the estimators we consider are consistent, we focus evaluation on precision and inference. In particular, for each FGLS variant we compute three key statistics to evaluate performance: empirical root mean squared error (RMSE), 95% confidence interval coverage probability (Cov. Prob.), and ratio of confidence interval length to OLS confidence interval length (CI ratio). RMSE is the square root of the average squared difference between the point estimate and true parameter value across the B Monte Carlo runs:

$$RMSE = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\hat{\beta}_{FGLS,b} - \beta)^2}. \tag{9}$$

The coverage probability is the fraction of Monte Carlo runs for which the estimated confidence interval included the true parameter value. The CI ratio is the confidence interval length for the given estimator divided by that for the OLS estimator.

We compare eight estimators: (1) a baseline of OLS, as well as (2-8) FGLS with the relationship between $Var(u|x)$ and x estimated seven different ways. The FGLS variants estimate the skedastic function in the following ways: (2) OLS on the correct functional form relating $Var(u)$ and x, (3) OLS assuming $Var(u|x)$ has the same functional form as $y(x)$ (the main model), (4) WLS-S1 from Romano & Wolf (2017) (5) WLS-S2 from Romano & Wolf (2017) (6) kernel estimation, (7) k nearest neighbor estimation, and (8) support vector regression estimation (SVR).

For each Monte Carlo run, we generate a random sample of covariates x, draw a sample of error variables u from a distribution with variance that may depend on x, compute the resulting y variables according to a known data generating process, and estimate all models using that dataset. In all cases, the structural model is correctly specified, so

perturbation approaches (see, e.g. Ye (1998)). At that point, however, the advantage of our proposed correction is unclear versus simply bootstrapping standard errors or confidence intervals directly.

that the focus is solely on the effects of heteroskedasticity. We repeat the Monte Carlo exercise 50,000 times for each form of heteroskedasticity. In all cases, we construct confidence intervals using heteroskedasticity-robust standard errors and a critical value, based on a normal approximation, of 1.96. In particular, we employ both the HC3 variant of heteroskedasticity-consistent standard errors suggested by MacKinnon & White (1985), and our HCFGLS estimate employing the correction in (5).

We repeat this process for three data generating processes, which differ only in the error variance. In all cases, the data generating process takes the following form:

$$\begin{aligned} y =& \beta_0 + \beta_1 x_1 + u, \\ u =& \sqrt{h(x_1)}\theta, \\ x_1 \sim& \mathcal{N}(0, \sigma_x^2), \\ \theta \sim& \mathcal{N}(0, \sigma^2). \end{aligned} \tag{10}$$

Note that for this model $Var(u|x_1) = \sigma^2 h(x_1)$. In the different scenarios, we simply choose different forms for $h(x_1)$, which are summarized in Table 1.

Scenario	True $Var(u\|x_1)$
Homoskedasticity	σ^2
Moderate heteroskedasticity	$\sigma^2(\alpha_0 + \alpha_1 x_1^2)$
Severe heteroskedasticity	$\sigma^2 exp(\alpha_0 + \alpha_1 x_1)$

Table 1: Scenarios: true model for $Var(u|x)$, along with implicitly assumed form if, in the second step of FGLS, a researcher simply estimates a model of the log of squared residuals as a function of the main (structural) model regressors.

In all scenarios, OLS and all FGLS variants remain consistent. For the two heteroskedasticity scenarios, we expect OLS to be inefficient and for FGLS with correctly specified error variance function to offer efficiency gains over OLS. The focus of the exercise is on the remaining FGLS variants, since misspecification or poor variance estimation could result in weights that actually reduce the efficiency of the estimator.

All simulations are performed using the R programming language. The `npreg`, `kknn`, and `e1071` packages are used for kernel, nearest neighbor, and SVR implementations, respectively (Racine & Li 2004, Li & Racine 2004). Kernel regression SVR both use Gaussian kernels, and all tuning parameters for nonparametric methods and support vector regression (e.g. bandwidths, number of neighbors, ϵ, λ, ϕ) are chosen via cross-validation.[6]

4.2 Boston housing data

To complement our simulated data investigations, we also evaluate the performance of FGLS using SVR on a more realistic dataset, turning to the commonly-used Boston housing dataset from Wooldridge (2015) to do so. In particular, we model the log of community median housing prices in Boston in 1970 as a function of several characteristics:

$$log(price) = \beta_0 + \beta_1 log(nox) + \beta_2 log(dist) + \beta_3 rooms + \beta_4 stratio + u. \tag{11}$$

Here *nox* is nitrogen oxide concentration (parts per million), *dist* is weighted distance (miles) from employment centers, *rooms* is the mean rooms per house, and *stratio* is the mean student-to-teacher ratio.

We choose this dataset both because it is well known and for comparison with recent work by Romano & Wolf (2017). We employ the same wild bootstrap approach as Romano & Wolf (2017), generating 50,000 replicates and evaluating RMSE and size of candidate estimators. Note that in the wild bootstrap approach, the "true" parameter values are those generated via OLS estimation on the main dataset. Thus RMSE is calculated based on deviation of the FGLS estimates from the OLS estimates in the original data. For this exercise, we limit our attention to the SVR-based FGLS estimator, a kernel estimator, and the WLS-S1 and WLS-S2 estimators in order to emphasize comparison of the SVR estimator to proposals in Romano & Wolf (2017).

[6] Optimization of cross-validation error over the tuning parameter space can be performed via grid search or any optimization routine. For example, the `npreg` package uses Powell's method to search over tuning parameters.

5 Results

5.1 Simple data generating process

The primary results from our simulations are presented in Tables 2-3. Each table presents relative RMSE along with coverage probability and average CI length for both HC3 standard errors and our new proposed correction to standard errors. These metrics are presented for each combination of heteroskedasticity scenario (row groups) and the estimator used for the skedastic function (columns). In all cases, metrics are calculated across 50,000 Monte Carlo replications, and the data generating process is as given earlier, with $\beta_0 = \beta_1 = 1$, $\sigma^2 = 4$, and $\sigma_x^2 = 5$. Table 2 presents results for $n = 100$, while Table 3 increases n to 500.

			Parametric				Nonparametric		Machine Learning
Heterosk.	Quantity	OLS	Correct	Main	WLS-S1	WLS-S2	KNN	Kernel	SVR
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
None	Rel. RMSE	1	1	1.085	1.007	1.08	1.157	1.051	1.077
	Coverage prob.	(0.947)	0.960(0.947)	0.950(0.936)	0.959(0.946)	0.950(0.936)	0.963(0.924)	0.936(0.918)	0.943(0.915)
	Rel. CI length	(1.000)	1.068(1.021)	1.070(1.022)	1.068(1.021)	1.067(1.020)	1.181(1.037)	1.010(0.971)	1.048(0.968)
Moderate	Rel. RMSE	1	0.43	1.147	0.439	1.136	0.563	0.513	0.479
	Coverage prob.	(0.939)	0.957(0.945)	0.950(0.932)	0.957(0.944)	0.949(0.932)	0.950(0.900)	0.920(0.865)	0.946(0.916)
	Rel. CI length	(1.000)	0.692(0.668)	1.125(1.057)	0.700(0.675)	1.122(1.053)	0.774(0.683)	0.678(0.612)	0.703(0.652)
Severe	Rel. RMSE	1	0.033	0.033	0.367	0.033	0.108	0.06	0.047
	Coverage prob.	(0.950)	0.967(0.956)	0.967(0.956)	0.969(0.952)	0.967(0.957)	0.936(0.883)	0.941(0.904)	0.955(0.932)
	Rel. CI length	(1.000)	0.160(0.153)	0.160(0.153)	0.826(0.765)	0.161(0.153)	0.263(0.226)	0.217(0.203)	0.212(0.198)

Table 2: FGLS estimates of slope parameter β_1 with various weight estimation methods using 100 observations per sample and 50,000 Monte Carlo samples. Row groups correspond to error covariance scenarios: homoskedastic, moderate heteroskedasticity, severe heteroskedasticity. Coverage probabilities and CI lengths are for HCFGLS estimator with HC3 results in parentheses.

			Parametric				Nonparametric		Machine Learning
Heterosk.	Quantity	OLS	Correct	Main	WLS-S1	WLS-S2	KNN	Kernel	SVR
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
None	Rel. RMSE	1	1	1.019	1.002	1.018	1.223	1.027	1.045
	Coverage prob.	(0.948)	0.952(0.948)	0.950(0.947)	0.952(0.948)	0.950(0.947)	0.965(0.927)	0.940(0.935)	0.946(0.935)
	Rel. CI length	(1.000)	1.015(1.004)	1.017(1.006)	1.016(1.004)	1.017(1.005)	1.206(1.051)	0.997(0.984)	1.017(0.987)
Moderate	Rel. RMSE	1	0.371	1.055	0.387	1.052	0.461	0.434	0.414
	Coverage prob.	(0.949)	0.952(0.949)	0.952(0.947)	0.952(0.950)	0.952(0.947)	0.952(0.922)	0.934(0.909)	0.951(0.943)
	Rel. CI length	(1.000)	0.615(0.611)	1.043(1.024)	0.629(0.624)	1.041(1.022)	0.683(0.630)	0.630(0.590)	0.645(0.631)
Severe	Rel. RMSE	1	0.008	0.008	0.483	0.008	0.029	0.017	0.016
	Coverage prob.	(0.954)	0.956(0.953)	0.956(0.953)	0.959(0.954)	0.956(0.954)	0.932(0.902)	0.950(0.941)	0.953(0.947)
	Rel. CI length	(1.000)	0.074(0.074)	0.074(0.074)	0.840(0.820)	0.075(0.074)	0.147(0.137)	0.131(0.128)	0.126(0.125)

Table 3: FGLS estimates of slope parameter β_1 with various weight estimation methods using 500 observations per sample and 50,000 Monte Carlo samples. Row groups correspond to error covariance scenarios: homoskedastic, moderate heteroskedasticity, severe heteroskedasticity. Coverage probabilities and CI lengths are for HCFGLS estimator with HC3 results in parentheses.

The first group of rows in Table 2, which come from a homoskedastic data generating process, match expectations based on classical econometric theory. OLS and FGLS with the correct functional form (which will be equivalent to OLS in this case) provide the most precise estimates and have coverage probabilities near the nominal levels. All other FGLS variants necessarily have higher RMSE due to the increased variance introduced by estimating an incorrect skedastic model. Absent heteroskedasticity, we note that SVR has lower RMSE than the nonparametric variants and only slightly higher RMSE than the parametric estimators. Our proposed corrected standard errors offer coverage probabilities near nominal levels, successfully addressing under-coverage of HC3 standard errors. We note that under homoskedasticity and $n = 100$, the average CI length for the SVR estimator is approximately 5% larger than that under OLS. The data-hungry machine learning approach we propose clearly has costs in small samples when there is no heteroskedasticity, though those costs appear minor.

When the data generating processes is characterized by heteroskedasticity, we begin to see the advantage of the machine learning approach. The second group of results in Table 2 correspond to moderate heteroskedasticity. The SVR-based estimator and the nonparametric estimators outperform OLS in terms of RMSE but do not achieve the efficiency of the FGLS estimator that uses the correct functional form. Further, SVR achieves over 91% of the reduction in RMSE achieved by FGLS with the correctly assumed functional form. That reduction is larger than all other approaches except for WLS-S1, which is nearly identical to the correct functional form. The principal advantage is that the efficiency gains under SVR come without functional form assumptions. The test size issues remain when using HC3 standard errors, but our proposed correction again yields the proper coverage probability.

As the form of heteroskedasticity gets more severe (last group of results in Table 2), we see that SVR continues to outperform OLS, this time achieving over 98% of the reduction in RMSE that would result from FGLS with correctly specified functional form. Our proposed standard error correction again leads to correctly sized tests.

Increasing the sample size to $n = 500$ yields similar results (Table 3). SVR and kernel regression continue to offer favorable RMSE performance. More importantly, the improvements in RMSE over OLS grow larger. Further, coverage based on HC3 standard errors moves toward the nominal rate in all cases, while our proposed correction continues to offer correct size.

5.2 Boston housing data

Estimation results for the wild bootstrapped Boston housing data are presented in Table 4. The SVR-based FGLS estimator again has small empirical RMSE and confidence intervals in comparison to OLS with HC3 standard errors. Similarly, the coverage probabilities for the corrected standard errors we proposed are near their nominal levels. For comparison, Table 4 also includes the same metrics for the WLS-S1 and WLS-S2 estimators proposed by Romano & Wolf (2017). The proposed SVR estimator has favorable RMSE and confidence interval length and similar coverage properties. Thus, FGLS using support vector regression to estimate the skedastic function, together with our proposed correction to standard errors, appears to have desirable properties on a realistic dataset. FGLS based on kernel regression slightly outperforms SVR in terms of RMSE, but coverage is superior for SVR.

		SVR	kernel	WLS-S1	WLS-S2
(Intercept)	Rel. RMSE	0.459	0.432	0.615	0.606
	Coverage prob.	0.944(0.907)	0.927(0.885)	0.954(0.947)	0.952(0.946)
	Rel. CI length	0.661(0.611)	0.611(0.563)	0.803(0.787)	0.793(0.778)
log(nox)	Rel. RMSE	0.51	0.471	0.674	0.651
	Coverage prob.	0.951(0.915)	0.943(0.901)	0.955(0.949)	0.953(0.947)
	Rel. CI length	0.714(0.653)	0.676(0.611)	0.836(0.821)	0.816(0.802)
log(dist)	Rel. RMSE	0.402	0.372	0.51	0.509
	Coverage prob.	0.945(0.911)	0.923(0.882)	0.955(0.949)	0.954(0.948)
	Rel. CI length	0.621(0.577)	0.562(0.519)	0.732(0.721)	0.730(0.718)
rooms	Rel. RMSE	0.373	0.358	0.504	0.484
	Coverage prob.	0.936(0.889)	0.905(0.855)	0.957(0.949)	0.954(0.946)
	Rel. CI length	0.587(0.532)	0.533(0.484)	0.736(0.716)	0.707(0.690)
stratio	Rel. RMSE	0.719	0.716	0.932	1.131
	Coverage prob.	0.948(0.916)	0.937(0.896)	0.956(0.949)	0.954(0.947)
	Rel. CI length	0.832(0.767)	0.799(0.732)	0.980(0.963)	1.069(1.050)

Table 4: FGLS estimates of Boston housing data model with 50,000 replicates. RMSE and CI length are reported relative to the RMSE and CI length of OLS with HC3 standard erors. Coverage probabilities and CI lengths are for HCFGLS estimator with HC3 results in parentheses.

6 Conclusion

In keeping with the trend in econometrics to relax assumptions, the use of feasible generalized least squares has declined in favor of the use of ordinary least squares with robust standard errors. As is well known, that approach comes at the cost of some efficiency. In this paper, we have proposed and investigated the use of support vector regression to estimate the form of heteroskedasticity as part of FGLS estimation. Our proposal is motivated by the flexibility of SVR coupled with its controls on variance and lower susceptibility to outliers. This compromise is likely to prove useful in the context of FGLS: skedastic function estimators which are not variable enough cannot capture differences in weights across observations, while high variance estimators may introduce too much variability in the final estimates. Further, to account for the flexibility afforded by machine learning tools, we have also proposed corrections to robust standard errors computed after FGLS.

Our simulation results suggest that using support vector regression as a part of FGLS imposes minor efficiency costs under homoskedasticity, but may offer substantial efficiency improvements in the presence of heteroskedasticity. The SVR estimator we propose offers comparable performance to FGLS using kernel estimation of the skedastic function, but tends to outperform the kernel estimator as the heteroskedasticity becomes more severe. This accords with intuition: SVR and other machine learning tools penalize model complexity, which makes them less susceptible to extreme observations. Together with recent work by Romano & Wolf (2017) providing guidance on inference after FGLS, we believe our approach can contribute to making FGLS a viable, modern tool in researchers' econometrics toolkits.

Appendix

Standard error correction

The first part of the derivation below is based closely on Rothenberg (1988), but approximates the FGLS estimator variance rather than the distribution of a test statistic. In addition, the motivation below considers the skedastic function to be parameterized directly by observation-specific error variances rather than a smaller set of parameters and a smoother model. While individual observation variances cannot themselves be consistently estimated, the formulation is useful in illustrating the correction.

Let C be a binary vector of length k containing zeros except for a one in the cth position. The FGLS estimator of the cth coefficient in $\hat{\beta}_{FGLS}$, which we call $\hat{\beta}_{FGLS,c}$, can be approximated via Taylor expansion around the cth element in the GLS estimator:

$$\hat{\beta}_{FGLS,c} \approx \hat{\beta}_{GLS,c} + \sum_i (\hat{\sigma}_i^2 - \sigma_i^2) \frac{\partial}{\partial \sigma_i^2} C (X'W^{-1}X)^{-1} X'W^{-1}Y,$$

with the sum over observations i. Evaluating partial derivatives and simplifying yields:

$$\begin{aligned}\hat{\beta}_{FGLS,c} \approx & \hat{\beta}_{GLS,c} + \\ & \sum_i C(\hat{\sigma}_i^2 - \sigma_i^2)(X'W^{-1}X)^{-1}X'W^{-1}W_i \left\{W^{-1}X(X'W^{-1}X)^{-1}X'W^{-1} - W^{-1}\right\} u,\end{aligned}$$

where W_i is a matrix with a one in the ith diagonal and zeros elsewhere, which comes from the partial derivative of the matrix W with respect to σ_i^2. This approximation can be used to express the variance of the difference between these estimators:

$$Var(\sqrt{n}(\hat{\beta}_{FGLS,c} - \hat{\beta}_{GLS,c})) \approx \sum_i \sum_j A'[W^{-1}(I_n - H_{GLS})]_{ij} cov(\tilde{u}_i^2, \tilde{u}_j^2) A. \tag{A.1}$$

where $A = W^{-1}X(X'W^{-1}X)^{-1}C$, I_n is the n by n identity matrix, and H_{GLS} is the hat matrix of the GLS estimator. Therefore $[W^{-1}(I_n - H_{GLS})]_{ij}$ measures the influence of outcome j's influence on the weighted GLS residual of observation i. Importantly, $cov(\tilde{u}_i^2, \tilde{u}_j^2)$ depends on the method used to estimate the skedastic function. If that quantity is known or estimable, this expression can be used to estimate the FGLS variances directly.

We seek a correction when closed form expressions for the covariance terms $cov(\tilde{u}_i^2, \tilde{u}_j^2)$ may not be known. First, note that the summand in (A.1) takes a familiar sandwich form with the matrix A also appearing in the standard HC estimators as applied to FGLS in Romano & Wolf (2017). We thus focus exclusively on the terms $[W^{-1}(I_n - H_{GLS})]_{ij} cov(\tilde{u}_i^2, \tilde{u}_j^2)$. Second, we make the assumption that $cov(\tilde{u}_i^2, \tilde{u}_j^2) \approx 0$ for $i \neq j$ so that we may focus mostly on the role of variance of the skedastic function estimator rather than covariance. For more flexible models such as kernel estimators or support vector regression, this approximation is likely reasonable, since variance estimates are influenced only by nearby points, especially if the data are sufficiently dense in covariate space. This assumption allows us to rewrite the approximation as:

$$Var(\sqrt{n}(\hat{\beta}_{FGLS,c} - \hat{\beta}_{GLS,c}) \approx \sum_i A'[W^{-1}(I_n - H_{GLS})]_{ii} Var(\tilde{u}_i^2) A.$$

We continue to simplify this expression by approximating $Var(\tilde{u}_i^2)$. In many implementations of FGLS, including all that we investigate here, $\tilde{u}_i^2$ is derived by fitting a model to $log(\hat{u}_i^2)$ exponentiating the resulting predictions, i.e., $\tilde{u}_i^2 = e^{\widehat{log(\hat{u}_i^2)}}$. If, as intended, $\widehat{log(\hat{u}_i^2)}$ is a reasonable approximation to $log(\sigma_i^2)$, then by the delta method:

$$Var(\tilde{u}_i^2) = Var\left(e^{\widehat{log(\hat{u}_i^2)}}\right) \approx \left(e^{log(\sigma_i^2)}\right)^2 Var\left(\widehat{log(\hat{u}_i^2)}\right) = \left(\sigma_i^2\right)^2 Var\left(\widehat{log(\hat{u}_i^2)}\right).$$

Next, note that the (i,i) element of $W^{-1}(I_n - H_{GLS})$ is simply $\frac{(1-h_{i,GLS})}{\sigma_i^2}$, where h_i is the ith diagonal element of H_{GLS}. Therefore

$$[W^{-1}(I_n - H_{GLS})]_{ii} Var(\tilde{u}_i^2) \approx \sigma_i^2(1 - h_{i,GLS}) Var\left(\widehat{log(\hat{u}_i^2)}\right).$$

Adding this correction to the meat of the HC3 sandwich estimator and simplifying yields corrected diagonal entries:

$$\hat{\Sigma}_{ii}^{FGLS} = \frac{\hat{u}_{i,FGLS}^2}{(\tilde{u}_i^2)^2}\left(\frac{1}{(1-h_{i,FGLS})^2} + \frac{\sigma_i^2}{\frac{\hat{u}_{i,FGLS}^2}{(1-h_{i,GLS})}} Var\left(\widehat{log(\hat{u}_i^2)}\right)\right).$$

The denominator of the fraction involving σ_i^2 is itself an estimate of σ_i^2, so we might approximate the expression by

$$\hat{\Sigma}_{ii}^{FGLS} = \frac{\hat{u}_{i,FGLS}^2}{(\tilde{u}_i^2)^2}\left(\frac{1}{(1-h_{i,FGLS})^2} + Var\left(\widehat{log(\hat{u}_i^2)}\right)\right).$$

To estimate $Var\left(\widehat{log(\hat{u}_i^2)}\right)$, let $\mathcal{M}(\log(\hat{u}^2), X)$ denote the method used to estimate the (log) skedastic function, and let $\mathcal{M}(\log(\hat{u}^2), X)_i$ denote the fitted value for the ith observation. Again applying the delta method and simplifying:

$$Var\left(\widehat{log(\hat{u}_i^2)}\right) \approx \sum_j 4\frac{\left(\frac{\partial\mathcal{M}(\log(\hat{u}^2),X)_i}{\partial log(\hat{u}_j^2)}\right)^2 Var(\hat{u}_j)}{\hat{u}_j^2}.$$

Since $Var(\hat{u}_j) = (1 - h_{j,OLS})\sigma_j^2$ and $\frac{\hat{u}_j^2}{(1-h_{j,OLS})}$ is an estimate of σ_j^2, we approximate:

$$Var\left(\widehat{log(\hat{u}_i^2)}\right) \approx 4\sum_j \left(\frac{\partial\mathcal{M}(\log(\hat{u}^2),X)_i}{\partial log(\hat{u}_j^2)}\right)^2.$$

The final step of our approximation is motivated by the case in which the (log) skedastic function is estimated via OLS. In that case, the partial derivatives are simply the h_{ij} entries in the hat matrix of the skedastic model. Then since $h_i = \sum_j h_{ij}^2$:

$$Var\left(\widehat{log(\hat{u}_i^2)}\right) \approx 4h_{i,sked} = 4\frac{\partial\mathcal{M}(\log(\hat{u}^2), X)_i}{\partial log(\hat{u}_i^2)} \tag{A.2}$$

We are not proposing to estimate the skedastic function via OLS, but the last expression hints at an approximation. We do not necessarily know $\frac{\partial\mathcal{M}(\log(\hat{u}^2),X)_i}{\partial log(\hat{u}_i^2)}$ for an arbitrary model, but it is the summand in the generalized degrees of freedom (GDF) in Ye (1998). With this in mind, we propose:

$$Var\left(\widehat{log(\hat{u}_i^2)}\right) \approx 4\frac{h_{i,OLS}}{k}\hat{df} = 4\frac{h_{i,OLS}}{\sum_j h_{j,OLS}}\hat{df},$$

where $\hat{df}$ is the GDF for the method used to estimate the skedastic function. The fraction $\frac{h_{i,OLS}}{k}$ captures the relative influence of observation i if OLS with the main model covariates were used to estimate the skedastic function, while $\hat{df}$ captures the aggregate sensitivity of whatever skedastic model is actually used. If OLS is in fact used to estimate the skedastic function, $\hat{df} = k$ and the expression collapses back down to (A.2).

Finally, plugging this into our correction to the meat term of the sandwich estimator, we have:

$$\hat{\Sigma}_{ii}^{FGLS} = \frac{\hat{u}_{i,fgls}^2}{(\tilde{u}_i^2)^2}\left(\frac{1}{(1-h_{i,FGLS})^2} + 4h_{i,OLS}\frac{\hat{df}}{k}\right),$$

which is the correction proposed in the main text and used in our simulations.

SUPPLEMENTARY MATERIAL

R implementation: R code is provided to reproduce results contained in the paper and to facilitate use of the proposed method. The .zip file also includes the dataset used for the Boston housing example. (.zip file)

References

Carroll, R. J. (1982), 'Adapting for heteroscedasticity in linear models', *The Annals of Statistics* pp. 1224–1233.

Cribari-Neto, F. (2004), 'Asymptotic inference under heteroskedasticity of unknown form', *Computational Statistics & Data Analysis* **45**(2), 215–233.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., Vapnik, V. et al. (1997), 'Support vector regression machines', *Advances in neural information processing systems* **9**, 155–161.

Eicker, F. (1963), 'Asymptotic normality and consistency of the least squares estimators for families of linear regressions', *The Annals of Mathematical Statistics* pp. 447–456.

Friedman, J., Hastie, T. & Tibshirani, R. (2001), *The elements of statistical learning*, Vol. 1, Springer series in statistics Springer, Berlin.

Gunter, L. & Zhu, J. (2007), 'Efficient computation and model selection for the support vector regression', *Neural Computation* **19**(6), 1633–1655.

Harrison, D. & Rubinfeld, D. L. (1978), 'Hedonic housing prices and the demand for clean air', *Journal of environmental economics and management* **5**(1), 81–102.

Huber, P. J. (1967), The behavior of maximum likelihood estimates under nonstandard conditions, *in* 'Proceedings of the fifth Berkeley symposium on mathematical statistics and probability', Vol. 1, pp. 221–233.

Li, Q. & Racine, J. (2004), 'Cross-validated local linear nonparametric regression', *Statistica Sinica* pp. 485–512.

MacKinnon, J. G. & White, H. (1985), 'Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties', *Journal of econometrics* **29**(3), 305–325.

Micchelli, C. A., Xu, Y. & Zhang, H. (2006), 'Universal kernels', *Journal of Machine Learning Research* **7**(Dec), 2651–2667.

Newey, W. K. (1994), 'Series estimation of regression functionals', *Econometric Theory* **10**(01), 1–28.

O'Hara, M. & Parmeter, C. F. (2013), 'Nonparametric generalized least squares in applied regression analysis', *Pacific Economic Review* **18**(4), 456–474.

Racine, J. & Li, Q. (2004), 'Nonparametric estimation of regression functions with both categorical and continuous data', *Journal of Econometrics* **119**(1), 99–130.

Robinson, P. M. (1987), 'Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form', *Econometrica: Journal of the Econometric Society* pp. 875–891.

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

Rothenberg, T. J. (1988), 'Approximate power functions for some robust tests of regression coefficients', *Econometrica: Journal of the Econometric Society* pp. 997–1019.

Steinwart, I. & Christmann, A. (2008), *Support vector machines*, Springer Science & Business Media.

Vapnik, V. N. (1995), *The nature of statistical learning theory*, Springer.

White, H. (1980), 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity', *Econometrica: Journal of the Econometric Society* pp. 817–838.

Windmeijer, F. (2005), 'A finite sample correction for the variance of linear efficient two-step gmm estimators', *Journal of econometrics* **126**(1), 25–51.

Wooldridge, J. M. (2010), *Econometric analysis of cross section and panel data*, MIT press.

Wooldridge, J. M. (2015), *Introductory econometrics: A modern approach*, Nelson Education.

Ye, J. (1998), 'On measuring and correcting the effects of data mining and model selection', *Journal of the American Statistical Association* **93**(441), 120–131.