Heteroskedastic linear regression: steps towards adaptivity, efficiency, and robustness

Dimitris N. Politis and Stefanos Poulis

Abstract In linear regression with heteroscedastic errors, the Generalized Least Squares (GLS) estimator is optimal, i.e., it is the Best Linear Unbiased Estimator (BLUE). The Ordinary Least Squares (OLS) estimator is suboptimal but still valid, i.e., unbiased and consistent. Halbert White, in his seminal paper (Econometrica, 1980) used the OLS residuals in order to obtain an estimate of the standard error of the OLS estimator under an unknown structure of the underlying heteroscedasticity. The GLS estimator similarly depends on the unknown heteroscedasticity, and is thus intractable. In this paper, we introduce two different approximations to the optimal GLS estimator; the starting point for both approaches is in the spirit of White's correction, i.e., using the OLS residuals to get a rough estimate of the underlying heteroscedasticity. We show how the new estimators can benefit from the Wild Bootstrap both in terms of optimising them, but also in terms of providing valid standard errors for them despite their complicated construction. The performance of the new estimators is compared via simulations to the OLS and to the exact (but intractable) GLS.

1 Introduction

Standard regression methods rely on the assumption that the regression errors are either independent, identically distributed (i.i.d.), or at least being uncorrelated having the same variance; this latter property is called *homoscedasticity*. The Generalized Least Squares (GLS) estimator is Best Linear Unbiased Estimator (BLUE) but its computation depends on the structure of the underlying heteroscedasticity. Typically, this structure is unknown, and the GLS estimator is intractable; in this case, practitioners may be forced to use the traditional Ordinary Least Squares (OLS) estimator which will still be valid, i.e., unbiased and consistent, under general conditions.

Under the assumption that the error variance is an unknown but smooth function of the regressors it is possible to give an approximation to the GLS estimator. For example, Caroll (1982) showed that one can construct estimates of regression parameters that are asymptotically equivalent to the weighted least squares estimates. Chatterjee and Mächler (1997) proposed an iterative weighted least squares algorithm to approximate the optimal GLS. In a different spirit, Yuan and Whaba (2004) introduced a penalized likelihood proce-

Dimitris N. Politis
University of California at San Diego, La Jolla, CA 92093-0112, USA , e-mail: dpolitis@ucsd.edu

Stefanos Poulis
University of California at San Diego, La Jolla, CA 92093-0112, USA e-mail: spoulis@ucsd.edu

dure to estimate the conditional mean and variance simultaneously. In the same framework, Le, Smola, and Canu (2005), using Gaussian process regression, estimate the conditional mean and variance by estimating the natural parameters of the exponential family representation of the Gaussian distribution.

What happens if no smoothness assumption on the error variance can be made? In his seminal paper, White (1980) used the OLS residuals in order to get an estimate of the standard error of the OLS estimator that is valid, i.e., consistent, under an unknown structure of the underlying heteroscedasticity. In the paper at hand, we introduce two new approximations to the optimal GLS estimator; the starting point for both approaches is in the spirit of White's (1980) correction, i.e., using the OLS residuals to get a rough estimate of the underlying heteroscedasticity. The paper is structured as follows; in Section 2 we formally state the problem, introduce our first estimator, and show how a convex combination of the *under-correcting* OLS and the *over-correcting* approximate GLS estimators can yield improved results. In Section 3, in our effort to approximate the quantities needed for the GLS, we introduce a second estimator. In Section 4 we present a series of simulation experiments to compare performance of the new estimators to the OLS and to the exact (but intractable) GLS.

2 Adaptive estimation: A first attempt

Consider the general linear regression set-up where the data vector $Y = (Y_1, \ldots, Y_n)'$ satisfies:

$$Y = X\beta + \varepsilon. \tag{1}$$

As usual, β is a $p \times 1$ unknown parameter vector, X is an $n \times p$ matrix of observable regressors, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ is an unobservable error vector. The regressors may be fixed or random variables (r.v.); in either case it will be assumed that the regressor matrix X is independent from the error vector ε, and that the $n \times p$ matrix is of full rank almost surely.

Letting x_i denote the ith row of the matrix X, we will further assume that

$$\{(x_i, \varepsilon_i) \text{ for } i = 1, \ldots, n\} \text{ is a sequence of independent r.v.'s} \tag{2}$$

and that the first two moments of ε are finite and satisfy:

$$E(\varepsilon) = 0 \text{ and } V(\varepsilon) = \Sigma \text{ where } \Sigma = diag(\sigma_1^2, \ldots, \sigma_n^2), \tag{3}$$

i.e., the ε_i errors are mean zero, uncorrelated but with heteroskedasticity of arbitrary form. The OLS estimator of β is $\hat{\beta}_{LS} = (X'X)^{-1}X'Y$. Its variance-covariance matrix, conditional on X, is given by

$$V(\hat{\beta}_{LS}) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}$$

that can be estimated consistently by White's (1980) heteroskedasticity consistent estimator (HCE):

$$(X'X)^{-1}X'\hat{\Sigma}X(X'X)^{-1}$$

where

$$\hat{\Sigma} = diag(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_n^2) \text{ with } \hat{\sigma}_i^2 = r_i^2/(1-h_i)^{\gamma}. \tag{4}$$

In the above, r_i is the ith element of the residual vector $r = Y - X\hat{\beta}_{LS}$, and h_i is the 'leverage' of point x_i, i.e., the ith diagonal element of the projection ('hat') matrix $H = X(X'X)^{-1}X'$. White's (1980) original proposal

had $\gamma = 0$ in eq. (4); later proposals recommended $\gamma = 1$, i.e., studentized residuals, or $\gamma = 2$, i.e., delete-1 jackknife residuals; see MacKinnon (2012) for a review.

Nevertheless, in the presence of heteroskedasticity, $\hat{\beta}_{LS}$ is not optimal. Efficiency in this case is attained by the GLS estimator $\hat{\beta}_{GLS}$ which is the solution of

$$(X'\Sigma^{-1}X)\,\hat{\beta}_{GLS} = X'\Sigma^{-1}Y. \tag{5}$$

Under the stated assumptions, the variance-covariance matrix of $\hat{\beta}_{GLS}$, conditional on X, is given by

$$V(\hat{\beta}_{GLS}) = (X'\Sigma^{-1}X)^{-1}. \tag{6}$$

The problem, of course, is that Σ is unknown, so $\hat{\beta}_{GLS}$ is unobtainable. However, despite the fact that $\hat{\Sigma}$ is an inconsistent estimator of Σ, we may still construct estimators of the matrices $X'\Sigma^{-1}X$ and $X'\Sigma^{-1}$ that are needed to compute $\hat{\beta}_{GLS}$; this is done in the spirit of White's (1980) HCE.

Before doing that though it is important to consider the possibility that $\hat{\beta}_{GLS}$ is not well-defined due to the fact that Σ might be non-invertible. We can define a small perturbation of Σ that is invertible; to do this, let $\delta > 0$ and define $\Sigma_\delta = diag(\sigma_1^2 + \delta, \ldots, \sigma_n^2 + \delta)$. We now define $\hat{\beta}_{GLS,\delta}$ as the solution of

$$(X'\Sigma_\delta^{-1}X)\,\hat{\beta}_{GLS,\delta} = X'\Sigma_\delta^{-1}Y.$$

Note that $\hat{\beta}_{GLS,\delta}$ is always well-defined, and—for δ small enough—is close to $\hat{\beta}_{GLS}$ when it is well-defined.

In the same spirit, let $\delta > 0$ and define

$$\tilde{\Sigma}_\delta = diag(\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_n^2) \;\text{ with }\; \tilde{\sigma}_i^2 = \frac{r_i^2 + \delta}{(1-h_i)^\gamma} \tag{7}$$

where r and γ are as in eq. (4). Note that $\tilde{\Sigma}_\delta$ reduces to $\hat{\Sigma}$ when $\delta = 0$; the advantage of $\tilde{\Sigma}_\delta$, however, is that it is always invertible with $\tilde{\Sigma}_\delta^{-1} = diag(\tilde{\sigma}_1^{-2}, \ldots, \tilde{\sigma}_n^{-2})$.

Remark 2.1 In practice, δ could/should be taken to be a fraction of the residual sample variance $S^2 = (n-p)^{-1}\sum_{i=1}^n r_i^2$, e.g., $\delta = 0.01\,S^2$ or $\delta = 0.001\,S^2$; among other things, this ensures equivariance of $\hat{\Sigma}_\delta$.

We now define a preliminary estimator $\tilde{\beta}_\delta$ as the solution of

$$(X'\tilde{\Sigma}_\delta^{-1}X)\,\tilde{\beta}_\delta = X'\tilde{\Sigma}_\delta^{-1}Y. \tag{8}$$

To investigate the behavior of the preliminary estimator $\tilde{\beta}_\delta$ we need to define the diagonal matrix W_δ that has as ith element the quantity $E[(r_i^2 + \delta)^{-1}]$. Now let $\tilde{\beta}_{W,\delta}$ as the solution of

$$(X'W_\delta X)\,\tilde{\beta}_{W,\delta} = X'W_\delta Y. \tag{9}$$

Under conditions similar to the assumptions of Theorem 1 of White (1980), we can now claim the following *large-sample* approximations:

$$X'\tilde{\Sigma}_\delta^{-1}X \approx_c X'W_\delta X \;\text{ and }\; X'\tilde{\Sigma}_\delta^{-1} \approx_c X'W_\delta \tag{10}$$

where symbol $A \approx_c B$ for matrices $A = (a_{ij})$ and $B = (b_{ij})$ is short-hand to denote that $a_{ij} \approx b_{ij}$ for all i, j, i.e., coordinate-wise approximation. As a consequence, it follows that $\tilde{\beta}_\delta \approx \tilde{\beta}_{W,\delta}$ for large enough samples.

Now if we could claim that $W_\delta \approx_c \tilde{\Sigma}_\delta^{-1}$, then we would have that $\tilde{\beta}_{W,\delta} \approx \hat{\beta}_{GLS,\delta}$, and thus $\tilde{\beta}_\delta$ would be a consistent approximation to the GLS estimator. However, the approximation $W_\delta \approx_c \tilde{\Sigma}_\delta^{-1}$ is not a good one. In

fact, by Jensen's inequality we have $Er_i^{-2} \geq 1/Er_i^2 = 1/\sigma_i^2$; hence, it follows that W_δ will be biased upward as an estimator of Σ_δ^{-1}. In this sense, $\tilde{\beta}_\delta$ is an *over-correction* in trying to take account of the covariance of the errors. Since the OLS estimator $\hat{\beta}_{LS}$ is an *under-correction*, it is interesting to explore the hybrid estimator $\tilde{\beta}_{\delta,\lambda}$ whose kth coordinate is given by

$$\tilde{\beta}_{\delta,\lambda;k} = \lambda_k \tilde{\beta}_{\delta;k} + (1-\lambda_k)\hat{\beta}_{LS;k} \tag{11}$$

where $\lambda_k, \tilde{\beta}_{\delta;k}, \hat{\beta}_{LS;k}$ denote the kth coordinates of $\lambda, \tilde{\beta}_\delta, \hat{\beta}_{LS}$ respectively, and λ is a *mixing* weight vector of the practitioner's choice. By choosing the tuning parameter λ to be multi-dimensional, we allow each coordinate to receive a different weight; this might be especially helpful if/when heteroscedasticity is due to only a subset of the predictors. In this case, eq. (11) becomes

$$\tilde{\beta}_{\delta,\lambda} = \lambda \circ \tilde{\beta}_\delta + (\mathbf{1}-\lambda) \circ \hat{\beta}_{LS} \tag{12}$$

where $(\circ)$ denotes the *Hadamard* (point-wise) vector product.

Even if efficiency is not attained, $\tilde{\beta}_{\delta,\lambda}$ is still a step in the right direction towards the GLS estimator. In addition, it satisfies the two robustness criteria of Chatterjee and Mächler (1997) as it gives reduced weight to (Y_i, x_i) points that involve either an outlier in the error ε_i, or an outlier in the regressor x_i, i.e., a high 'leverage' point.

Remark 2.2 To fine-tune the choice of the weight vector λ, the *wild bootstrap* of Wu (1986) is found useful. Here we will use the simplest version proposed by Liu (1988) and further discussed by Davidson and Flachaire (2008), namely bootstrapping the signs of the residuals. Specifically, define the bootstrap residuals

$$r_i^* = s_i \, r_i/(1-h_i)^{\gamma/2} \;\text{ for }\; i = 1,\ldots,n$$

where $s_1,\ldots,s_n$ are i.i.d. sign changes with $Prob\{s_i = 1\} = Prob\{s_i = -1\} = 1/2$, and $\gamma = 1$ or 2, i.e., studentized or jackknife residuals respectively as in eq. (4). The use of jackknife residuals—also known as predictive residuals—in bootstrapping has also been recommended by Politis (2010, 2013).

The wild bootstrap datasets are generated via eq. (1) using the same X matrix, and β replaced by $\hat{\beta}_{LS}$. Letting $r^* = (r_1^*,\ldots,r_n^*)'$, the bootstrap data vector Y^* sastifies the equation:

$$Y^* = X\hat{\beta}_{LS} + r^*. \tag{13}$$

Then the pseudo-data Y^* can be treated as (approximate) replicates of the original data. Of course, this is justified under the additional assumption that the errors ε_i are (approximately) symmetrically distributed around zero; if this is not true, then the error skewness can be incorporated in the wild bootstrap by using a non-symmetric auxiliary distribution for the s_i random variables used above.

The bootstrap procedure for choosing the weight vector λ is summarized in Algorithm 1 below that works as follows. Firstly, d denotes the number of candidate λ vectors that the bootstrap procedure should choose from. Typically, a grid will be formed spanning the allowed [0,1] interval for each coordinate of λ; a large d corresponds to a fine grid. Then, B bootstrap datasets are generated based on eq. (13). For a particular choice of the vector λ, $\tilde{\beta}_{\delta,\lambda}$ is computed as in eq. (12) on every bootstrap dataset. The algorithm finds λ_{opt} that minimizes the empirical MSE with respect to the $\hat{\beta}_{LS}$, as estimated in the original dataset. In the algorithm, k denotes the kth parameter vector, and j denotes the jth bootstrap dataset.

Algorithm 1 Find optimal tuning parameter vector λ for $\tilde{\beta}_{\delta,\lambda}$ via Wild Bootstrap

Choose d candidate tuning parameter vectors $\lambda^{(1)},\ldots,\lambda^{(d)}$, where $\lambda^{(k)} \in (0,1)^p$
Create B wild bootstrap datasets by eq. (13)
for k in 1 to d **do**
  for j in 1 to B **do**
    $\tilde{\beta}_{\delta}^{(j)} = (X'\tilde{\Sigma}_{\delta}^{-1(j)}X)^{-1}X'\tilde{\Sigma}_{\delta}^{-1(j)}Y^{(j)}$
    $\hat{\beta}_{LS}^{(j)} = (X'X)^{-1}X'Y^{(j)}$
    $\tilde{\beta}_{\delta,\lambda}^{(kj)} = \lambda^{(k)} \circ \tilde{\beta}_{\delta}^{(j)} + (\mathbf{1}-\lambda^{(k)}) \circ \hat{\beta}_{LS}^{(j)}$
  end for
end for
$\lambda_{opt} = \underset{1\le k\le d}{\operatorname{argmin}} \frac{1}{B}\sum_{j=1}^{B}(\tilde{\beta}_{\delta,\lambda}^{(kj)} - \hat{\beta}_{LS})^2$
output $\tilde{\beta}_{\delta,\lambda} = \lambda_{opt} \circ \hat{\beta}_{\delta} + (\mathbf{1}-\lambda_{opt}) \circ \hat{\beta}_{LS}$

3 Adaptive estimation: a second attempt

One may consider linearization as an alternative approximation to adaptive GLS. To do this, consider linearizing the $1/x$ function around a point x_0, i.e.,

$$1/x \approx 1/x_0 - (1/x_0^2) * (x - x_0) = 2/x_0 - (1/x_0^2) * x. \tag{14}$$

Using the above approximation on each of the elements of the diagonal matrix Σ, it follows that we may be able to approximate the $X'\Sigma^{-1}X$ needed for GLS by: $2x_0^{-1}X'X - x_0^{-2}X'\Sigma X$. But the last expression is estimable by a usual HCE procedure, i.e., by:

$$2x_0^{-1}X'X - x_0^{-2}X'\hat{\Sigma}X \tag{15}$$

where $\hat{\Sigma}$ is defined as in eq. (4). Similarly, $X'\Sigma^{-1} \approx 2x_0^{-1}X' - x_0^{-2}X'\Sigma$ which is estimable by:

$$2x_0^{-1}X' - x_0^{-2}X'\hat{\Sigma}. \tag{16}$$

Because of the convexity of the $1/x$ function the linear approximation (14) becomes a supporting line, i.e.,

$$1/x > 2/x_0 - (1/x_0^2) * x$$

for all positive $x \neq x_0$ (and becomes an approximation for x close to x_0). Hence, the quantity in eq. (15) *under*-estimates $X'\Sigma^{-1}X$ in some sense.

Note however, that the RHS of (14) goes negative when $x > 2x_0$ in which case it is not only useless, but may also cause positive definitess problems if the approximation is applied to $1/\sigma_i^2$ or $1/r_i^2$ type quantities. To address this, we propose a two-step solution:

(a) Truncate (rather Winsorize) the r_i^2 by letting $\tilde{r}_i = r_i$ when $|r_i| < \zeta$, and $\tilde{r}_i = \zeta \cdot sign(r_i)$ when $|r_i| \geq \zeta$ for some $\zeta > 0$. This implies that the influence of too large residuals will be bounded. The number ζ can be picked in a data-based manner, e.g. we can take ζ^2 as the λ quantile of the empirical distribution of the r_i^2 for some $\lambda \in (0,1)$ but close to 1; for example, $\lambda = 0.9$ may be reasonable in which case ζ^2 is the upper decile.

(b) Choose a big enough linearization point x_0 to ensure positive defininess of the quantity in eq. (15). To do this, let $x_0 = (\zeta^2 - \delta)/2$ where δ is a small, positive number; as in Remark 2.1, δ can be chosen to be a fraction of S^2 or–in this case–a fraction of ζ^2, e.g. $\delta = 0.001\zeta^2$.

With choices of x_0 and λ as in parts (a) and (b) above, we now define $\breve{\beta}_{\delta,\lambda}$ as the solution of

$$(2x_0^{-1}X'X - x_0^{-2}X'\hat{\Sigma}X)\check{\beta}_{\delta,\lambda} = 2x_0^{-1}X'Y - x_0^{-2}X'\hat{\Sigma}Y$$

which is equivalent to

$$(2X'X - x_0^{-1}X'\hat{\Sigma}X)\check{\beta}_{\delta,\lambda} = 2X'Y - x_0^{-1}X'\hat{\Sigma}Y. \tag{17}$$

The construction of estimator $\check{\beta}_{\delta,\lambda}$ is described in full detail in Algorithm 2. The procedure for the choice of the optimal linearization point (or truncation point) is similar to that of Algorithm 1. B bootstrap datasets are generated based on eq. (13). For a particular choice of truncation point, $\check{\beta}_{\delta,\lambda}$ is computed as in eq. (17) on every bootstrap dataset. The algorithm finds ζ_{opt} that minimizes the empirical MSE with respect to the $\hat{\beta}_{LS}$, as estimated in the original dataset.

In Algorithm 2, d denotes the total number of possible truncation points to be examined and denotes how many quantiles of the OLS squared residuals are considered as candidates; for example, one may consider the 70%, 80%, and 90% quantiles, i.e., $d = 3$. In the Algorithm, k denotes the kth candidate truncation point and j denotes the jth bootstrap dataset.

Algorithm 2 Find optimal linearization point x_0 for $\check{\beta}_{\delta,\lambda}$ via Wild Bootstrap

Choose d candidate truncation parameters $\zeta^{(1)}, \ldots, \zeta^{(d)}$.
Create B wild bootstrap datasets by eq. (13)
for k in 1 to d **do**
    Set $\tilde{r}_i \leftarrow r_i$ when $|r_i| < \zeta^{(k)}$ and $\tilde{r}_i \leftarrow \zeta^{(k)} sign(r_i)$ otherwise
    $x_0^{(k)} = \frac{(\zeta^{(k)})^2 - \delta}{2}$, for some small δ
    for j in 1 to B **do**
        $\check{\beta}_{\delta,\lambda}^{(kj)} = (X'X - \frac{1}{x_0^{(k)}}X'\hat{\Sigma}^{(j)}X)^{-1}(X'Y^{(j)} - \frac{1}{x_0^{(k)}}X'\hat{\Sigma}^{(j)}Y^{(j)})$
    end for
end for
$\zeta_{opt} = \underset{1\le k\le d}{\operatorname{argmin}} \frac{1}{B}\sum_{j=1}^{B}(\check{\beta}_{\delta,\lambda}^{(kj)} - \hat{\beta}_{LS})^2$
$x_0 \leftarrow \frac{\zeta_{opt}^2 - \delta}{2}$
$\check{\beta}_{\delta,\lambda} = (X'X - \frac{1}{x_0}X'\hat{\Sigma}X)^{-1}(X'Y - \frac{1}{x_0}X'\hat{\Sigma}Y)$

Remark 3.1 Computing the distribution of $\tilde{\beta}_\delta$, $\tilde{\beta}_{\delta,\lambda}$, and $\check{\beta}_{\delta,\lambda}$ analytically seems a daunting task even under simplifying assumptions such as normality; the same is true regarding their asymptotic distribution. In order to estimate/approximate the intractable distribution of our estimators, we propose the simple *wild bootstrap* procedure mentioned in Section 2, i.e., changing the signs of the ith residual with probability $1/2$.

Luckily, it seems that the bias of the above estimators—although not identically zero—is negligible; see the empirical results of Section 4.4. Hence, the main concern is estimating the variance of our estimators. Section 4.3 provides a simulation experiment showing that estimation of variance can be successfully performed via the wild bootstrap.

4 Simulation experiments

To empirically study the performance of the proposed estimators, extensive finite-sample simulations were carried through. The setup for the simulation experiments was the same as in MacKinnon (2012). In particular, the model employed was

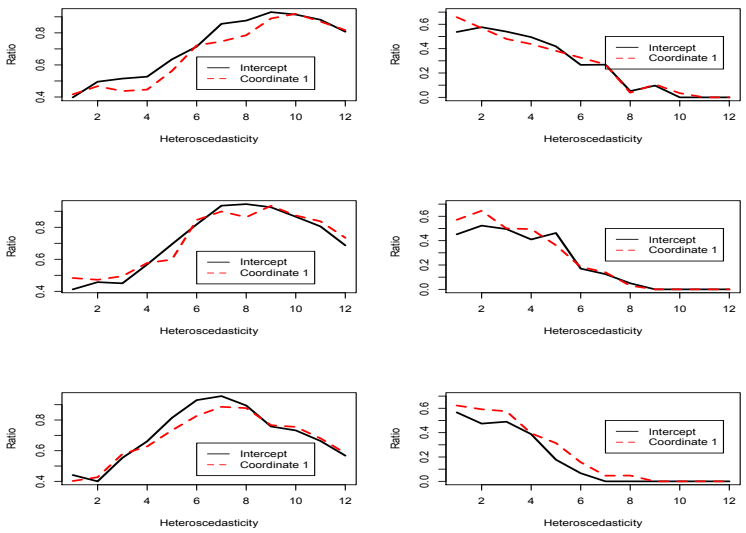

Fig. 1 Ratio of counts when $\lambda = 1$ and when $\lambda = 0$ to the counts of times when the corresponding coefficient was closer to the true parameter in L_2. Plots on the left show the ratio when $\tilde{\beta}_\delta$ was optimal, while plots on the right show the ratio when $\hat{\beta}_{LS}$ was optimal. Each row in the figure corresponds to an experiment.

$$y_i = \beta_1 + \sum_{k=2}^{p} (\beta_k X_{ik}) + u_i, \tag{18}$$

where $X_k \sim$ lognormal(0,1), and $u_i = \sigma_i \varepsilon_i$ where $\sigma_i = |1 + \beta_k X_{ik}|^\eta$ and $\varepsilon_i \sim N(0,1)$. Experiments were conducted for $p = 2$. The heteroscedasticity inducing factor $\eta \in [0,2]$ was chosen in increments of 0.2.

4.1 Choosing the tuning parameter vector λ

In this experiment, we explore the empirical behavior of the tuning parameter vector λ for the estimator of Section 2. We first choose λ to be simply a binary vector, that is, its kth coordinate is either 0 or 1. Here, we simply count the number of *correctly* captured λ's. *Correctly*, means closeness to the true parameter as measured by the squared difference (L_2 distance). Specifically, when $\hat{\beta}_{LS;k}$ is closer to the true parameter in L_2 than what $\tilde{\beta}_{\delta;k}$ is, λ_k should be equal to 0. When the opposite is true, λ_k should be equal to 1. The above strategy is equivalent to a test procedure, in the sense that only the optimal parameter stays in the final model. We carry out the experiment for three cases; Case I, $\beta_{TRUE} = (1, 0.5)'$; Case II, $\beta_{TRUE} = (1,1)'$; Case III, $\beta_{TRUE} = (1, 1.5)'$.

Eq. (18) was used for the generation of the response variable Y. The experiment was carried out with $n = 200$ and it was repeated 199 times while the number of bootstrap samples was 201. For all the experiments, we choose the perturbation on the diagonal of $\hat{\Sigma}$ to be $\delta = 0.001\, S^2$. Results for the first part of the experiment are shown in Figure 1. We report the ratio of counts of times when $\lambda = 1$ and when $\lambda = 0$ to the counts of times when the corresponding estimate was closer to the true parameter in L_2. As expected, as heteroscedasticity increases, $\tilde{\beta}_{\delta;k}$ is closer to the kth coordinate of the true parameter than what $\hat{\beta}_{LS;k}$ is. As it can be seen, the

ratio of counts of times when $\lambda_k = 1$ to the counts of times when $\tilde{\beta}_{\delta;k}$ is optimal, increases as a function of heteroscedasticity. On the other hand, the ratio of counts of times when $\lambda_k = 0$ to the counts of times when $\hat{\beta}_{LS;k}$ is optimal, decreases as a function of heteroscedasticity. This suggested that our procedure for estimating $\tilde{\beta}_{\delta,\lambda}$ is functioning properly.

4.2 Comparing the performance of the new estimators

We now compare the performance of our estimators in terms of their average square difference to the true parameter, i.e., their Mean Squared Error (MSE); our empirical results are summarized in Table 1. For $\tilde{\beta}_{\delta,\lambda}$, the *mixing* parameter vector was the same as above. For $\check{\beta}_{\delta,\lambda}$, the search for the optimal truncation point begun at the 70% quantile of the EDF of the squared OLS residuals. In several cases where heteroscedasticity is mild, $\tilde{\beta}_{\delta,\lambda}$ and $\check{\beta}_{\delta,\lambda}$ outperform $\hat{\beta}_{LS}$ and they are closer to the optimal $\hat{\beta}_{GLS}$. In cases of heavy heteroscedasticity, $\tilde{\beta}_{\delta}$ is always closer to the true parameter.

4.3 Variance estimation via wild bootstrap

As suggested in Remark 3.1, it may be possible to estimate the variance of our estimators via the wild bootstrap. In this simulation, data were generated as before using Case II. The sample sizes were $n = 50, 100$, and 200, while the heteroscedasticity inducing factor was $\eta = 0, 1$, and 2. It is important to note that we modified our data generating model in eq. (18) by changing the parameters of the lognormal distribution from $(0, 1)$ to $(0, .25)$. In our experiments we found that the former parameters generate a heavy tall distribution, therefore possible outliers in the predictor variables could show up. This might cause problems in the variance estimation procedure via the wild bootstrap that are magnified in the presence of heteroscedasticity. By modifying the parameters and controlling for heavy tails, we alleviate this problem. We compare the true variances of our estimators to their bootstrap counterparts. Specifically, we evaluate the true variance of our estimators, denoted by $var\beta$ by a Monte Carlo simulation based on 100 datasets. Secondly, to obtain estimates of $var\beta$ we use the wild bootstrap. Our bootstrapped estimates are denoted by $var\beta^*$ and were averaged over 100 bootstrap samples. Our results our shown in Table 2. As it can be seen, the ratio $\frac{var\beta^*}{var\beta}$, is nearly always, very close to 1. This suggests that the wild bootstrap can be safely used to estimate standard errors. For this experiment, given that our estimators are already computationally intensive, we did not do any fine tuning.

4.4 The bias of the proposed estimators

If the bias of the proposed estimators were appreciable, then we would need to also estimate it via bootstrap, otherwise the standard errors developed in Section 4.3 would be to no avail. Luckily, as hinted at in Remark 3.1, the bias of the proposed estimators appears to be negligible; this implies that a practitioner need not worry about the bias (or estimation thereof) when conducting inference using the new estimators. To empirically validate this claim, we performed the same simulation experiment as in Section 4.3 but with a larger number of Monte Carlo samples (1000) in order to accurately gauge the bias. As can be seen in Table 3, all estimators have a bias that may be considered negligible. Also reported are the inter-quantile range associated with the

1000 replications; they are small enough to ensure that if the true bias is different from zero, it can not be too far away.

5 Concluding remarks

In our attempt to approximate the intractable GLS estimator, we presented several different estimators. The advantage of our estimators lies on the fact that they do not rely on smoothness assumptions regarding the error variance. Our simulation results show that our estimators largely outperform the traditional OLS in the presence of heteroscedasticity of unknown structure. In particular, our preliminary analysis shows that *mixing* different estimators can yield good results. However, a limitation of this approach is computational time, since the optimal grid of the *mixing* parameter is unknown and an extensive search is needed. Nevertheless, the direct estimator $\tilde{\beta}_\delta$ is not computer-intensive, and can always be used since it has been empirically shown to be quite efficient in our simulation. Also, improved optimization techniques can be used for estimating the tuning parameters of both our estimators, i.e. an adaptive grid, as well as parallelization for speed-up.

Further work may include comparisons to techniques that rely on smoothness assumptions, as well as extending these ideas to the time series case where the underlying covariance structure is unknown. Finally, in addition to adapting to heteroscedasticity, any of our procedures can be easily modified to perform model selection by introducing L_2 or L_1 regularizers on our objective function. The latter could be particularly interesting in high-dimensional scenarios, where the rank of the design matrix is less than p and the data are contaminated with outliers. All the above are open directions for future research.

References

1. Chatterjee, S. and Mächler, M. (1997). Robust regression: a Weighted Least Squares approach, *Comm. Statist.—Theory and Methods*, vol. 26, no. 6, pp. 1381–1394.
2. Carol, J. R. (1982). Adapting for heteroscedasticity in linear models, *Annals Statist.*, vol. 10, no. 4, pp. 1224-1233.
3. Davidson, R. and Flachaire, E. (2008). The wild bootstrap, tamed at last, *J. of Econometrics*, vol. 146, 162-169.
4. Liu, R.Y. (1988). Bootstrap procedures under some non-iid models, *Annals Statist.*, 16, 1696-1708.
5. Mammen, E. (1992). *When does bootstrap work? Asymptotic results and simulations*, Springer Lecture Notes in Statistics, Springer, New York.
6. MacKinnon, J.G. (2012). Thirty years of heteroskedasticity-robust inference, in *Recent advances and future directions in causality, prediction, and specification analysis*, (Chen, X. and Swanson, N.R., Eds.), Springer, New York, pp. 437-462.
7. Politis, D.N. (2010), Model-free Model-fitting and Predictive Distributions, Discussion Paper, Department of Economics, Univ. of California—San Diego. Retrievable from: `http://escholarship.org/uc/item/67j6s174`.
8. Politis, D.N. (2013), Model-free Model-fitting and Predictive Distributions, Invited Discussion paper in journal *TEST,* vol. 22, no. 2, 183-221.
9. Le, Q.V., Smola, A.J., and Canu, S. (2005), Heteroscedastic gaussian process regression, *in Proceedings of the* 22*nd International Conference on Machine Learning*, Bonn, Germany.
10. White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica*, vol. 48, 817-838.
11. Wu, C.F.J.(1986). Jackknife, bootstrap and other resampling methods in regression analysis, *Annals Statist.*, 14, 1261-1295.
12. Yuan, M., and Wahba, G. (2004). Doubly penalized likelihood estimator in heteroscedastic regression, *Statist. Probab. Letters*, vol 34, pp. 603-617.

	$\eta=0$					$\eta=0.2$				
MSE	$\breve{\beta}_{\delta,\lambda}$	$\hat{\beta}_{LS}$	$\tilde{\beta}_{\delta}$	$\tilde{\beta}_{\delta,\lambda}$	$\hat{\beta}_{GLS}$	$\breve{\beta}_{\delta,\lambda}$	$\hat{\beta}_{LS}$	$\tilde{\beta}_{\delta}$	$\tilde{\beta}_{\delta,\lambda}$	$\hat{\beta}_{GLS}$
	0.05378	0.04339	0.04570	0.04319	0.04339	0.05417	0.05447	0.05474	0.05399	0.05366
Case I	0.01531	0.00835	0.00899	0.00832	0.00835	0.01334	0.01161	0.01187	0.01148	0.01121
	0.04263	0.04344	0.04323	0.04359	0.04344	0.06832	0.051	0.05021	0.05107	0.04852
Case II	0.00763	0.00761	0.00769	0.00762	0.007615	0.01966	0.01147	0.01204	0.01154	0.010781
	0.05459	0.04348	0.04329	0.04318	0.04348	0.05838	0.05778	0.06082	0.05887	0.05542
Case III	0.01609	0.00788	0.00788	0.00771	0.007888	0.02273	0.01841	0.01931	0.01889	0.01705
	$\eta=0.6$					$\eta=0.8$				
MSE	$\breve{\beta}_{\delta,\lambda}$	$\hat{\beta}_{LS}$	$\tilde{\beta}_{\delta}$	$\tilde{\beta}_{\delta,\lambda}$	$\hat{\beta}_{GLS}$	$\breve{\beta}_{\delta,\lambda}$	$\hat{\beta}_{LS}$	$\tilde{\beta}_{\delta}$	$\tilde{\beta}_{\delta,\lambda}$	$\hat{\beta}_{GLS}$
	0.37866	0.12574	0.12319	0.12078	0.06432	0.18478	0.17819	0.1645	0.17111	0.07202
Case I	0.21391	0.06465	0.05943	0.06233	0.03657	0.10541	0.09371	0.08488	0.08947	0.04991
	0.36578	0.27534	0.24874	0.26025	0.13435	0.75059	0.61153	0.55498	0.57962	0.1176
Case II	0.2253	0.15584	0.15005	0.15085	0.0819	0.49334	0.31246	0.29469	0.29966	0.12589
	0.70688	0.78015	0.73473	0.74693	0.19797	2.29265	2.44659	1.89875	2.21672	0.1991
Case III	0.44353	0.48215	0.44644	0.45875	0.17111	1.79619	1.88462	1.59625	1.81567	0.40225
	$\eta=1.0$					$\eta=1.2$				
MSE	$\breve{\beta}_{\delta,\lambda}$	$\hat{\beta}_{LS}$	$\tilde{\beta}_{\delta}$	$\tilde{\beta}_{\delta,\lambda}$	$\hat{\beta}_{GLS}$	$\breve{\beta}_{\delta,\lambda}$	$\hat{\beta}_{LS}$	$\tilde{\beta}_{\delta}$	$\tilde{\beta}_{\delta,\lambda}$	$\hat{\beta}_{GLS}$
	0.59474	0.60417	0.51433	0.56085	0.10105	9.79127	0.82338	0.73297	0.7738	0.15385
Case I	0.32983	0.30275	0.26862	0.29096	0.08065	12.49075	0.53116	0.48447	0.5065	0.13754
	1.26204	1.36173	1.01231	1.24952	0.14238	4.47709	4.60854	3.35955	4.02106	0.25482
Case II	0.72283	0.75983	0.64193	0.75699	0.16291	3.32433	3.40715	2.67437	3.18252	0.45618
	18.09569	18.2068	9.06757	17.16014	0.40708	22.72211	22.98654	15.51993	21.06294	0.53693
Case III	7.70597	7.76887	5.53547	7.37507	0.76265	14.18275	14.37411	10.60903	12.6128	1.11989
	$\eta=1.8$					$\eta=2.0$				
MSE	$\breve{\beta}_{\delta,\lambda}$	$\hat{\beta}_{LS}$	$\tilde{\beta}_{\delta}$	$\tilde{\beta}_{\delta,\lambda}$	$\hat{\beta}_{GLS}$	$\breve{\beta}_{\delta,\lambda}$	$\hat{\beta}_{LS}$	$\tilde{\beta}_{\delta}$	$\tilde{\beta}_{\delta,\lambda}$	$\hat{\beta}_{GLS}$
	19.833	20.417	5.631	19.372	0.168	26.271	26.914	11.45	25.286	0.196
Case I	12.044	12.364	3.782	11.976	0.248	14.655	15.073	7.516	13.816	0.432
	100.48	100.756	31.418	95.755	0.437	1213.308	1213.636	118.394	1187.726	0.363
Case II	70.156	70.326	24.712	67.256	1.441	520.303	520.536	73.39	506.098	1.259
	5221.765	5222.393	1020.255	5048.798	1.077	10365.66	10366.458	1639.68	9932.69	1.26
Case III	2504.414	2504.788	598.076	2403.024	6.202	4975.22	4975.733	1056.367	4719.818	8.529

Table 1 Comparison of MSE's of various estimators. Highlighted numbers indicate the estimator that had the least MSE in each case.

$\frac{var\beta^*}{var\hat{\beta}}$	$n=50$			$n=100$			$n=200$		
	$\eta=0$	$\eta=1$	$\eta=2$	$\eta=0$	$\eta=1$	$\eta=2$	$\eta=0$	$\eta=1$	$\eta=2$
	0.98452	0.92712	0.90215	0.97599	1.00946	0.98036	1.03372	0.95064	0.97939
$\hat{\beta}_{LS}$	0.94926	0.94358	0.90952	1.00786	1.00354	0.97330	1.02632	0.94079	0.97425
	0.97823	0.93632	0.89858	0.96645	1.00206	0.98862	1.01824	0.93914	0.99269
$\tilde{\beta}_{\delta}$	0.94020	0.94542	0.90791	0.99466	0.99601	0.98258	1.01044	0.92939	0.98661
	0.91100	0.90898	0.88396	0.69562	0.88212	1.15183	1.07295	0.96827	0.94053
$\tilde{\beta}_{\delta,\lambda}$	0.89474	0.88480	0.87706	0.69291	0.89955	1.06307	1.10091	1.01367	0.92281
	0.88959	0.91235	0.89738	0.70893	0.85388	1.17216	1.02872	0.91514	0.94677
$\breve{\beta}_{\delta,\lambda}$	0.87843	0.88626	0.89392	0.70797	0.86385	1.07663	1.02124	0.96083	0.92529

Table 2 Comparison of the true variance to the variance estimated by the wild bootstrap for the various estimators.

$\eta = 0$	$\eta = 1$	$\eta = 2$
$n = 50$		
0.028 (0.024)	-0.031 (0.047)	-0.043 (0.095)
$\hat{\beta}_{LS}$ -0.017 (0.023)	0.018 (0.048)	0.018 (0.103)
0.029 (0.025)	-0.038 (0.046)	-0.031 (0.09)
$\tilde{\beta}_{\delta}$ -0.017 (0.024)	0.026 (0.049)	0.002 (0.101)
0.027 (0.024)	-0.036 (0.047)	-0.038 (0.093)
$\tilde{\beta}_{\delta,\lambda}$ -0.017 (0.023)	0.021 (0.047)	0.009 (0.104)
0.028 (0.024)	-0.031 (0.047)	-0.043 (0.095)
$\check{\beta}_{\delta,\lambda}$ -0.016 (0.024)	0.017 (0.049)	0.017 (0.103)
$n = 100$		
-0.001 (0.018)	0.022 (0.036)	0.016 (0.082)
$\hat{\beta}_{LS}$ -0.003 (0.016)	-0.031 (0.036)	-0.028 (0.087)
0.000 (0.018)	0.014 (0.037)	0.014 (0.082)
$\tilde{\beta}_{\delta}$ -0.004 (0.017)	-0.025 (0.037)	-0.027 (0.086)
0.000 (0.017)	0.018 (0.037)	0.007 (0.082)
$\tilde{\beta}_{\delta,\lambda}$ -0.004 (0.017)	-0.027 (0.036)	-0.021 (0.087)
-0.001 (0.017)	0.021 (0.036)	0.016 (0.082)
$\check{\beta}_{\delta,\lambda}$ -0.002 (0.017)	-0.031 (0.036)	-0.027 (0.087)
$n = 200$		
0.012 (0.012)	0.003 (0.025)	-0.04 (0.055)
$\hat{\beta}_{LS}$ -0.013 (0.011)	-0.005 (0.024)	0.032 (0.056)
0.013 (0.012)	0.000 (0.025)	-0.039 (0.054)
$\tilde{\beta}_{\delta}$ -0.014 (0.011)	-0.002 (0.024)	0.03 (0.055)
0.012 (0.012)	0.001 (0.025)	-0.038 (0.054)
$\tilde{\beta}_{\delta,\lambda}$ -0.013 (0.011)	-0.003 (0.024)	0.031 (0.055)
0.012 (0.012)	0.003 (0.025)	-0.039 (0.055)
$\check{\beta}_{\delta,\lambda}$ -0.014 (0.011)	-0.005 (0.024)	0.032 (0.056)

Table 3 Bias of the different estimators with inter-quantile ranges divided by $\sqrt{1000}$ (number of Monte Carlo Samples) in parentheses. As it can be seen in most cases the bias is negligible.