DiCiccio, Cyrus J.; Romano, Joseph P.; Wolf, Michael

Working Paper

Improving weighted least squares inference

Working Paper, No. 232

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

Suggested Citation: DiCiccio, Cyrus J.; Romano, Joseph P.; Wolf, Michael (2016) : Improving weighted least squares inference, Working Paper, No. 232, University of Zurich, Department of Economics, Zurich,
https://doi.org/10.5167/uzh-125468

This Version is available at:
https://hdl.handle.net/10419/162435

University of Zurich

Department of Economics

Working Paper Series

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

Working Paper No. 232

Improving Weighted Least Squares Inference

Cyrus J. DiCiccio, Joseph P. Romano and Michael Wolf

August 2016

Improving Weighted Least Squares Inference*

Cyrus J. DiCiccio
Department of Statistics
Stanford University
cyrusd@stanford.edu

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

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

August 2016

Abstract

In the presence of conditional heteroskedasticity, inference about the coefficients in a linear regression model these days is typically based on the ordinary least squares estimator in conjunction with using heteroskedasticity consistent standard errors. Similarly, even when the true form of heteroskedasticity is unknown, heteroskedasticity consistent standard errors can be used to base valid inference on a weighted least squares estimator. Using a weighted least squares estimator can provide large gains in efficiency over the ordinary least squares estimator. However, intervals based on plug-in standard errors often have coverage that is below the nominal level, especially for small sample sizes. In this paper, it is shown that a bootstrap approximation to the sampling distribution of the weighted least squares estimate is valid, which allows for inference with improved finite-sample properties. Furthermore, when the model used to estimate the unknown form of the heteroskedasticity is misspecified, the weighted least squares estimator may be less efficient than the ordinary least squares estimator. To address this problem, a new estimator is proposed that is asymptotically at least as efficient as both the ordinary and the weighted least squares estimator. Simulation studies demonstrate the attractive finite-sample properties of this new estimator as well as the improvements in performance realized by bootstrap confidence intervals.

KEY WORDS: Bootstrap, conditional heteroskedasticity, HC standard errors.

JEL classification codes: C12, C13, C21.

*Research of the first two authors supported by NSF grant DMS-1307973.

1 Introduction

In this paper, we consider the problem of inference in a linear regression model. Under conditional homoskedasticity, the ordinary least squares (OLS) estimator is the best linear unbiased estimator. Traditional inference based upon the ordinary least squares estimator, such as the F test or t confidence intervals for individual coefficients relies on estimators of asymptotic variance that are only consistent when the model is homoskedastic. In many applications, the assumption of homoskedasticity is unrealistic. When instead the model exhibits conditional heteroskedasticity, traditional inference based on the ordinary least squares estimator may fail to be (asymptotically) valid.

If the skedastic function is known (that is, the function that determines the conditional heteroskedasticty of the error term given the values of the regressors), the best linear unbiased estimator (BLUE) is obtained by computing the ordinary least squares estimator after weighting the data by the inverse of square root of the value of the skedastic function. Unfortunately, in all but the most ideal examples, the heteroskedasticity is of unknown form, and this estimator cannot be used. However, if the skedastic function can be estimated, then weighting the model by the inverse square root of the estimate of the skedastic function produces a feasible weighted least squares (WLS) estimator. Although this estimator is no longer unbiased, it can often give improvements in efficiency over the weighted least squares estimator. Even so, estimating the skedastic function is often challenging, and a poorly estimated skedastic function may produce an estimator that is less efficient than the ordinary least squares estimator. Furthermore, when the estimated skedastic function is not consistent, traditional inference based on the weighted least squares estimator may not be valid. Because of these difficulties the weighted least squares estimator has largely fallen out of favor with practitioners.

As an alternative, White (1980) developed heteroskedasticity consistent (HC) standard errors which allow for asymptotically valid inference, based on the ordinary least squares estimator, in the presence of conditional heteroskedasticity of unknown form. Although this approach abandons any efficiency gains that could be achieved from weighting, the standard errors are consistent under minimal model assumptions.

Simulation studies, such as MacKinnon and White (1985) who investigated the performance of several different heteroskedasticity consistent standard errors, show that inference based on normal or even t approximations can be misleading in small samples. In such cases, it is useful to consider bootstrap methods.

Following the proposal of White's heteroskedasticity consistent covariance estimators, resampling methods have been developed that give valid inference based on the ordinary least squares estimator. Freedman (1981) proposed the pairs bootstrap which resamples pairs of predictor and response variables from the original data. Another popular technique is the wild bootstrap which was suggested by Wu (1986). This method generates bootstrap samples by simulating error terms whose

variance are an estimate of the conditional variance for each predictor variable. Recent numerical work comparing the pairs bootstrap and the wild bootstrap to asymptotic approximations is given in Flachaire (2005) and Cribari-Neto (2004). Godfrey and Orne (2004) conducted simulations suggesting that combining heteroskedasticity consistent standard errors with the wild bootstrap produces tests that are more reliable in small samples than using the normal approximation. Despite the improvements that the resampling methods produce over asymptotic approximations, inference based on the ordinary least squares estimator may still not be as efficient as weighted least squares.

Neither the solution of using heteroscedasticity consistent covariance estimates, nor using weighted least squares with traditional inference seem entirely satisfactory. Even recently there has been debate about the merits of weighting. Angrist and Pischke (2010) are of the belief that any potential efficiency gains from using a weighted least squares estimator are not substantial enough to risk the harm that could be done by poorly estimated weights. On the other hand, Leamer (2010) contends that researchers should be working to model the heteroskedasticity in order to determine whether sensible reweighting changes estimates or confidence intervals.

Even in examples where the estimated skedastic function is not consistent for the true skedastic function, the weighted least squares estimator can be more efficient than the ordinary least squares estimator. Arguably, a more satisfying approach to inference than simply abandoning weighting is to base inference on the weighted least squares estimator in conjunction HC errors. This proposal goes back to at least Wooldridge (2012) and was made rigorous in Romano and Wolf (2015). Regardless of whether or not the parametric family used to estimate the skedastic function is correctly specified or not, the weighted least squares estimator has an asymptotically normal distribution with mean zero and a variance that can be estimated consistently estimated by the means of HC standard errors (as long as some technical conditions are satisfied).

There are two difficulties with basing inference on these consistent standard errors. As is the case with using White's standard errors, using heteroskedasticity consistent standard errors with the weighted least squares estimator leads to inference that can be misleading in small samples. This problem is even more severe with the weighted estimator than with the ordinary least squares estimator because the plug-in standard errors use the estimated skedastic function, and are the same estimators that would be used if it had been known *a priori* that the model would be weighted by this particular estimated skedastic function. Confidence intervals, for example, do not account for the randomness in estimating the skedastic function and for this reason tend to have coverage that is below the nominal level, especially in small samples.

The other trouble is that inference based on the weighted least squares estimator using consistent standard errors may not be particularly efficient, and investing effort in modeling the conditional variance may be counterproductive. In fact, when the family of skedastic functions is misspecified (or the estimated skedastic function is not consistent for the true skedastic function), the weighted least squares estimator can be less efficient than the ordinary least squares estimator, even when

conditional heteroskedasticity is present. Although this possibility seems rare, it is theoretically unsatisfying and has been given as a reason to abandon the approach altogether.

In this paper, we will address these limitations of the weighted least squares estimator. Thus, the general goal is to improve the methodology in Romano and Wolf (2015) by constructing methods with improved accuracy and efficiency. In particular, we show that the bootstrap approximation to the sampling distribution of the weighted least squares estimator is consistent and we provide numerical evidence that using the bootstrap leads to more reliable inference. We also propose a new estimator that is a convex combination of the ordinary least squares estimator and the weighted least squares estimator and is at least as efficient (asymptotically) as both the weighted and the ordinary least squares estimator.

Model assumptions are given in Section 2. Consistency of both the pairs and wild bootstrap approximations to the distribution of the weighted least squares estimator is given in Section 3; notably, the bootstrap accounts for estimation of the skedastic function as it is re-estimated in each bootstrap sample. Tests for linear constraints of the coefficient vector using both bootstrap methods, as well as a randomization test, are given in Section 4. Estimators based on a convex combination of the ordinary and weighted least squares estimators that are asymptotically no worse, but potentially more efficient than the ordinary least squares estimator, as well as the consistency of the bootstrap distribution of these estimators, are given in Section 5. Here, the bootstrap is useful not only to account for the randomness in the skedastic function but also the randomness in the convex weights. Section 6 provides an example where the convex combination estimator is strictly more efficient than either the ordinary or weighted least squares estimators. Simulations to examine finite-sample performance are provided in Section 7. Proofs are given in the appendix.

2 Model and Notation

Throughout the paper, we will be concerned with the heteroskedastic linear regression model specified by the following assumptions.

(A1) The model can be written

$$y_i = x_i^\top \beta + \varepsilon_i \ ,$$

$i = 1, ..., n$, where $x_i \in \mathbb{R}^p$ is a vector of predictor variables, and ε_i is an unobservable error term with properties specified below.

(A2) $\{(y_i, x_i)\}$ are independent and identically distributed (i.i.d.) according to a distribution P.

(A3) The error terms have conditional mean zero given the predictor variables:

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

(A4) $\Sigma_{xx} := \mathbb{E}(x_i x_i^\top)$ is nonsingular and $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ is almost surely invertible.

(A5) $\Omega \coloneqq \mathbb{E}(\varepsilon_i^2 x_i x_i^\top)$ is nonsingular.

(A6) There exists a function $v(\cdot)$, called the skedastic function, such that

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

It is also convenient to write the linear model specified by assumption (A1) in vector-matrix notation.

$$Y = X\beta + \varepsilon$$

where

$$Y \coloneqq \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \ , \quad \varepsilon \coloneqq \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \ , \quad \text{and} \quad X \coloneqq \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \dots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix} \ .$$

Finally, following the notation of Romano and Wolf (2015), define

$$\Omega_{a/b} \coloneqq \mathbb{E}\left(x_i^\top x_i \frac{a(x_i)}{b(x_i)} \right)$$

for any functions $a, b : \mathbb{R}^p \to \mathbb{R}$. Using this convention, $\Sigma_{xx} = \Omega_{1/1}$ and $\Omega = \Omega_{v/1}$.

3 Estimators

Under the model assumptions given in Section 2, it is common to use the ordinary least squares (OLS) estimator

$$\hat{\beta}_{\text{OLS}} \coloneqq \left(X^\top X \right)^{-1} X^\top Y$$

to estimate β. Although this estimator is unbiased, it is not efficient when the model is not conditionally homoskedastic. Ideally, one would use the best linear unbiased estimator (BLUE) which is obtained by regressing $y_i / \sqrt{v(x_i)}$ on $x_i / \sqrt{v(x_i)}$ by OLS. But this estimator requires knowledge of the true skedastic function and thus is not feasible in most applications.

Instead, one can estimate the skedastic function and weight the observations by the estimate of the skedastic function. Typically, the skedastic function is estimated by $v_{\hat{\theta}}(\cdot)$, a member of a parametric family $\left\{ v_\theta(\cdot) : \theta \in \mathbb{R}^d \right\}$ of skedastic functions. For instance, a popular choice for the family of skedastic functions is

$$v_\theta(x_i) \coloneqq \exp\big(\theta_0 + \gamma_2 \log|x_{i,1}| + \ldots + \theta_p \log|x_{i,p}|\big) \ , \quad \text{with} \quad \theta \coloneqq (\theta_0, \theta_1, \ldots, \theta_p) \in \mathbb{R}^{p+1} \ . \tag{3.1}$$

The weighted least squares (WLS) estimator based on the estimated skedastic function is obtained by regressing $y_i / \sqrt{v_{\hat{\theta}}(x_i)}$ on $x_i / \sqrt{v_{\hat{\theta}}(x_i)}$ by OLS and thus given by

$$\hat{\beta}_{\text{WLS}} \coloneqq (X^\top V_{\hat{\theta}}^{-1} X)^{-1} X^\top V_{\hat{\theta}}^{-1} Y$$

where $V_\theta := \text{diag}\{v_\theta(x_1), ..., v_\theta(x_n)\}$.

Provided the estimated skedastic function $v_{\hat{\theta}}(\cdot)$ is suitably close to some limiting estimated skedastic function, say $v_{\theta_0}(\cdot)$ for n large, then the weighted least squares estimator has an asymptotically normal distribution. Note that $v_{\theta_0}(\cdot)$ need not correspond to the true skedastic function, which of course happens if the family of skedastic functions is not well specified. Romano and Wolf (2015) assume that $\hat{\theta}$ is a consistent estimator of some θ_0 in the sense that

$$n^{1/4}(\hat{\theta} - \theta_0) \xrightarrow{P} 0 \ , \tag{3.2}$$

where $\xrightarrow{P}$ denote convergence in probability. They also assume that at this θ_0, $1/v_\theta(\cdot)$ is differentiable in the sense that there exists a d-dimensional vector-valued function

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

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

$$\left| \frac{1}{v_\theta(x)} - \frac{1}{v_{\theta_0}(x)} - r_{\theta_0}(x)(\theta - \theta_0) \right| \leq \frac{1}{2}|\theta - \theta_0|^2 s_{\theta_0}(x) \ , \tag{3.3}$$

for all θ in some small open ball around θ_0 and all x.

If (3.2) and (3.3) are satisfied, then under some further regularity conditions,

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

where $w(\cdot) := v_{\theta_0}(\cdot)$ and $\xrightarrow{d}$ denotes convergence in distribution.

The matrices $\Omega_{1/w}$ and Ω_{v/w^2} appearing in the asymptotic variance can be consistently estimated by

$$\hat{\Omega}_{1/w} := \frac{X' V_{\hat{\theta}}^{-1} X}{n} \ ,$$

and

$$\hat{\Omega}_{v/w^2} := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\tilde{\varepsilon}_i^2}{v_{\hat{\theta}}^2(x_i)} \cdot x_i x_i^\top\right)$$

respectively, for suitable residuals $\tilde{\varepsilon}$ that are consistent for the true error terms ε. Then the asymptotic variance of the weighted least squares estimator, denoted by $\text{Avar}(\hat{\beta}_{\text{WLS}})$, can be consistently estimated by

$$\widehat{\text{Avar}}\,(\beta_{\text{WLS}}) = \hat{\Omega}_{1/w}^{-1}\hat{\Omega}_{v/w^2}\hat{\Omega}_{1/w}^{-1} \ . \tag{3.4}$$

Remark 3.1. When the 'raw' OLS residuals, $\hat{\varepsilon}_i := y_i - x_i\hat{\beta}_{\text{OLS}}$, are used, the estimator (3.4) is commonly referred to as the HC0 estimator. To improve finite-sample performance other variants of HC used scaled residuals instead. The HC1 estimator scales the OLS residuals by $\sqrt{n/(n-p)}$, which reduces bias. When the errors are homoskedastic, the variance of the OLS residual $\hat{\varepsilon}_i$ is proportional to $1/(1-h_i)$, where h_i is the i^{th} diagonal entry of the 'hat' matrix $H = X(X^\top X)^{-1}X^\top$. The HC2 estimator uses the OLS residuals scaled by $1/\sqrt{(1-h_i)}$. The HC3 estimator uses the OLS residuals scaled by $1/(1-h_i)$. ■

Using this plug-in estimator of asymptotic variance gives t confidence intervals for the coefficients having the form

$$\hat{\beta}_{\mathrm{WLS},k} \pm t_{n-p,1-\alpha/2} \cdot \mathrm{SE}(\hat{\beta}_{\mathrm{WLS},k})$$

where

$$\mathrm{SE}(\hat{\beta}_{\mathrm{WLS},k}) := \sqrt{\widehat{\mathrm{Avar}}(\hat{\beta}_{\mathrm{WLS},k})/n} \ ,$$

and $t_{n-p,1-\alpha/2}$ is the $1-\alpha/2$ quantile of the t-distribution with $n-p$ degrees of freedom. These intervals are asymptotically valid; however, simulations suggest that the true coverage rates are often smaller than the nominal level, especially in small samples. The standard errors for these confidence intervals are the same standard errors that would be used if we had known before observing any data that the model would be weighted by $1/\sqrt{v_{\hat{\theta}}(\cdot)}$ and the intervals do not account for variability in the estimation of the skedastic function. The coverage can be improved by reporting intervals based on the "pairs" bootstrap confidence intervals where the skedastic function is estimated on each bootstrap sample separately.

The empirical distribution of a sample $(x_1, y_1), ..., (x_n, y_n)$ is

$$\hat{P}_n(s,t) := \frac{1}{n}\sum_{i=1}^{n} I\left\{x_i \le s, y_i \le t\right\} \ .$$

The pairs bootstrap, which is commonly used for heteroskedastic regression models, generates bootstrap samples, $(x_1^*, y_1^*), ..., (x_n^*, y_n^*)$ from $\hat{P}_n$. Alternatively, one could generate bootstrap samples $(x_1, y_1^*), ..., (x_n, y_n^*)$ using the wild bootstrap which simulates new response variables

$$y_i^* := x_i\hat{\beta}_{\mathrm{WLS}} + \varepsilon_i^*$$

where ε_i^* are sampled from any distribution with mean zero and variance $\hat{\varepsilon}_i^2$. It is common to use $\varepsilon_i^* := u_i \cdot \hat{\varepsilon}_i$ where u_i is a random variable taking values ± 1, each with probability $1/2$.

When computing the weighted least squares estimator $\hat{\beta}_{\mathrm{WLS}}$, the parameter for the estimated skedastic function is re-estimated on the bootstrap sample by $\hat{\theta}^*$. The following theorem establishes that the distribution of $\sqrt{n}\big(\hat{\beta}^*_{\mathrm{WLS}} - \hat{\beta}_{\mathrm{WLS}}\big)$, using the pairs or the wild bootstrap, is a consistent approximation of the sampling distribution of $\sqrt{n}\big(\hat{\beta}_{\mathrm{WLS}} - \beta\big)$.

Theorem 3.1. *Suppose that $(x_1, y_1), ..., (x_n, y_n)$ are i.i.d. satisfying assumptions $(A1)$–$(A6)$ above, and that $\left\{v_\theta(\cdot) : \theta \in \mathbb{R}^d\right\}$ is a family of continuous skedastic functions satisfying (3.3) at some θ_0 with $r(\cdot)$ and $s(\cdot)$ such that*

$$\mathbb{E}\left|x_1 y_1 r(x_1)\right|^2 < \infty \quad \textit{and} \quad \mathbb{E}\left|x_1 y_1 s(x_1)\right|^2 < \infty \ .$$

Let $\hat{\theta}$ be an estimator satisfying (3.2). Further suppose that $n^{1/4}\big(\hat{\theta}^* - \hat{\theta}_0\big)$ converges to zero in conditional probability. Let $\hat{\beta}_{WLS} := (X^\top V_{\hat{\theta}}^{-1} X)^{-1} X^\top V_{\hat{\theta}}^{-1} Y$ and $v_{\theta_0} =: w$ so that $W = diag(v_{\theta_0}(x_1), ..., v_{\theta_0}(x_n))$. If

$$\mathbb{E}\left(\left\|\left(\frac{x_{i1}}{\sqrt{w(x_i)}}, ..., \frac{x_{ip}}{\sqrt{w(x_i)}}, \frac{y_i}{\sqrt{w(x_i)}}\right)\right\|_2^4\right) < \infty \ ,$$

where $\|\cdot\|_2$ *is the Euclidean norm, then the conditional law of* $\sqrt{n}\big(\hat{\beta}^*_{WLS} - \hat{\beta}_{WLS}\big)$*, based on a pairs bootstrap sample or a wild bootstrap sample, converges weakly to the multivariate normal distribution with mean zero and covariance matrix* $\Omega^{-1}_{1/w}\Omega_{v/w^2}\Omega^{-1}_{1/w}$ *in probability.*

Remark 3.2. Of course, the bootstrap distribution is random and hence its weak convergence properties hold in a probabilistic sense. As is customary, when we say that a sequence of random distributions, say $\hat{G}_n$ converges weakly to G in probability, we mean that $\rho(\hat{G}_n, G) \xrightarrow{P} 0$ where ρ is any metric metrizing weak convergence. We also say that a sequence $T_n(X, Y)$ converges in conditional probability to zero almost surely if for almost every sequence $\{x_i, y_i\}$, $T_n(X^*, Y^*) \to 0$ in $\hat{P}_n$ probability. ■

In Theorem 3.1, it was assumed that we have a family of skedastic functions $\{v_\theta(\cdot)\}$, and an estimator of θ, say $\hat{\theta}$, such that $n^{1/4}\big(\hat{\theta}^* - \theta_0\big)$ converges in conditional probability to zero. We will now verify this assumption for a flexible family of skedastic functions which includes the family specified in (3.1).

Lemma 3.1. *For any functions* $g_i : \mathbb{R}^d \to \mathbb{R}^d$*,* $i = 1, ..., d$*, define the family* $\big\{v_\theta : \theta \in \mathbb{R}^d\big\}$ *by*

$$v_\theta(x) := \exp\left[\sum_{i=1}^{d} \theta_j g_j(x)\right] ,$$

and let $\hat{\theta}$ *be the estimator obtained by regressing* $h_\delta(\hat{\varepsilon}_i) := \log\left(\max\left\{\delta^2, \hat{\varepsilon}_i^2\right\}\right)$ *on* $g(x_i) = \big(g_1(x_i), ..., g_d(x_i)$ *by OLS, where* $\delta > 0$ *is a small constant. Then,* $n^{1/4}\big(\hat{\theta}^* - \theta_0\big)$ *converges in conditional probability to zero for*

$$\theta_0 := E(g(x_i)g(x_i)')E(g(x_i)h_\delta(\varepsilon_i))$$

provided $E(g_j(x_i)g_k(x_i))^{4/3}$ *and* $E(g_j(x_i)h_\delta(\varepsilon_i))^{4/3}$ *are both finite for each* j *and* k*.*

4 Hypothesis Testing

Just as using a t approximation often produces confidence intervals with coverage below the nominal confidence level, especially for small samples using an F approximation to conduct F tests of linear constraints often gives rejection probabilities that are above the nominal significance level, especially for small samples. And as with confidence intervals, using the bootstrap can produce tests that have rejection probabilities that are closer to the nominal level. Consider the hypothesis

$$H_0 : R(\beta) = q$$

where R is a $J \times p$ matrix of full rank (with $J \leq p$) and q is a vector of length J. Two appropriate test statistics for this hypothesis are the Wald statistic

$$W_n(X, Y) := n \cdot \left(R\hat{\beta}_{\text{WLS}} - q\right)^\top \left[R\hat{\Omega}^{-1}_{1/w}\hat{\Omega}_{v/w^2}\hat{\Omega}^{-1}_{1/w}R^\top\right]^{-1} \left(R\hat{\beta}_{\text{WLS}} - q\right) , \tag{4.1}$$

and the maximum statistic,

$$M_n(X,Y) := \max_{1\leq k\leq p}\left\{\frac{\left([R\hat{\beta}_{\text{WLS}}]_k - q_k\right)}{\left[R\widehat{\Omega}_{1/w}^{-1}\hat{\Omega}_{v/w^2}\hat{\Omega}_{1/w}^{-1}R^\top\right]_{k,k}}\right\}. \tag{4.2}$$

It follows immediately from the results of Romano and Wolf (2015) that, under the null, the sampling distribution of $W_n(X,Y)$ is asymptotically chi-squared with J degrees of freedom and the sampling distribution of $M_n(X,Y)$ is asymptotically distributed as the maximum of k correlated standard normal variables. Let $G_n(x,P)$ denote the sampling distribution of W_n when (X_1,Y_1) are distributed according to P.

Define $c_n(1-\alpha,\hat{P})$ to be the $1-\alpha$ quantile of the distribution of

$$\left(R\left(\hat{\beta}^*_{\text{WLS}} - \hat{\beta}_{\text{WLS}}\right)\right)^\top \left[R\hat{\Omega}^{*-1}_{1/w}\hat{\Omega}^*_{v/w^2}\hat{\Omega}^{*-1}_{1/w}R^\top\right]^{-1}\left(R\left(\hat{\beta}^*_{\text{WLS}} - \hat{\beta}_{\text{WLS}}\right)\right)$$

and $d_n(1-\alpha,\hat{P})$ to be the $1-\alpha$ quantile of the distribution of

$$\max_{1\leq k\leq p}\left\{\frac{\left([R\hat{\beta}^*_{\text{WLS}}]_k - [R\hat{\beta}_{\text{WLS}}]_k\right)}{\left[R\hat{\Omega}^{*-1}_{1/w}\hat{\Omega}^*_{v/w^2}\hat{\Omega}^{*-1}_{1/w}R^\top\right]_{k,k}}\right\}$$

using the pairs or wild bootstrap.

Theorem 4.1. *Suppose that $(x_1,y_1),...,(x_n,y_n)$ are i.i.d. according to a distribution P such that $R\beta = q$. Then, under the assumptions of Theorem 3.1,*

$$P\left(W_n(X,Y) > c_n(1-\alpha,\hat{P}_n)\right) \to \alpha$$

as $n\to\infty$. That is, the bootstrap quantiles of the Wald statistic converge to the corresponding quantiles of a chi-squared distribution with J degrees of freedom when $R\beta = q$. Similarly,

$$P\left(M_n(X,Y) > d_n(1-\alpha,\hat{P}_n)\right) \to \alpha$$

as $n\to\infty$.

We point out that hypothesis testing using the wild bootstrap is closely related to a commonly used randomization test under symmetry assumptions.

Suppose that the ε_i follow a symmetric distribution conditional on X_i in the sense that the distribution of ε_i given X_i is the same as the distribution of $-\varepsilon_i$ given X_i. Then under $H:\beta=0$, the joint distribution of the (X_i,Y_i) is invariant under the group of transformations $\mathbf{G}_n := \{g_\delta : \delta\in\{1,-1\}^n\}$ such that $g_\delta((x_1,y_1),...,(x_n,y_n)) = ((x_1,\delta_1 y_1),...,(x_n,\delta_n y_n))$ for any $x,y\in\mathbb{R}^n$. Given a test statistic T_n used to test the hypothesis $H:\beta=0$, the permutation test rejects if $T_n(X,Y)$ exceeds the appropriate quantiles of the permutation distribution of T_n, which is given by

$$\hat{R}^{T_n}_n(t) := \frac{1}{2^n}\sum_{g_\delta\in\mathbf{G}_n} I\left\{U_n(X,g_\delta(Y)) \leq t\right\}$$

For any choice of test statistic, the invariance of the distribution of the data under the group of transformations is sufficient to ensure that the randomization test is exact; see Lehmann and Romano (2005, Chapter 15) for details.

Typically for regression problems, the test statistic is chosen to be the usual F-statistic in homoskedastic models, or the Wald statistic in heteroskedastic models. While under the symmetry assumption this test is exact in either setting, Janssen (1999) shows that this test is robust against violations of the symmetry assumptions (in the sense that the test is still asymptotically valid when the distribution of the Y_i is not symmetric).

When the symmetry assumption is satisfied, the randomization test using W_n or M_n — as defined in equations (4.1) and (4.2), respectively — as the test statistic is also exact. Even when this assumption is not satisfied, the test is still asymptotically valid, as the following theorem demonstrates.

Theorem 4.2. *Suppose that $(x_1, y_1), ..., (x_n, y_n)$ are i.i.d. according to a distribution P such that $\beta = 0$. Suppose that $\sqrt{n}(\hat{\theta}(g_\delta(X, Y)) - \theta_0)$ converges in probability to zero conditionally on the X's and Y's for any uniformly randomly chosen $g_\delta \in \mathbb{G}_n$. Then, under the assumptions of Theorem 3.1, the permutation distribution $\hat{R}_n^{W_n}$ of W_n satisfies*

$$\sup_{t \in \mathbb{R}} \left| \hat{R}_n^{W_n}(t) - J_n^{W_n}(t, P) \right| \to 0$$

in probability as $n \to \infty$ where $J_n^{W_n}(\cdot, P)$ is the sampling distribution of W_n under P. Similarly, the permutation distribution $\hat{R}_n^{M_n}$ of M_n satisfies

$$\sup_{t \in \mathbb{R}} \left| \hat{R}_n^{M_n}(t) - J_n^{M_n}(t, P) \right| \to 0$$

in probability as $n \to \infty$ where $J_n^{M_n}(\cdot, P)$ is the sampling distribution of M_n under P.

Once again, this theorem makes assumptions about the consistency of the estimate of the parameter in the skedastic function. We verify this assumption for a particular family of skedastic functions.

Lemma 4.1. *For any functions $g_i : \mathbb{R}^d \to \mathbb{R}^d$, $i = 1, ..., d$, define the family $\left\{ v_\theta : \theta \in \mathbb{R}^d \right\}$ by*

$$v_\theta(x) := \exp\left[\sum_{i=1}^{d} \theta_j g_j(x) \right] ,$$

and let $\hat{\theta}$ be the estimator obtained by regressing $h_\delta(\hat{\varepsilon}_i) := \log\left(\max\left\{ \delta^2, \hat{\varepsilon}_i^2 \right\} \right)$ on $g(x_i) = \left(g_1(x_i), ..., g_d(x_i) \right)$ by OLS, where $\delta > 0$ is a small constant. Then, for any randomly and uniformly chosen $g_\delta \in \mathbb{G}_n$, $n^{1/4}\left(\hat{\theta}(g_\delta(X, Y)) - \theta_0 \right)$ converges in conditional probability to zero for

$$\theta_0 := E(g(x_i) g(x_i)') E(g(x_i) h_\delta(\varepsilon_i))$$

provided $E(g_j(x_i) g_k(x_i))^{4/3}$ and $E(g_j(x_i) h_\delta(\varepsilon_i))^{4/3}$ are both finite for each j and k.

5 A convex linear combination of the ordinary and weighted least squares estimators

When the family of skedastic functions is misspecified, the weighted least squares estimator can be less efficient than the ordinary least squares estimator, even asymptotically.

When interested in inference for a particular coefficient, say β_k, practitioners might be tempted to decide between the ordinary and weighted least squares estimators based on which estimator has the smaller standard error In particular, it might be tempting to report the estimator

$$\hat{\beta}_{\text{MIN},k} \coloneqq \begin{cases} \hat{\beta}_{\text{WLS},k} & \text{if} \quad \widehat{\text{Avar}}(\hat{\beta}_{\text{OLS},k}) > \widehat{\text{Avar}}(\hat{\beta}_{\text{WLS},k}) \\ \hat{\beta}_{\text{OLS},k} & \text{if} \quad \widehat{\text{Avar}}(\hat{\beta}_{\text{OLS},k}) \leq \widehat{\text{Avar}}(\hat{\beta}_{\text{WLS},k}) \end{cases},$$

along with the corresponding confidence interval

$$\hat{\beta}_{\text{MIN},k} \pm t_{n-p,1-\alpha/2} \cdot \sqrt{\frac{1}{n} \min\left\{\widehat{\text{Avar}}(\hat{\beta}_{\text{WLS},k}), \widehat{\text{Avar}}(\hat{\beta}_{\text{OLS},k})\right\}}. \tag{5.1}$$

Asymptotically, this estimator has the same efficiency as the better of the ordinary least squares and weighted estimators. However, the confidence interval (5.1) tends to undercover in finite samples due to the minimizing over the standard error. The next theorem established consistency of the bootstrap distribution, which can be used to produce confidence intervals with better finite-sample coverage than those given by (5.1).

Theorem 5.1. *Under the conditions of Theorem 3.1, the sampling distribution of* $\sqrt{n}\big(\hat{\beta}_{MIN,k} - \beta_k\big)$ *converges weakly to the normal distribution with mean zero and variance*

$$\sigma^2_{MIN} \coloneqq \min\Big\{Avar(\hat{\beta}_{WLS,k}), Avar(\hat{\beta}_{OLS,k})\Big\}$$

The distribution of $\sqrt{n}\big(\hat{\beta}^*_{MIN,k} - \beta_k\big)$*, where the samples* (x^*_i, y^*_i) *are generated according to the pairs bootstrap or the wild bootstrap, converges weakly to the normal distribution having mean zero and variance* σ^2_{MIN} *in probability.*

When the estimated skedastic function is consistent for the true skedastic function, the estimator $\hat{\beta}_{\text{MIN},k}$ is asymptotically as efficient as the best linear unbiased estimator. On the other hand, when the skedastic function is misspecified, one can find an estimator which is at least as efficient as $\hat{\beta}_{\text{MIN}}$, regardless of whether the skedastic function is well modeled, but can potentially have smaller asymptotic variance. With the aim of creating such an estimator, consider estimators of the form

$$\hat{\beta}_\lambda \coloneqq \lambda\hat{\beta}_{\text{OLS}} + (1-\lambda)\hat{\beta}_{\text{WLS}} \tag{5.2}$$

for $\lambda \in [0,1]$, which are convex combinations of the ordinary and weighted least squares estimators. To study the asymptotic behavior of these estimators, it is helpful to first find the asymptotic joint distribution of the ordinary and weighted least squares estimators.

Theorem 5.2. *Under the assumptions of Theorem 3.1,*

$$\sqrt{n}\left(\begin{pmatrix}\hat{\beta}_{WLS}\\ \hat{\beta}_{OLS}\end{pmatrix}-\begin{pmatrix}\beta\\ \beta\end{pmatrix}\right)\xrightarrow{d} N\left(\begin{pmatrix}0\\ 0\end{pmatrix},\begin{pmatrix}\Omega_{1/w}^{-1}\Omega_{v/w^2}\Omega_{1/w}^{-1} & \Omega_{1/w}^{-1}\Omega_{v/w}\Omega_{1/1}^{-1}\\ \Omega_{1/1}^{-1}\Omega_{v/w}\Omega_{1/w}^{-1} & \Omega_{1/1}^{-1}\Omega_{v/1}\Omega_{1/1}^{-1}\end{pmatrix}\right)$$

as $n\to\infty$.

It follows that for any $\lambda\in[0,1]$, $\sqrt{n}\big(\hat{\beta}_\lambda-\beta\big)$ asymptotically has a normal distribution with mean zero and covariance matrix

$$\mathrm{Avar}(\hat{\beta}_\lambda):=\lambda^2\Omega_{1/w}^{-1}\Omega_{v/w^2}\Omega_{1/w}^{-1}+2\lambda(1-\lambda)\Omega_{1/w}^{-1}\Omega_{v/w}\Omega_{1/1}^{-1}+(1-\lambda)^2\Omega_{1/1}^{-1}\Omega_{v/1}\Omega_{1/1}^{-1}\ ,$$

which can be consistently estimated by

$$\widehat{\mathrm{Avar}}(\hat{\beta}_\lambda):=\left[\lambda^2\hat{\Omega}_{1/w}^{-1}\hat{\Omega}_{v/w^2}\hat{\Omega}_{1/w}^{-1}+2\lambda(1-\lambda)\hat{\Omega}_{1/w}^{-1}\hat{\Omega}_{v/w}\hat{\Omega}_{1/1}^{-1}+(1-\lambda)^2\hat{\Omega}_{1/1}^{-1}\hat{\Omega}_{v/1}\hat{\Omega}_{1/1}^{-1}\right]\ .$$

For any particular coefficient β_k, it then holds that $\sqrt{n}\big(\hat{\beta}_{\lambda,k}-\beta_k\big)$ is asymptotically normal with mean zero and variance $\mathrm{Avar}(\hat{\beta}_{\lambda,k})$, which denotes the k^{th} diagonal entry of $\mathrm{Avar}(\hat{\beta}_\lambda)$. This variance can be consistently estimated by $\widehat{\mathrm{Avar}}(\hat{\beta}_{\lambda,k})$, the k^{th} diagonal entry of $\widehat{\mathrm{Avar}}(\hat{\beta}_\lambda)$. In conjunction with this standard error, the estimator $\hat{\beta}_{\lambda,k}$ can be used for inference about β_k. For instance, asymptotically valid t confidence intervals are given by

$$\hat{\beta}_{\lambda,k}\pm t_{n-p,1-\alpha/2}\cdot\sqrt{\widehat{\mathrm{Avar}}(\hat{\beta}_{\lambda,k})/n}\ .$$

These intervals suffer from the same shortcomings as the asymptotic confidence intervals based on the weighted least squares estimator. But using the bootstrap can once again lead to improved finite-sample performance, and the following theorem establishes consistency of the bootstrap (and also bootstrap-t) distribution.

Theorem 5.3. *Under the conditions of Theorem 3.1,* $\sqrt{n}\big(\hat{\beta}^*_\lambda-\hat{\beta}_\lambda\big)$*, using the pairs or the wild bootstrap, converges weakly to the normal distribution with mean zero and variance* $Avar(\hat{\beta}_\lambda)$*, in probability for any fixed* λ*. Furthermore, for any* k*, the distribution of* $\sqrt{n}\big(\hat{\beta}^*_{\lambda,k}-\hat{\beta}_{\lambda,k}\big)/\widehat{Avar}(\hat{\beta}_{\lambda,k})^*$ *is asymptotically standard normal in probability.*

Although inference for β_k can be based on $\hat{\beta}_\lambda$ for any $\lambda\in[0,1]$, we would like to choose a value of λ that results in an efficient estimator. The asymptotic variance $\mathrm{Avar}(\hat{\beta}_{\lambda,k})$ is a quadratic function of λ, and therefore has a unique minimum, say λ_0, over the interval $[0,1]$ unless $\mathrm{Avar}(\hat{\beta}_{\lambda,k})$ is constant in λ (which may occur if there is homoskedasticity). In this case, define $\lambda_0=1$. Asymptotically, $\hat{\beta}_{\lambda_0,k}$ is the most efficient estimate of β_k amongst the collection $\big\{\hat{\beta}_{\lambda,k}:\lambda\in[0,1]\big\}$. Because this collection includes both the weighted and ordinary least squares estimators, $\hat{\beta}_{\lambda_0,k}$ is at least as efficient as the ordinary least squares estimator, and may have considerably smaller asymptotic variance when the skedastic function is well modeled. In fact, this estimate can have

smaller asymptotic variance than both the ordinary and weighted least squares estimators. Unfortunately, without knowing the asymptotic variance, we cannot find λ_0 and we cannot compute the estimate $\hat{\beta}_{\lambda_0,k}$. Instead, we can estimate λ_0 by $\hat{\lambda}_0$, the minimum of $\widehat{\text{Avar}}(\hat{\beta}_{\lambda,k})$ over the interval $[0,1]$, provided there is a unique minimum (otherwise set $\hat{\lambda}_0 = 1$). In particular, the minimizer is given by

$$\hat{\lambda}_0 = \frac{\left[\hat{\Omega}_{1/1}^{-1}\hat{\Omega}_{v/1}\hat{\Omega}_{1/1}^{-1} - \hat{\Omega}_{1/w}^{-1}\hat{\Omega}_{v/w}\hat{\Omega}_{1/1}^{-1}\right]_{k,k}}{\left[\hat{\Omega}_{1/w}^{-1}\hat{\Omega}_{v/w^2}\hat{\Omega}_{1/w}^{-1} - 2\cdot\hat{\Omega}_{1/w}^{-1}\hat{\Omega}_{v/w}\hat{\Omega}_{1/1}^{-1} + \hat{\Omega}_{1/1}^{-1}\hat{\Omega}_{v/1}\hat{\Omega}_{1/1}^{-1}\right]_{k,k}},$$

if this quantity lies in the interval [0,1], or otherwise $\hat{\lambda}_0$ is zero or one depending on which gives a smaller variance. If we choose to use the estimator, $\hat{\beta}_{\hat{\lambda}_0,k}$, then the confidence interval

$$\hat{\beta}_{\hat{\lambda}_0,k} \pm t_{n-p,1-\alpha/2}\cdot\sqrt{\frac{1}{n}\widehat{\text{Avar}}(\hat{\beta}_{\hat{\lambda}_0,k})}$$

will tend to have a coverage rate that is (much) smaller than the nominal level in finite samples, since the smallest estimated variance is likely downward biased for the true variance. Instead, reporting bootstrapped confidence intervals where the $\hat{\lambda}_0$ is recomputed for each bootstrap sample may give more reliable confidence intervals. The next theorem demonstrates that the bootstrap distribution of $\sqrt{n}\big(\hat{\beta}^*_{\hat{\lambda}^*_0,k} - \hat{\beta}_{\hat{\lambda}_0,k}\big)$ consistently approximates the sampling distribution of $\sqrt{n}\big(\hat{\beta}_{\hat{\lambda}_0,k} - \beta_k\big)$.

Theorem 5.4. *Under the conditions of Theorem 3.1, the sampling distribution of $\sqrt{n}\big(\hat{\beta}_{\hat{\lambda}_0,k} - \beta_k\big)$ converges weakly to the normal distribution with mean zero and variance $Avar(\hat{\beta}_{\lambda_0,k})$ and the bootstrap distribution of $\sqrt{n}\big(\hat{\beta}^*_{\hat{\lambda}^*_0,k} - \hat{\beta}_{\hat{\lambda}_0,k}\big)$ also converges weakly to the normal distribution with mean zero and variance $Avar(\hat{\beta}_{\lambda_0,k})$ in probability. Also, for any k, the distribution of $\sqrt{n}\big(\hat{\beta}^*_{\hat{\lambda}_0,k} - \hat{\beta}_{\lambda_0,k}\big)/\widehat{Avar}(\hat{\beta}_{\hat{\lambda},k})^*$ converges to the standard normal distribution in probability.*

6 Toy examples of linear combinations with lower variance

We will now give and example of a regression model where the optimal λ is in $[0,1]$ followed by an example where the optimal λ is outside of $[0,1]$.

For both examples, we will consider the simplest case, namely univariate regression through the origin:

$$y_i = \beta x_i + \varepsilon_i \ .$$

For the first example, let x_i be uniform on the interval $[-1,1]$ and ε_i have conditional mean zero and conditional variance $\text{var}(\varepsilon_i|x_i) = \sqrt{|x_i|}$. In this example, we will estimate the skedastic function from the family $\{v_\theta(x) = \theta\cdot|x| : \theta\in\mathbb{R}\}$. Consequently,

$$\theta_0 = \mathbb{E}\big(|x_i|^2\big)^{-1}\mathbb{E}\big(|x_i|\varepsilon_i^2\big) = \mathbb{E}\big(|x_i|^2\big)^{-1}\mathbb{E}\big(|x_i|^{3/2}\big) = \frac{6}{5}$$

λ:		0	.25	.50	.75	1	14/23
n = 20	eMSE	0.1449	0.1380	0.1345	0.1344	0.1378	0.1340
	Coverage	0.9613	0.9596	0.9575	0.9553	0.9527	0.9573
	Width	1.6645	1.6267	1.6066	1.6057	1.6247	1.6038
n = 50	eMSE	0.0564	0.0539	0.0527	0.0528	0.0540	0.0525
	Coverage	0.9524	0.9487	0.9465	0.9449	0.9448	0.9465
	Width	0.9589	0.9371	0.9258	0.9253	0.9360	0.9242
n = 100	eMSE	0.0270	0.0259	0.0254	0.0254	0.0261	0.0255
	Coverage	0.9520	0.9514	0.9506	0.9486	0.9481	0.9483
	Width	0.6592	0.6448	0.6375	0.6376	0.6450	0.6366

Table 6.1: Empirical mean squared error of estimators of β as well as coverage and width of confidence intervals based on the normal approximation

The estimator $(1-\lambda)\hat{\beta}_{\text{WLS}} + \lambda\hat{\beta}_{\text{OLS}}$ has variance

$$(1-\lambda)^2 \frac{\mathbb{E}\sqrt{|x_i|}}{\left(\mathbb{E}\,|x_i|\right)^2} + 2\lambda(1-\lambda)\frac{\mathbb{E}\,|x_i|^{3/2}}{\mathbb{E}\,|x_i|\,\mathbb{E}x_i^2} + \lambda^2 \frac{\mathbb{E}\,|x_i|^{5/2}}{\left(\mathbb{E}x_i^2\right)^2} \ ,$$

which is minimized by

$$\begin{aligned}
\lambda_0 &= 1 - \frac{-\frac{\mathbb{E}|x_i|^{3/2}}{\mathbb{E}|x_i|\mathbb{E}x_i^2} + \frac{\mathbb{E}|x_i|^{5/2}}{\left(\mathbb{E}x_i^2\right)^2}}{\frac{\mathbb{E}\sqrt{|x_i|}}{\left(\mathbb{E}|x_i|\right)^2} - 2\frac{\mathbb{E}|x_i|^{3/2}}{\mathbb{E}|x_i|\mathbb{E}x_i^2} + \frac{\mathbb{E}|x_i|^{5/2}}{\left(\mathbb{E}x_i^2\right)^2}} \\
&= 1 - \frac{-\frac{12}{5} + \frac{18}{7}}{\frac{8}{3} - 2\frac{12}{5} + \frac{18}{7}} \\
&= \frac{14}{23} \ .
\end{aligned}$$

Table 6.1 presents the empirical mean squared error of this estimator for various λ, as well as the coverage and average width of confidence intervals based on the normal approximation. For these simulations, the error terms are normally distributed.

For the second example, let the x_i be standard normal, and ε_i have conditional mean zero and conditional variance $\text{var}(\varepsilon_i|x_i) = x_i^2$. For the weighted least squares estimator, we will again use the incorrectly specified family of skedastic functions $\{v_\theta(x) = \theta \cdot |x| : \theta \in \mathbb{R}\}$.

In this example, the value of λ minimizing the asymptotic variance of $(1-\lambda)\hat{\beta}_{\text{WLS}} + \lambda\hat{\beta}_{\text{OLS}}$ is

$$\begin{aligned}\lambda_0 &= 1 - \frac{\mathbb{E}\left(x_i^2\right)^{-1}\mathbb{E}\left(x_i^4\right)\mathbb{E}\left(x_i^2\right)^{-1} - \mathbb{E}\left(|x_i|\right)^{-1}\mathbb{E}\left(|x_i|^3\right)\mathbb{E}\left(x_i^2\right)^{-1}}{\mathbb{E}\left(x_i^2\right)^{-1}\mathbb{E}\left(x_i^4\right)\mathbb{E}\left(x_i^2\right)^{-1} - 2 + \mathbb{E}\left(|x_i|\right)^{-1}\mathbb{E}x_i^2\mathbb{E}\left(|x_i|\right)^{-1}} \\ &= 1 - \frac{3-2}{\pi/2 - 4 + 3} \\ &\approx -0.75 \ .\end{aligned}$$

Although choosing values of lambda outside the interval $[0,1]$ may give estimators with lower variance, we recommend restricting lambda to the interval $[0,1]$. In situations where $\text{Avar}(\hat{\beta}_\lambda)$ is nearly constant in lambda (such as homoskedastic models), the estimates of λ can be highly unstable when not restricted, and the resulting intervals can have poor coverage. We recommend choosing $\hat{\lambda} = 0$ if the minimizing λ is negative, or $\hat{\lambda} = 1$ if the minimizing λ is positive. Even if the optimal lambda is outside the interval [0.1], choosing estimators in this way gives an estimator that asymptotically has the same variance as the better of the ordinary and weighted least squares estimators.

7 Simulations for confidence intervals

In this section, we present simulations studying the width and coverage of bootstrap and asymptotic approximation confidence intervals for regression coefficients. Simulations are given using the model

$$y_i = \alpha + x_i\beta + \sqrt{v(x_i)}\varepsilon_i$$

where $x_i \sim U(1,4)$ and ε_i are i.i.d. according to a distribution specified in several scenarios below. Several forms of the true skedastic function $v(\cdot)$ are used, and are specified in the tables. In each of the simulations, $(\alpha, \beta) = (0,0)$ and a confidence interval is constructed for β. The parametric family used to estimate the skedastic function is

$$v_\theta(x) := \exp\left(\theta_1 + \theta_2 \log|x|\right) \ .$$

The tables presented in this section compare the ordinary least squares estimator, the weighted least squares estimator, the estimator chosen between the ordinary and weighted estimators based on which has smaller sample variance, and the convex combination estimator giving smallest sample variance (referred to as OLS, WLS, Min, and Optimal, respectively). Simulations are presented using both the HC0 covariance estimator which is "the most commonly used heteroskedasticity-consistent covariance matrix estimator" (Cribari-Neto (2004)) as well as the HC3 estimator. Intervals based on a t-approximation use 10,000 simulations, while bootstrap intervals use 10,000 simulations with 1,000 bootstrap samples. For the wild bootstrap simulations, we scale the residuals by $1/(1-h_t)$ when generating bootstrap samples, where the h_t are defined in Remark 3.1.

Table 7.1 gives the empirical mean squared error when the errors, ε_i, are $N(0,1)$. Table 7.2 gives the coverage of and length of t-intervals using the HC0 covariance estimator. Table 7.4 repeats the simulations in table 7.2, but instead uses the HC3 estimator. These simulations are repeated using exponential (with parameter one, centered to have mean zero) errors in Table 7.2 (with HC0 estimators) and 7.4 (with HC3 estimators).

Tables 7.3 and 7.5 give the coverage and length of wild bootstrap-t intervals using the HC0 and HC3 estimators, respectively, when the errors are $N(0,1)$. Simulations with exponential errors are given in Table 7.7 (with HC0 estimators) and Table 7.7 (with HC3 estimators). In each of these tables, the residuals used for the wild bootstrap samples are calculated using the ordinary least squares estimator.

The empirical mean squared error of the weighted least squares estimator can be considerably smaller than that of the ordinary least squares estimator when the skedasticity is well modeled. When the family of skedastic functions is misspecified or there is conditional homoskedasticity, the weighted least squares may have worse mean squared error. While in several of the simulations, the empirical mean squared error of the weighted least squares estimator can be reduced by the ordinary least squares estimator, using the optimal combination, or the estimator with smallest variance gives similar performance to the better of the ordinary and weighted least squares estimators. Similarly, in each of the simulations, the width of intervals based on the convex combination estimator, or the estimator with smallest variance is close to the narrower of the intervals based on the ordinary and weighted least squares estimators. By using either the convex combination estimator or the estimator with minimum variance, there is little loss in efficiency when the model is homoskedastic. But, these estimators provide improvements in efficiency that are comparable to those realized by the weighted least squares estimator when the weighted estimator outperforms the ordinary least squares estimator.

When the errors are normally distributed, the t-intervals using HC0 standard errors can have coverage that is much lower than the nominal level, especially in small sample sizes. Furthermore, coverage of the t-intervals based on either the minimum variance or optimal convex combination estimator is considerably lower than the coverage of intervals based on either the ordinary or weighted least squares estimators. The wild-bootstrap-t intervals (with the HCO estimator) have coverage that is very close to the nominal level, regardless of sample size, for each of the estimators used. For the asymptotic approximation intervals, using the HC3 estimator substantially improved coverage over the HC0 estimator. With the HC3 estimators, t-intervals have coverage that is very close to the nominal level when using the ordinary least squares estimator. But for intervals based on each of the other estimators, the asymptotic intervals still have coverage that is slightly under the nominal level (especially in small samples) when using the HC3 estimators. The coverage of these intervals is not as close to the nominal level as the wild bootstrap-t intervals using the HC0 estimator. The wild bootstrap-t intervals using the HC3 estimators are more conservative than

λ:		OLS	WLS	Min	Optimal
n = 20, $v(x) = 1$	eMSE	0.0754	0.0838	0.0795	0.0794
n = 50, $v(x) = 1$	eMSE	0.0284	0.0297	0.0294	0.0292
n = 100, $v(x) = 1$	eMSE	0.0136	0.0140	0.0140	0.0138
n = 20, $v(x) = x^2$	eMSE	0.5611	0.4550	0.4824	0.4775
n = 50, $v(x) = x^2$	eMSE	0.2107	0.1555	0.1637	0.1627
n = 100, $v(x) = x^2$	eMSE	0.0511	0.0352	0.0363	0.0360
n = 20, $v(x) = \log(x)^2$	eMSE	0.0654	0.0457	0.0483	0.0487
n = 50, $v(x) = \log(x)^2$	eMSE	0.0249	0.0137	0.0138	0.0146
n = 100, $v(x) = \log(x)^2$	eMSE	0.0123	0.0063	0.0062	0.0065
n = 20, $v(x) = 4\exp(.02x + .02x^2)$	eMSE	0.3613	0.4088	0.3943	0.3816
n = 50, $v(x) = 4\exp(.02x + .02x^2)$	eMSE	0.1368	0.1450	0.1390	0.1405
n = 100, $v(x) = 4\exp(.02x + .02x^2)$	eMSE	0.0667	0.0686	0.0682	0.0677

Table 7.1: Empirical mean squared error of estimators of β as well as average coverage and width of confidence intervals based on an asymptotic approximation

those using the HC0 estimator, but are also wider.

As with normal errors, the wild bootstrap-t using the HC0 estimator have coverage that is better than the asymptotic intervals using either the HC0 or HC3 estimators when the errors have an exponential distribution. In this setting, the wild bootstrap-t method, using the HC3 estimator, gave intervals that have similar coverage to those using the HC0 estimator, but are somewhat wider.

Basing intervals on the minimum variance or optimal convex combination estimators performs similarly to using the weighted least squares estimator in situations when this estimator is more efficient, but never performs noticeably worse than intervals based on the ordinary least squares estimator. However, when using these estimators, intervals based on asymptotic approximations tend to under-cover. Using the wild-bootstrap-t method (especially with the HC0 estimator) produces intervals based on these estimators which have coverage that is closer to the nominal level. In each of the simulations, using the minimum variance or convex combination estimator produces confidence intervals whose width is similar to those given by weighting when there are improvements in efficiency to be had, but that are never substantially wider than those given by the ordinary least squares estimator.

λ:		OLS	WLS	Min	Optimal
n = 20, $v(x) = 1$	Coverage	0.9190	0.8976	0.8966	0.8962
	Width	1.0240	1.0004	0.9766	0.9728
n = 50, $v(x) = 1$	Coverage	0.9394	0.9323	0.9324	0.9319
	Width	0.6424	0.6381	0.6304	0.6294
n = 100, $v(x) = 1$	Coverage	0.9446	0.9385	0.9391	0.9387
	Width	0.4535	0.4520	0.4491	0.4488
n = 20, $v(x) = x^2$	Coverage	0.9076	0.9039	0.8908	0.8902
	Width	2.7364	2.4102	2.3589	2.3292
n = 50, $v(x) = x^2$	Coverage	0.9275	0.9341	0.9263	0.9263
	Width	1.7481	1.4848	1.4779	1.4640
n = 100, $v(x) = x^2$	Coverage	0.9387	0.9410	0.9396	0.9367
	Width	1.2414	1.0385	1.0375	1.0315
n = 20, $v(x) = \log(x)^2$	Coverage	0.9067	0.9197	0.9042	0.9022
	Width	0.9440	0.7703	0.7613	0.7513
n = 50, $v(x) = \log(x)^2$	Coverage	0.9308	0.9409	0.9384	0.9330
	Width	0.6001	0.4501	0.4498	0.4462
n = 100, $v(x) = \log(x)^2$	Coverage	0.9443	0.9460	0.9459	0.9430
	Width	0.4260	0.3071	0.3071	0.3060
n = 20, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9201	0.8980	0.8983	0.8977
	Width	2.2601	2.2107	2.1448	2.1333
n = 50, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9413	0.9317	0.9317	0.9318
	Width	1.4274	1.4154	1.3938	1.3910
n = 100, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9470	0.9429	0.9424	0.9432
	Width	1.0054	1.0003	0.9921	0.9910

Table 7.2: Average coverage and width of confidence intervals for β based on an asymptotic approximation using HC0 standard errors

λ:		OLS	WLS	Min	Optimal
	Coverage	0.9459	0.9466	0.9437	0.9433
n = 20, $v(x) = 1$	Width	1.1975	1.2613	1.2469	1.2333
	Coverage	0.9465	0.9473	0.9478	0.9459
n = 50, $v(x) = 1$	Width	0.6779	0.6962	0.6886	0.6871
	Coverage	0.9488	0.9499	0.9487	0.9483
n = 100, $v(x) = 1$	Width	0.4649	0.4713	0.4683	0.4682
	Coverage	0.9447	0.9472	0.9471	0.9456
n = 20, $v(x) = x^2$	Width	3.3458	3.0195	3.0330	3.0321
	Coverage	0.9416	0.9458	0.9487	0.9467
n = 50, $v(x) = x^2$	Width	1.8864	1.6009	1.6192	1.6152
	Coverage	0.9476	0.9515	0.9488	0.9510
n = 100, $v(x) = x^2$	Width	1.2863	1.0724	1.0796	1.0774
	Coverage	0.9461	0.9520	0.9487	0.9486
n = 20, $v(x) = \log(x)^2$	Width	1.1553	0.9336	0.9596	0.9594
	Coverage	0.9504	0.9523	0.9514	0.9511
n = 50, $v(x) = \log(x)^2$	Width	0.6476	0.4706	0.4838	0.4800
	Coverage	0.9513	0.9527	0.9523	0.9541
n = 100, $v(x) = \log(x)^2$	Width	0.4421	0.3117	0.3175	0.3144
	Coverage	0.9487	0.9462	0.9469	0.9467
n = 20, $v(x) = 4\exp(.02x + .02x^2)$	Width	2.6898	2.8535	2.7960	2.7791
	Coverage	0.9431	0.9444	0.9493	0.9443
n = 50, $v(x) = 4\exp(.02x + .02x^2)$	Width	1.5050	1.5495	1.5220	1.5212
	Coverage	0.9475	0.9471	0.9493	0.9487
n = 100, $v(x) = 4\exp(.02x + .02x^2)$	Width	1.0341	1.0476	1.0392	1.0390

Table 7.3: Average coverage and width of confidence intervals for β based on the bootstrap-t method using the wild bootstrap with HC0 covariance estimators

λ:		OLS	WLS	Min	Optimal
n = 20, $v(x) = 1$	Coverage	0.9507	0.9353	0.9340	0.9338
	Width	1.1950	1.1608	1.1341	1.1301
n = 50, $v(x) = 1$	Coverage	0.9491	0.9423	0.9412	0.9411
	Width	0.6805	0.6755	0.6669	0.6659
n = 100, $v(x) = 1$	Coverage	0.9500	0.9449	0.9457	0.9463
	Width	0.4661	0.4646	0.4616	0.4612
n = 20, $v(x) = x^2$	Coverage	0.9495	0.9445	0.9364	0.9362
	Width	3.2361	2.8017	2.7418	2.7117
n = 50, $v(x) = x^2$	Coverage	0.9490	0.9481	0.9451	0.9434
	Width	1.8600	1.5711	1.5637	1.5500
n = 100, $v(x) = x^2$	Coverage	0.9465	0.9482	0.9469	0.9458
	Width	1.2761	1.0641	1.0634	1.0574
n = 20, $v(x) = \log(x)^2$	Coverage	0.9494	0.9496	0.9401	0.9408
	Width	1.1017	0.8774	0.8687	0.8595
n = 50, $v(x) = \log(x)^2$	Coverage	0.9461	0.9516	0.9498	0.9466
	Width	0.6375	0.4706	0.4704	0.4675
n = 100, $v(x) = \log(x)^2$	Coverage	0.9465	0.9498	0.9496	0.9477
	Width	0.4379	0.3134	0.3134	0.3125
n = 20, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9548	0.9388	0.9358	0.9368
	Width	2.6677	2.6016	2.5252	2.5134
n = 50, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9512	0.9431	0.9435	0.9437
	Width	1.5151	1.5042	1.4807	1.4778
n = 100, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9516	0.9497	0.9484	0.9492
	Width	1.0375	1.0338	1.0245	1.0234

Table 7.4: Average coverage and width of confidence intervals for β based on an asymptotic approximation using HC3 standard errors

λ:		OLS	WLS	Min	Optimal
n = 20, $v(x) = 1$	Coverage	0.9560	0.9527	0.9435	0.9433
	Width	1.3105	1.3789	1.2533	1.2519
n = 50, $v(x) = 1$	Coverage	0.9546	0.9537	0.9487	0.9486
	Width	0.6971	0.7156	0.6863	0.6855
n = 100, $v(x) = 1$	Coverage	0.9511	0.9521	0.9487	0.9489
	Width	0.4722	0.4788	0.4685	0.4684
n = 20, $v(x) = x^2$	Coverage	0.9566	0.9566	0.9421	0.9418
	Width	3.6474	3.2782	3.0676	3.0574
n = 50, $v(x) = x^2$	Coverage	0.9535	0.9559	0.9528	0.9497
	Width	1.9403	1.6507	1.6262	1.6197
n = 100, $v(x) = x^2$	Coverage	0.9521	0.9498	0.9512	0.9474
	Width	1.3060	1.0882	1.0817	1.0786
n = 20, $v(x) = \log(x)^2$	Coverage	0.9527	0.9620	0.9472	0.9466
	Width	1.2643	1.0030	0.9704	0.9599
n = 50, $v(x) = \log(x)^2$	Coverage	0.9536	0.9536	0.9525	0.9502
	Width	0.6702	0.4828	0.4817	0.4801
n = 100, $v(x) = \log(x)^2$	Coverage	0.9537	0.9529	0.9523	0.9511
	Width	0.4485	0.3147	0.3156	0.3136
n = 20, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9604	0.9580	0.9476	0.9471
	Width	2.9172	3.0692	2.7876	2.7802
n = 50, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9556	0.9571	0.9415	0.9501
	Width	1.5494	1.5941	1.5285	1.5221
n = 100, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9479	0.9477	0.9463	0.9436
	Width	1.0476	1.0612	‘1.0406	1.0370

Table 7.5: Average coverage and width of confidence intervals for β based on the wild bootstrap-t method with HC3 covariance estimates

λ:		OLS	WLS	min	Optimal
n = 20, $v(x) = 1$	Coverage	0.9302	0.8841	0.8862	0.8857
	Width	0.9788	0.9250	0.8980	0.8927
n = 20, $v(x) = x^2$	Coverage	0.8870	0.8810	0.8720	0.8694
	Width	2.6241	2.2717	2.2002	2.1681
n = 20, $v(x) = \log(x)^2$	Coverage	0.8705	0.8784	0.8621	0.8621
	Width	0.8967	0.7344	0.7178	0.7075
n = 20, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9306	0.8860	0.8869	0.8840
	Width	2.1813	2.0628	1.9916	1.9754

Table 7.6: Average coverage and width of confidence intervals for β based on the asymptotic approximation using the HC0 covariance estimator with exponential errors

λ:		OLS	WLS	Min	Optimal
n = 20, $v(x) = 1$	Coverage	0.9570	0.9276	0.9342	0.9340
	Width	1.1415	1.1468	1.0981	1.1200
n = 20, $v(x) = x^2$	Coverage	0.9285	0.9247	0.9268	0.9262
	Width	3.1013	2.8088	2.7857	2.7766
n = 20, $v(x) = \log(x)^2$	Coverage	0.9024	0.9132	0.9120	0.9074
	Width	1.0724	0.8889	0.8920	0.8969
n = 20, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9565	0.9292	0.9345	0.9377
	Width	2.4942	2.5328	2.4396	2.4653

Table 7.7: Average coverage and width of wild bootstrap-t confidence intervals for β using the HC0 covariance estimator with exponential errors

λ:		OLS	WLS	Min	Optimal
n = 20, $v(x) = 1$	Coverage	0.9648	0.9310	0.9312	0.9319
	Width	1.1487	1.0768	1.0483	1.0428
n = 20, $v(x) = x^2$	Coverage	0.9243	0.9178	0.9125	0.9100
	Width	3.0674	2.6113	2.5353	2.5031
n = 20, $v(x) = \log(x)^2$	Coverage	0.9104	0.9089	0.9009	0.8996
	Width	1.0411	0.8350	0.8191	0.8092
n = 20, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9673	0.9284	0.9287	0.9265
	Width	2.5522	2.3858	2.3098	2.2948

Table 7.8: Average coverage and width of confidence intervals for β based on the asymptotic approximation using the HC3 covariance estimator with exponential errors

λ:		OLS	WLS	Min	Optimal
n = 20, $v(x) = 1$	Coverage	0.9661	0.9393	0.9352	0.9333
	Width	1.2303	1.2262	1.1314	1.1172
n = 20, $v(x) = x^2$	Coverage	0.9316	0.9355	0.9221	0.9201
	Width	3.3637	3.0195	2.8386	2.7818
n = 20, $v(x) = \log(x)^2$	Coverage	0.9171	0.9238	0.9040	0.9070
	Width	1.1605	0.9568	0.9093	0.9009
n = 20, $v(x) = 4\exp(.02x + .02x^2)$	Coverage	0.9680	0.9429	0.9357	0.9363
	Width	2.7516	2.7648	2.5510	2.5103

Table 7.9: Average coverage and width of wild bootstrap-t confidence intervals for β using the HC0 covariance estimator with exponential errors

8 Appendix

Proof of Theorem 3.1. For a fixed function $w(\cdot)$, define $W \coloneqq \operatorname{diag}\{w(x_1), ..., w(x_n)\}$ and

$$\hat{\beta}_W \coloneqq (X^\top W^{-1} X)^{-1} X^\top W^{-1} Y \ .$$

If the skedastic function is estimated from a family $\{v_\theta\}$ by $v_{\hat{\theta}}$, the weighted least squares estimator is given by by

$$\hat{\beta}_{\text{WLS}} \coloneqq (X^\top V_{\hat{\theta}}^{-1} X)^{-1} X^\top V_{\hat{\theta}}^{-1} Y$$

where $V_\theta \coloneqq \operatorname{diag}\{v_\theta(x_1), ..., v_\theta(x_n)\}$. We would like to show that the bootstrap distribution $\sqrt{n}\big(\hat{\beta}^*_{WLS} - \hat{\beta}_{\text{WLS}}\big)$ (conditional on the data) consistently approximates the sampling distribution of $\sqrt{n}\big(\hat{\beta}_{\text{WLS}} - \beta\big)$. To do this, we will first show that the distribution of $\sqrt{n}\big(\hat{\beta}^*_W - \hat{\beta}_W\big)$ consistently approximates the distribution of $\sqrt{n}\big(\hat{\beta}_W - \beta\big)$ for a fixed W (satisfying some regularity conditions). We will then show that $\sqrt{n}\big(\hat{\beta}^*_{WLS} - \hat{\beta}_{\text{WLS}}\big) - \sqrt{n}\big(\hat{\beta}^*_W - \hat{\beta}_W\big)$ converges in conditional probability to zero for $W = V_{\theta_0}$, assuming that the estimate $\hat{\theta}^*$ of the variance parameter is conditionally consistent for some fixed θ_0. That is, the proof of Theorem 3.1 will rely on Lemmas 8.1 and 8.2 which are stated below. ■

Lemma 8.1. *Suppose that $(x_1, y_1), ..., (x_n, y_n)$ are i.i.d. satisfying assumptions (A1)–(A6). Suppose that $w : \mathbb{R}^d \to \mathbb{R}^+$ is a fixed and known function (although not necessarily the true skedastic function) and satisfies*

$$\mathbb{E}\left(\left\|\left(\frac{x_{i1}}{\sqrt{w(x_i)}}, ..., \frac{x_{ip}}{\sqrt{w(x_i)}}, \frac{y_i}{\sqrt{w(x_i)}}\right)\right\|_2^4\right) < \infty \ .$$

Define $W \coloneqq diag(w(x_1), ..., w(x_n))$, and let $\hat{\beta}_W \coloneqq (X^\top W^{-1} X)^{-1} X^\top W^{-1} Y$. Then, for almost all sample sequences, the conditional law of $\sqrt{n}\big(\hat{\beta}^*_W - \hat{\beta}_W\big)$ converges weakly to the normal distribution with mean 0 and variance $\Omega_{1/w}^{-1} \Omega_{v/w^2} \Omega_{1/w}^{-1}$.

Proof of Lemma 8.1 using the pairs bootstrap. Let C_P be the set of sequences $\{P_n\}$ such that

(B1) P_n converges weakly to P (the distribution of (x_i, y_i)).

(B2) $\beta_W(P_n) \coloneqq \left(\int \frac{1}{w(x)} x x^\top dP_n\right)^{-1} \cdot \int \frac{1}{w(x)} xy dP_n \to \beta$.

(B3) $\int \frac{1}{w(x)} x x^\top dP_n \to \Omega_{1/w}$.

(B4) $\int \big(1/w(x) x^\top (y - x\beta_W(P_n))\big)^\top \big(1/w(x) x^\top (y - x\beta_W(P_n))\big) dP_n \to \Omega_{v/w^2}$.

To prove the lemma, we will first show that the distribution of $\sqrt{n}\big(\hat{\beta}_W - \beta_W(P_n)\big)$ under P_n converges weakly to the normal distribution with mean 0 and variance $\Omega_{1/w}^{-1} \Omega_{v/w^2} \Omega_{1/w}^{-1}$ whenever $\{P_n\} \in C_p$, and then show that the empirical distribution is in C_p almost surely.

Let $(x_{n,i}, y_{n,i})$, $i = 1, ..., n$ be independent and identically distributed according to P_n such that $\{P_n\} \in C_P$.

Define residuals $\varepsilon_{n,i} \coloneqq Y_{n,i} - X_{n,i}\beta_W(P_n)$ so that

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_W - \beta_W(P_n)\right) &= \sqrt{n}\left(X_n^\top W^{-1} X_n\right)^{-1} X_n^\top W^{-1}\left(\varepsilon_n + X_n\beta_W(P_n)\right) - \beta_W(P_n) \\ &= \left(\frac{1}{n} X_n^\top W^{-1} X_n\right)^{-1} \sqrt{n} X_n^\top W^{-1}\varepsilon_n \ .\end{aligned}$$

It follows immediately from the assumptions that

$$\left(\frac{1}{n} X_n^\top W^{-1} X_n\right)^{-1} \xrightarrow{P} \Omega_{1/w}^{-1} \ ,$$

and we have the desired asymptotic normal distribution if we can show

$$\sqrt{n} X_n^\top W^{-1}\varepsilon_n \xrightarrow{d} N(0, \Omega_{v/w^2}) \ .$$

We will first consider the case of $x_i \in \mathbb{R}$. Because

$$\int x_{n,i}^\top \frac{1}{w(x_{n,i})}(y_{n,i} - x_{n,i}\beta_W(P_n)) dP_n = 0 \ ,$$

and

$$\int x_{n,i}^\top x_{n,i} \frac{1}{w^2(x_{n,i})}\varepsilon_{n,i}^2 dP_n \to \Omega_{v/w^2} \ ,$$

the asymptotic normality follows from the Lindeberg-Feller Central Limit Theorem if we can verify that

$$\mathbb{E}\left(x_{n,1}^2 \frac{1}{w^2(x_{n,1})}\varepsilon_{n,1}^2 \mathbb{1}\left\{x_{n,1}^2 \frac{1}{w^2(x_{n,1})}\varepsilon_{n,1}^2 > n\delta\right\}\right) \to 0$$

for all $\delta > 0$, where $\mathbb{1}\{\cdot\}$ denotes the indicator function of a set. Since $\beta_W(P_n) \to \beta$ and $(x_{n,i}, y_{n,i}) \xrightarrow{d} (X, Y) \sim P$,

$$x_{n,1}\frac{1}{w(x_{n,1})}\varepsilon_{n,1} \xrightarrow{d} \frac{X}{w(X)}(Y - X\beta) = \frac{X}{w(X)}\varepsilon \ .$$

Therefore, for any fixed γ that is a continuity point of the distribution of $X\varepsilon/w(X)$ and $n > \gamma/\delta$, we have that

$$\begin{aligned}\mathbb{E}\left(x_{n,1}^2 \frac{1}{w^2(x_{n,1})}\varepsilon_{n,1}^2 \mathbb{1}\left\{x_{n,1}^2 \frac{1}{w^2(x_{n,1})}\varepsilon_{n,1}^2 > n\delta\right\}\right) &\leq E\left(x_{n,1}^2 \frac{1}{w^2(x_{n,1})}\varepsilon_{n,1}^2 \mathbb{1}\left\{x_{n,1}^2 \frac{1}{w^2(x_{n,1})}\varepsilon_{n,1}^2 > \gamma\right\}\right) \\ &\to E\left(X^2 \frac{1}{w^2(X)}\varepsilon^2 \mathbb{1}\left\{X^2 \frac{1}{w^2(X)}\varepsilon^2 > \gamma\right\}\right) \ .\end{aligned}$$

The Lindeberg-Feller condition is satisfied, since the right-hand side of this equation can be made arbitrarily small by choosing γ sufficiently large. The multivariate case follows analogously using the Cramér-Wold device. For any vector of constants, $C \in \mathbb{R}^p$, we must show

$$\sum_{i=1}^n \frac{\varepsilon_{n,i}}{w(x_{n,i})} x_{n,i} C \xrightarrow{d} N(0, C^\top \Omega_{v/w^2} C) \ .$$

This convergence follows from the Lindeberg-Feller CLT if

$$\mathbb{E}\left(\left(\frac{\varepsilon_{n,i}}{w(x_{n,i})}x_{n,i}C\right)^2 \mathbb{1}\left\{\left(\frac{\varepsilon_{n,i}}{w(x_{n,i})}x_{n,i}C\right)^2 > n\delta\right\}\right) \to 0$$

for all $\delta > 0$. This convergence holds by the same argument as in one dimensional case given above. It is easily seen that the empirical distribution functions $\hat{P}_n$ are almost surely in C_P, and the result of the theorem follows. ■

Proof of Lemma 8.1 using the wild bootstrap. Let S be the set of sequences $\{x_i, y_i\}$ satisfying the following conditions:

(S1) $\hat{\beta}_W \to \beta$,

(S2) $\hat{\Omega}_{1/w} \to \Omega_{1/w}$,

(S3) $\hat{\Omega}_{v/w^2} \to \Omega_{v/w^2}$, and

(S4) $\sqrt{n}\left(\hat{\beta}_{\text{WLS}} - \hat{\beta}_W\right) \to 0$.

Write

$$\sqrt{n}\left(\hat{\beta}^*_W - \hat{\beta}_W\right) = \sqrt{n}\left(X_n^\top W^{-1} X_n\right)^{-1} X_n^\top W^{-1}\hat{\varepsilon}^* + \sqrt{n}\left(\hat{\beta}_{\text{WLS}} - \hat{\beta}_W\right) .$$

On S, $\left(\frac{1}{n}X_n^\top W^{-1}X_n\right)^{-1} \to \Omega_{1/w}$, and $\sqrt{n}\left(\hat{\beta}_{\text{WLS}} - \hat{\beta}_W\right) \to 0$. Thus, to show the desired asymptotic normality, it suffices to show that, on S, $W^{-1}\hat{\varepsilon}^* \xrightarrow{d} N(0, \Omega_{v/w^2})$ conditionally on the $x's$ and $y's$. This convergence holds using the Cramér-Wold device, since for each vector $c \in \mathbb{R}^p$,

$$c^\top X_n^\top W^{-1}\hat{\varepsilon}^* = \sum x_i c \frac{1}{w(x_i)}\hat{\varepsilon}^*$$

which is asymptotically normal with mean zero and variance $c^\top \Omega_{v/w^2} c$ by the Lindeberg-Feller Central Limit Theorem which is applicable because condition $(S3)$ holds.

The conditions specified by the set S do not hold almost surely, but they do hold in probability. By the Almost Sure Representation Theorem, there exist versions of the X's and Y's such that S holds almost surely. It follows that the asymptotic normality of the wild bootstrap distribution holds in probability. ■

Lemma 8.2. *Suppose that $\hat{\theta}^*$ is consistent for θ_0, in the sense that $n^{1/4}(\hat{\theta}^* - \theta_0)$ converges in conditional probability to zero. Suppose that $\hat{\beta}_{WLS} := (X^\top V_{\hat{\theta}}^{-1} X)^{-1} X^\top V_{\hat{\theta}}^{-1} Y$ and $v_{\theta_0} =: w$ so that $W := diag(v_{\theta_0}(X_1), ..., v_{\theta_0}(X_n))$. Under the assumptions of Theorem 3.1,*

$$\sqrt{n}(\hat{\beta}^*_{WLS} - \hat{\beta}_{WLS}) - \sqrt{n}(\hat{\beta}^*_W - \hat{\beta}_W) \xrightarrow{P} 0$$

in probability.

Proof of Lemma 8.2 using the pairs bootstrap. Let C_P be the set of sequences $\{P_n\}$ that satisfy the following conditions:

(C1) P_n converges weakly to P

(C2) $\int \frac{1}{w(x)} x x^\top dP_n \to \Omega_{1/w}$

(C3) $\int \left(1/w(x) x^\top (y - x\beta_W(P_n)\right)^\top \left(1/w(x) x^\top (y - x\beta_W(P_n)\right) dP_n \to \Omega_{v/w^2}$

(C4) $n^{1/4} \left(\beta_W(P_n) - \beta(P_n)\right) \to 0$

(C5) $n^{1/4} \mathbb{E}_{P_n} \left(x_i (y - x\beta(P_n)) r_{\theta_0,l}(x)\right) \to 0$ for each $i = 1, ..., p$, $l = 1, ..., d$

(C6) $\mathbb{E}_{P_n} \left| x_i \varepsilon r_{\theta_0,l}(x) \right|^2 \to \mathbb{E}_P(|x_i \varepsilon r_{\theta_0,l}(x)|^2)$ for each $i = 1, ..., p$, $l = 1, ..., d$

(C7) $\mathbb{E}_{P_n} |x_i \varepsilon s_{\theta_0}(x)|^2 \to \mathbb{E}_P(|x_i \varepsilon s_{\theta_0}(x)|^2)$ for each $i = 1, ..., p$, $l = 1, ..., d$

(C8) $n^{1/4}\left(\hat{\theta} - \theta_0\right)$ converges in P_n-probability to zero

Suppose that $(x_{n,i}, y_{n,i})$, $i = 1, ..., n$ are i.i.d. according to P_n where $\left\{P_n\right\}$ is any sequence in C_P.

Define the residuals

$$\varepsilon_{\hat{W},n,i} := y_{n,i} - x_{n,i}\beta_{\hat{W}}(P_n) \ ,$$
$$\varepsilon_{n,i} := y_{n,i} - x_{n,i}\beta(P_n) \ ,$$

and

$$\varepsilon_{W,n,i} := y_{n,i} - x_{n,i}\beta_W(P_n)$$

where

$$\beta_{\hat{W}}(P_n) := \left(\int \frac{1}{v_{\hat{\theta}}(x)} x x^\top dP_n\right)^{-1} \int \frac{1}{v_{\hat{\theta}}(x)} x y dP_n \ ,$$
$$\beta(P_n) := \left(\int x x^\top dP_n\right)^{-1} \int x y dP_n \ ,$$

and

$$\beta_W(P_n) := \left(\int \frac{1}{w(x)} x x^\top dP_n\right)^{-1} \int \frac{1}{w(x)} x y dP_n \ .$$

Then,

$$\sqrt{n}\left(\hat{\beta}_{\text{WLS}} - \beta_{\text{WLS}}(P_n)\right) - \sqrt{n}\left(\hat{\beta}_W - \beta_W(P_n)\right) = (X_n^\top \hat{W}^{-1} X_n)^{-1} X_n^\top \hat{W}^{-1} \varepsilon_{\hat{W},n}$$
$$- (X_n^\top W^{-1} X_n)^{-1} X_n^\top W^{-1} \varepsilon_{W,n} \ .$$

To show this quantity converges in probability to zero, it suffices to show that

$$\frac{1}{\sqrt{n}}\left(X_n^\top \hat{W}^{-1} \varepsilon_{\hat{W},n} - X_n^\top W^{-1} \varepsilon_{W,n}\right) \xrightarrow{P} 0$$

and

$$\frac{1}{n}\left(X_n^\top \hat{W}^{-1} X_n - X_n^\top W^{-1} X_n\right) \xrightarrow{P} 0 \ .$$

We can write the first expression as

$$\frac{1}{\sqrt{n}}\left[X_n^\top \left(\hat{W}^{-1} - W^{-1}\right)\varepsilon_{W,n} + X_n^\top \hat{W}^{-1} X_n \left(\beta_{\hat{W}}(P_n) - \beta_W(P_n)\right)\right] \ .$$

By the assumptions on sequences in C_P, $\sqrt{n}\left(\beta_{\hat{W}} - \beta_W\right) \xrightarrow{P} 0$. It will be seen later that $\frac{1}{n} X_n^\top \hat{W}^{-1} X_n \xrightarrow{P} \mathbb{E}(x^\top x / w(x))$, so the second tern in the above expression converges to zero in probability. The first term is

$$\frac{1}{\sqrt{n}} X_n^\top \left(\hat{W}^{-1} - W^{-1}\right)\varepsilon_{W,n} = \frac{1}{\sqrt{n}} \sum x_{n,i}^\top \left(\frac{1}{v_{\hat{\theta}}(x_{n,i})} - \frac{1}{v_{\theta_0}(x_{n,i})}\right)\varepsilon_{W,n,i}$$

which, as in Romano and Wolf (2015), can be written as $A + B$ where the j^{th} entry of A is

$$A_j = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{n,i,j}\varepsilon_{W,n,i} \sum_{l=1}^{K} r_{\theta_0,l}(x_{n,i})(\hat{\theta}_l - \theta_{0,l}) \ ,$$

and with probability tending to one,

$$|B_j| \leq \frac{1}{2\sqrt{n}} \left|\hat{\theta} - \theta_0\right|^2 \sum |x_{n,i,j}\varepsilon_{W,n,i} s_{\theta_0}(x_{n,i})| \ .$$

Because $n^{1/4}(\hat{\theta}_l - \theta_{0,l}) \xrightarrow{P} 0$, to show $A_j \xrightarrow{P} 0$, we only need to show that

$$n^{-3/4} \sum_{i=1}^{n} x_{n,i,j}\varepsilon_{W,n,i} r_{\theta_0,l}(x_{n,i}) \xrightarrow{P} 0$$

for each $l = 1, ..., K$. We will do this by showing that the mean and variance converge to zero.

The variance converges to zero since

$$\text{var}_{P_n}\left(n^{-3/4} \sum_{i=1}^{n} x_{n,i,j}\varepsilon_{W,n,i} r_{\theta_0,l}(x_{n,i})\right) = n^{-1/2} \text{var}_{F_n}\left(x_{n,i,j}\varepsilon_{W,n,i} r_{\theta_0,l}(x_{n,i})\right)$$

and, by the assumptions on C_P, the sequence of variances $\text{var}_{P_n}\left(x_{n,i,j}\varepsilon_{W,n,i} r_{\theta_0,l}(x_{n,i})\right)$ is bounded. To show that the mean converges to zero, write

$$n^{-3/4} \sum_{i=1}^{n} x_{n,i,j}\varepsilon_{W,n,i} r_{\theta_0,l}(x_{n,i}) = n^{-3/4} \sum_{i=1}^{n} x_{n,i,j}\varepsilon_{n,i} r_{\theta_0,l}(x_{n,i}) + n^{-3/4} \sum_{i=1}^{n} (\varepsilon_{W,n,i} - \varepsilon_{n,i}) x_{n,i,j} r_{\theta_0,l}(x_{n,i}) \ .$$

The expectation of the first term converges to zero by assumption and the expectation of the second term converges to zero, since

$$\mathbb{E}_{P_n}\left(n^{-3/4} \sum_{i=1}^{n} x_{n,i,j}\varepsilon_{n,i} r_{\theta_0,l}(x_{n,i})\right) = \mathbb{E}_{P_n}\left(\frac{1}{n} X_{n,i} x_{n,i,j} r_{\theta_0,l}(x_{n,i})\right) n^{1/4}\left(\hat{\beta}(P_n) - \hat{\beta}_W(P_n)\right) \to 0 \ .$$

Similarly, since $\sqrt{n}\left|\hat{\theta}-\theta_0\right|^2 \xrightarrow{P} 0$, we have that $|B_j| \xrightarrow{P} 0$ provided $\frac{1}{n}\sum\left|x_{n,i,j}\varepsilon_{W,n,i}s_{\theta_0}(x_{n,i})\right| = O_p(1)$. As in the argument for A_j, this last sum has expectation tending to a constant, and variance tending to zero, and so it converges in probability to a constant.

Finally we must show that

$$\frac{1}{n}\left(X_n^\top \hat{W}^{-1}X_n - X_n^\top W^{-1}X_n\right) = \frac{1}{n}\sum x_i^\top x_{n,i}\left(\frac{1}{v_{\hat{\theta}}(x_{n,i})} - \frac{1}{v_{\theta_0}(x_{n,i})}\right)$$

converges in probability to zero. The argument proceeds as above.

Since $\sqrt{n}\big(\hat{\beta}_{\hat{W}} - \hat{\beta}_W\big)$ converges to zero in probability, but not necessarily almost surely, the empirical distribution functions $\hat{P}_n$ do not lie in C_P almost surely. However, it is easily seen that the empirical distribution functions satisfy the moment conditions on C_P in probability, so the asymptotic normality of the bootstrap distribution holds in probability. ■

Proof of Lemma 8.2 using the wild bootstrap. Let S' be the set on which $(S1)$–$(S4)$ hold as well as

(S5) $\frac{1}{n}\sum_{i=1}^n \left|x_i\hat{y}_i r_{\theta_0,l}(x)\right|^2 \to \mathbb{E}_P(|x_i y_i r_{\theta_0,l}(x)|^2)$ for each $i = 1,...,p$, $l = 1,...,d$,

(S6) $\frac{1}{n}\sum_{i=1}^n |x_i\hat{y}_i s_{\theta_0}(x)|^2 \to \mathbb{E}_P(|x_i y_i s_{\theta_0}(x)|^2)$ for each $i = 1,...,p$, $l = 1,...,d$, and

(S7) $n^{1/4}\big(\hat{\theta}^* - \theta_0\big)$ converges in probability to zero.

We will show that

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}^*_{\mathrm{WLS}} - \hat{\beta}_{\mathrm{WLS}}\right) - \sqrt{n}\left(\hat{\beta}^*_W - \hat{\beta}_W\right) =& \sqrt{n}\left[\left(X^\top W^{*-1}X\right)^{-1}X^\top W^{*-1}\varepsilon^* \right.\\ &\left. - \left(X^\top W^{-1}X\right)^{-1}X^\top W^{-1}\varepsilon^*\right] + \sqrt{n}\left(\hat{\beta}_{\mathrm{WLS}} - \hat{\beta}_W\right)\end{aligned}$$

converges to probability to zero, conditional on any sequence of $x's$ and $y's$ in S'.

By assumption, the second term converges to zero on S'. To show the first term converges in probability to zero, we will show that

$$\frac{1}{\sqrt{n}}\left(X_n^\top \hat{W}^{*-1}\varepsilon^* - X_n^\top W^{-1}\varepsilon^*\right) \xrightarrow{P} 0$$

and

$$\frac{1}{n}\left(X_n^\top \hat{W}^{*-1}X_n - X_n^\top W^{-1}X_n\right) \xrightarrow{P} 0 \ .$$

The first quantity can be written as

$$\frac{1}{\sqrt{n}}X_n^\top\left(\hat{W}^{-1} - W^{-1}\right)\varepsilon^* = \frac{1}{\sqrt{n}}\sum x_{n,i}^\top\left(\frac{1}{v_{\hat{\theta}^*}(x_{n,i})} - \frac{1}{v_{\theta_0}(x_{n,i})}\right)\varepsilon_i^*$$

which again can be written as $A + B$ where the j^{th} entry of A is

$$A_j = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{n,i,j} \varepsilon_i^* \sum_{l=1}^{K} r_{\theta_0,l}(x_{n,i})(\hat{\theta}_l^* - \theta_{0,l}) \ ,$$

and with probability tending to one,

$$|B_j| \leq \frac{1}{2\sqrt{n}} \left| \hat{\theta}^* - \theta_0 \right|^2 \sum |x_{n,i,j} \varepsilon_i^* s_{\theta_0}(x_{n,i})| \ .$$

By assumption $(S7)$, $n^{1/4}(\hat{\theta}_l^* - \theta_{0,l}) \xrightarrow{p} 0$. Further, for each l, $n^{-3/4} \sum_{i=1}^{n} x_{n,i,j} \varepsilon_i^* \sum_{l=1}^{K} r_{\theta_0,l}(x_{n,i})$ converges in probability to zero since it has mean zero and variance

$$\text{var}\left(n^{-3/4} \sum_{i=1}^{n} x_{n,i,j} \varepsilon_i^* r_{\theta_0,l}(x_{n,i}) \right) = n^{-3/2} \sum_{i=1}^{n} (x_{n,i,j} \hat{\varepsilon}_i r_{\theta_0,l}(x_{n,i}))^2$$

which converges to zero on S' by assumption $(S5)$. Consequently, A_j converges in probability to zero for each j. Similarly, B_j converges in probability to zero since $\sqrt{n}(\hat{\theta}_l^* - \theta_{0,l})^2$ converges in probability to zero, and $\frac{1}{n} \sum |x_{n,i,j} \varepsilon_i^* s_{\theta_0}(x_{n,i})|$ converges in probability to a constant.

The other convergence,

$$\frac{1}{n} \left(X_n^\top \hat{W}^{*-1} X_n - X_n^\top W^{-1} X_n \right) \xrightarrow{P} 0 \ ,$$

follows from a similar argument. ■

Proof of Lemma 3.1. We will first consider the estimate $\tilde{\theta}$ obtained by regressing $h_\delta(\varepsilon_i)$ on $g(x_i)$. By a similar argument to Lemma 8.1, $\sqrt{n}\left(\tilde{\theta}^* - \tilde{\theta}\right)$ is almost surely asymptotically normal. Consequently, $n^{1/4}\left(\tilde{\theta}^* - \tilde{\theta}\right)$ converges in conditional probability to zero, almost surely. We can express

$$\begin{aligned} n^{1/4}\left(\tilde{\theta} - \theta_0\right) &= n^{1/4}\left((G^\top G)^{-1} G^\top h - \theta_0\right) \\ &= n^{1/4}(G^\top G)^{-1} G^\top e \ . \end{aligned}$$

where G and h are the matrix and vector containing the $g(x_i)$ and $h_\delta(\varepsilon_i)$, respectively, and e is the vector with entries $e_i = h_\delta(y_i) - g(x)\theta_0$. Since $(\frac{1}{n} G^\top G)^{-1}$ converges almost surely to $\mathbb{E}(g(x_i)^\top g(x_i))$ and $n^{-3/4} G^\top e$ converges in almost surely to zero, $n^{1/4}\left(\tilde{\theta} - \theta_0\right)$ converges almost surely to zero.

Writing

$$n^{1/4}\left(\tilde{\theta}^* - \theta_0\right) = n^{1/4}\left(\tilde{\theta}^* - \tilde{\theta}\right) + n^{1/4}\left(\tilde{\theta} - \theta_0\right) \ ,$$

we see this quantity converges in conditional probability to zero, almost surely.

Now,

$$\hat{\theta}^* - \tilde{\theta}^* = \left(\frac{1}{n} \sum g(x_i^*) g^\top(x_i^*) \right)^{-1} \frac{1}{n} \sum g(x_i^*) \left(h_\delta(\hat{\varepsilon}_i^*) - h_\delta(\varepsilon_i^*) \right) \ .$$

It is easily seen that $\left(\frac{1}{n} \sum g(x_i^*) g^\top(x_i^*)\right)$ converges in conditional probability to $\mathbb{E}(g(x)g(x)')$ and $n^{-3/4} \sum g(x_i^*) \left(h_\delta(\hat{\varepsilon}_i^*) - h_\delta(\varepsilon_i^*) \right)$ converges in conditional probability to zero, almost surely. ■

Proof of Theorem 4.1. The bootstrap estimator $\hat{\Omega}^{*-1}_{1/w}\hat{\Omega}^{*}_{v/w^2}\hat{\Omega}^{*-1}_{1/w}$ converges in conditional probability to $\Omega^{-1}_{1/w}\Omega_{v/w^2}\Omega^{-1}_{1/w}$. As a consequence of Theorem 2, the bootstrap distribution of $\sqrt{n}R(\beta^*_{WLS} - \hat{\beta}_{\text{WLS}})$ approximates the distribution of $\sqrt{n}(R\hat{\beta} - q)$. It follows that the bootstrap distribution of W^*_n consistently approximates the distribution of W_n. Moreover, both the bootstrap distribution of M^*_n and the sampling distribution of M_n are asymptotically distributed as $\max_i Z_i$ where Z is a multivariate normal random variable with mean zero and covariance matrix $V\Omega^{-1}_{1/w}\Omega_{v/w^2}\Omega^{-1}_{1/w}V$, with V a diagonal matrix whose diagonal entries are equal to the square root of the diagonal entries of $\Omega^{-1}_{1/w}\Omega_{v/w^2}\Omega^{-1}_{1/w}$. The claims of the theorem now follow from Slutsky's Theorem. ■

Proof of Theorem 4.2 and Lemma 4.1. These claims follow from the same arguments as the wild bootstrap counterparts, but with $\hat{\varepsilon}_i$ replaced by ε_i. ■

Proof of Theorem 5.1. For almost all sequences $\{(x_i, y_i)\}$, $\widehat{\text{Avar}}(\hat{\beta}_{\text{OLS},k})^*$ converges to $\text{Avar}(\hat{\beta}_{\text{OLS},k})$ and $\widehat{\text{Avar}}(\hat{\beta}_{\text{WLS},k})$ converges to $\text{Avar}(\hat{\beta}_{\text{WLS},k})$ in conditional probability. The claim follows from applying Slutsky's theorem conditionally. ■

Proof of Theorem 5.2. Following the argument of Theorem 3.1 of Romano and Wolf (2015), we must only find the asymptotic joint distribution of $\sqrt{n}(\hat{\beta}_W - \beta)$ and $\sqrt{n}(\hat{\beta}_{\text{OLS}} - \beta)$ since $\sqrt{n}(\hat{\beta}_{\text{WLS}} - \hat{\beta}_W) \xrightarrow{P} 0$. We can write $\sqrt{n}(\hat{\beta}_W - \beta) = \left(\frac{1}{n}X^\top W^{-1}X\right)^{-1}\frac{1}{\sqrt{n}}X^\top W^{-1}\varepsilon$ and $\sqrt{n}(\hat{\beta}_{\text{OLS}} - \beta) = \left(\frac{1}{n}X^\top X\right)^{-1}\frac{1}{\sqrt{n}}X^\top\varepsilon$. Because

$$\left(\frac{1}{n}X^\top W^{-1}X\right)^{-1} \xrightarrow{P} \mathbb{E}\left(\frac{1}{w(x_i)}x_i^\top x_i\right)^{-1} = \Omega^{-1}_{1/w} \ ,$$

and

$$\left(\frac{1}{n}X^\top X\right)^{-1} \xrightarrow{P} \mathbb{E}\left(x_i^\top x_i\right)^{-1} = \Omega^{-1}_{1/1} \ ,$$

it is enough to find the joint limiting distribution of $\frac{1}{\sqrt{n}}X^\top W^{-1}\varepsilon$ and $\frac{1}{\sqrt{n}}X^\top\varepsilon$. These are sums of i.i.d. mean zero random variables, so the Multivariate Central Limit Theorem gives

$$\sqrt{n}\begin{pmatrix} \frac{1}{\sqrt{n}}X^\top W^{-1}\varepsilon \\ \frac{1}{\sqrt{n}}X^\top\varepsilon \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbb{E}\left(x_i^\top x_i \frac{v(x_i)}{w^2(x_i)}\right) & \mathbb{E}\left(x_i^\top x_i \frac{v(x_i)}{w(x_i)}\right) \\ \mathbb{E}\left(x_i^\top x_i \frac{v(x_i)}{w(x_i)}\right) & \mathbb{E}\left(x_i^\top x_i v(x_i)\right) \end{pmatrix}\right) \ .$$

The claim follows from Slutsky's Theorem. ■

Proof of Theorem 5.3. An argument analogous to the proof of Theorem 3.1 to the one presented above shows that for any fixed λ, the bootstrap distribution of

$$\sqrt{n}(\lambda\hat{\beta}^*_{\text{WLS}} + (1-\lambda)\hat{\beta}^*_{\text{OLS}} - \lambda\hat{\beta}_{\text{WLS}} - (1-\lambda)\hat{\beta}_{\text{OLS}}) = \sqrt{n}(\hat{\beta}^*_\lambda - \hat{\beta}_\lambda) \ ,$$

is asymptotically normal with mean zero and covariance matrix $\text{Avar}(\hat{\beta}_\lambda)$ in probability.

It follows from the weak law of large numbers for triangular arrays that $\widehat{\text{Avar}}(\hat{\beta}_\lambda)^*$ converges in conditional probability to $\text{Avar}(\hat{\beta}_\lambda)$, almost surely. The second convergence follows from Slutsky's Theorem. ■

Proof of Theorem 5.4. We begin with the case where $\text{Avar}(\hat{\beta}_{\lambda,k})$ is non-constant. In order to show that $\sqrt{n}(\hat{\beta}_{\hat{\lambda}} - \beta) \xrightarrow{d} N(0, \text{Avar}(\hat{\beta}_{\lambda_0}))$, we will show that $\sqrt{n}(\hat{\beta}_{\hat{\lambda}_0} - \beta) - \sqrt{n}(\hat{\beta}_{\lambda_0} - \beta) \xrightarrow{P} 0$. Indeed,

$$\sqrt{n}\left(\hat{\beta}_{\hat{\lambda}_0} - \beta\right) - \sqrt{n}\left(\hat{\beta}_{\lambda_0} - \beta\right) = \sqrt{n}\left(\hat{\lambda}_0 - \lambda_0\right)\left[\hat{\beta}_{\text{OLS}} - \hat{\beta}_{\text{WLS}}\right]$$

which converges in probability to zero.

Theorem 5.3 gives that for any fixed λ, the bootstrap distribution of

$$\sqrt{n}(\lambda\hat{\beta}^*_{\text{WLS}} + (1-\lambda)\hat{\beta}^*_{\text{OLS}} - \lambda\hat{\beta}_{\text{WLS}} - (1-\lambda)\hat{\beta}_{\text{OLS}}) = \sqrt{n}(\hat{\beta}^*_\lambda - \hat{\beta}_\lambda)\ ,$$

is asymptotically normal with mean zero and covariance matrix $\text{Avar}(\hat{\beta}_\lambda)$ in conditional probability.

To prove the convergence of the bootstrap distribution stated in the theorem, we will first show that the bootstrap distribution of $\sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}^*} - \hat{\beta}_{\hat{\lambda}^*}\right)$ is asymptotically normal with mean 0 and covariance matrix $\text{Avar}(\hat{\beta}_\lambda)$ in probability and then show that $\sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}^*} - \hat{\beta}_{\hat{\lambda}}\right) - \sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}^*} - \hat{\beta}_{\hat{\lambda}^*}\right) \xrightarrow{p} 0$ in probability.

To show the desired asymptotic normality of $\sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}^*} - \hat{\beta}_{\hat{\lambda}^*}\right)$, we will show

$$\sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}_0} - \hat{\beta}_{\lambda_0}\right) - \sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}^*} - \hat{\beta}_{\hat{\lambda}^*}\right) \xrightarrow{P} 0\ .$$

We can write

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}_0} - \hat{\beta}_{\lambda_0}\right) - \sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}^*} - \hat{\beta}_{\hat{\lambda}^*}\right) =& \sqrt{n}(\hat{\lambda}^* - \lambda_0)\left[\hat{\beta}^*_{\text{WLS}} - \hat{\beta}_{\text{WLS}}\right] \\ &+ \sqrt{n}\left((1-\hat{\lambda}^*) - (1-\lambda_0)\right)\left[\hat{\beta}^*_{\text{OLS}} - \hat{\beta}_{\text{OLS}}\right]\ .\end{aligned}$$

Because $\sqrt{n}(\hat{\beta}^*_{WLS} - \hat{\beta}_{\text{WLS}})$ and $\sqrt{n}(\hat{\beta}^*_{\text{OLS}} - \hat{\beta}_{\text{OLS}})$ are asymptotically normal (in probability), the desired convergence follows from Slutsky's Theorem if we can show $\hat{\lambda}^* \xrightarrow{P} \lambda_0$. Note that $\hat{\lambda}^*$ is a continuous function of $\left[\hat{\Omega}^{*-1}_{1/w}\hat{\Omega}^*_{v/w^2}\hat{\Omega}^{*-1}_{1/w}\right]_{k,k}$, $\left[\hat{\Omega}^{*-1}_{1/w}\hat{\Omega}^*_{v/w}\hat{\Omega}^{*-1}_{1/1}\right]_{k,k}$, and $\left[\hat{\Omega}^{*-1}_{1/1}\hat{\Omega}^*_{v/1}\hat{\Omega}^{*-1}_{1/1}\right]_{k,k}$. Because these quantities converge in probability to the population versions almost surely, it follows from the continuous mapping theorem that $\hat{\lambda}^*$ converges in conditional probability to λ_0.

Similarly,

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}^*} - \hat{\beta}_{\hat{\lambda}^*}\right) - \sqrt{n}\left(\hat{\beta}^*_{\hat{\lambda}^*} - \hat{\beta}_{\hat{\lambda}}\right) =& \sqrt{n}\left(\hat{\beta}_{\hat{\lambda}^*} - \hat{\beta}_{\hat{\lambda}_0}\right) \\ =& \sqrt{n}(\hat{\lambda}^* - \hat{\lambda}_0)\left[\hat{\beta}^*_{\text{WLS}} - \hat{\beta}_{\text{WLS}}\right] \\ \xrightarrow{P}& 0\end{aligned}$$

in conditional probability.

The case where $\text{Avar}(\hat{\beta}_{\lambda,k})$ is similar, but follows from a simpler argument. ■

References

Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24:3–30.

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

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

Freedman, D. A. (1981). Bootstrapping regression models. *Annals of Statistics*, 9(6):1218–1228.

Godfrey, L. and Orne, C. (2004). Controlling the finite sample significance levels of heteroskedasticity-robust tests of several linear restrictions on regression coefficients. *Economics Letters*, 82:281–287.

Janssen, A. (1999). Nonparametric symmetry tests for statistical functionals. *Mathematical Methods of Statistics*, 8:320–343.

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

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, New York, third edition.

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

Romano, J. P. and Wolf, M. (2015). Resurrecting weighted least squares. Working Paper ECON 172, Department of Economics, University of Zurich.

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

Wooldridge, J. M. (2012). *Introductory Econometrics*. South-Western, Mason, Ohio, fifth edition.

Wu, C. F. J. (1986). Jacknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, 14:1261–1350.