The Canadian Journal of Statistics
Vol. 43, No. 3, 2015, Pages 358–377
La revue canadienne de statistique

Profiling heteroscedasticity in linear regression models

Qian M. ZHOU[1]*, Peter X.-K. SONG[2] and Mary E. THOMPSON[3]

[1]*Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada*
[2]*Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, U.S.A.*
[3]*Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada*

Key words and phrases: Heteroscedasticity; hybrid test; information ratio; linear regression models; model-based estimators; sandwich estimators; screening; weighted least squares.

MSC 2010: Primary 62J05; secondary 62J20

Abstract: Diagnostics for heteroscedasticity in linear regression models have been intensively investigated in the literature. However, limited attention has been paid on how to identify covariates associated with heteroscedastic error variances. This problem is critical in correctly modelling the variance structure in weighted least squares estimation, which leads to improved estimation efficiency. We propose covariate-specific statistics based on information ratios formed as comparisons between the model-based and sandwich variance estimators. A two-step diagnostic procedure is established, first to detect heteroscedasticity in error variances, and then to identify covariates the error variance structure might depend on. This proposed method is generalized to accommodate practical complications, such as when covariates associated with the heteroscedastic variances might not be associated with the mean structure of the response variable, or when strong correlation is present amongst covariates. The performance of the proposed method is assessed via a simulation study and is illustrated through a data analysis in which we show the importance of correct identification of covariates associated with the variance structure in estimation and inference. *The Canadian Journal of Statistics* 43: 358–377; 2015 © 2015 Statistical Society of Canada

Résumé: Les outils de diagnostic pour l'hétéroscédasticité dans les modèles de régression linéaire sont largement étudiés dans la littérature. Toutefois, l'identification des covariables associées aux variances hétéroscédastiques n'a suscité que peu d'intérêt. Ce problème joue pourtant un rôle clé pour l'estimation par les moindres carrés pondérés, puisque la modélisation correcte de la structure de variance accroît l'efficacité de l'estimation. Les auteurs proposent des statistiques spécifiques aux covariables fondées sur un ratio d'information comparant l'estimateur de la variance basé sur le modèle à l'estimateur sandwich de la variance. Ils développent une procédure diagnostique en deux étapes, détectant d'abord l'hétéroscédasticité et identifiant ensuite les covariables dont peut dépendre la structure de variance. Ils généralisent la méthode proposée afin d'accommoder des complications pratiques telles que l'absence de lien entre la structure de la moyenne et une covariable associée avec l'hétéroscédasticité, ou la forte corrélation des covariables. Les auteurs évaluent la performance de la méthode proposée à l'aide d'une étude de simulation et l'illustrent en analysant un jeu de données montrant l'importance d'identifier correctement les covariables associées avec la structure de variance pour l'estimation et l'inférence. *La revue canadienne de statistique* 43: 358–377; 2015 © 2015 Société statistique du Canada

1. INTRODUCTION

A standard assumption in linear regression models (LM) is homogeneity of error variances, also known as homoscedasticity. However, we often encounter violations to this assumption where

* *Author to whom correspondence may be addressed.*
E-mail: qmzhou@sfu.ca

error variances differ across subjects, and in some situations, are functions of one or multiple explanatory variables. Such violations are referred to as heteroscedasticity in the literature. In the presence of heteroscedasticity, ordinary least squares (OLS) estimation might be problematic. One of the problems is that the classic so-called "model-based" covariance matrix estimators are not consistent estimators for the variances of regression coefficient estimators; the resulting confidence intervals would not attain the nominal level of coverage. White (1980) proposed a consistent covariance matrix estimator, the "sandwich estimator" (Kauermann & Carroll, 2001), which is robust against heteroscedasticity. However, even with such heteroscedasticity-consistent covariance estimators, OLS estimation can still lead to loss of efficiency or low power in hypothesis testing.

Weighted least squares (WLS) estimation offers a remedy, where the weights incorporate the information about the error variances. Since the variance structure is usually unknown in practice, several iterative WLS estimation procedures were proposed where error variances were modelled as functions of covariates using parametric or nonparametric methods Harvey, 1976; Carroll & Ruppert, 1982, 1988; Müller & Stadtmuller, 1987; Mak, 1992; Verbyla, 1993; Müller & Zhao, 1995; Ruppert et al., 1997; Kuk, 1999. If the variance function is correctly modelled, the resulting iterative WLS estimators of regression coefficients will be asymptotically equivalent to the WLS estimators with known optimal weights (Mak, 1992). The first and critical step towards successful modelling of the error variance is to correctly identify which covariate or which set of covariates the variance function depends on. Some graphical procedures have been suggested, for example, plotting the OLS residuals against fitted values or against a covariate (Cook & Weisberg, 1982). However, visual judgement is mostly subjective and may work only for simple patterns. Numerous diagnostic tests for heteroscedasticity have been proposed in the literature (Anscombe, 1961; Goldfeld & Quandt, 1965; Park, 1966; Glejser, 1969; Bickel, 1978; Breusch & Pagan, 1979; White, 1980; Cook & Weisberg, 1983; Diblasi & Bowman, 1997; Zhou, Song, & Thompson, 2012). Some of these tests could be modified to detect covariates associated with variations in the error variance. However, most of the tests rely on specific functional forms for how the error variance depends on the covariates. In this paper, we provide a reliable diagnostic tool to identify the direction in the covariate space that allows for differences in the variance of response variable across subjects. We propose a two-step procedure for profiling heteroscedasticity: in the first step, we test whether the error variances are homogeneous, and if heteroscedasticity is detected, in the second step, we identify covariates on which the error variance function might depend. This construction is based on so-called information ratio (IR) statistics which are defined via comparisons between the model-based and sandwich estimators of the variances of individual regression coefficient estimators. The resulting statistics are called covariate-specific IR statistics. Compared with the existing approaches, this new method is advantageous in two aspects: (i) it does not depend on the underlying distribution of the error variables; and (ii) it does not require any specific forms of error variance functions. In addition, the proposed method can be easily implemented with low computational burden. Through simulation studies, we compare our method with some of the existing tests including White's test, Goldfeld–Quandt test and Breusch–Pagan–Godfrey test (Gujarati & Porter , 2009). The results show that our method outperforms these competitive methods in terms of controlling type I error and having high sensitivity.

The rest of the paper is organized as follows. In Section 2, we introduce covariate-specific IR statistics, and present a two-step procedure for examining the homoscedasticity assumption and further identifying covariates associated with the heteroscedastic variances. Several generalizations of our method are presented in Section 3. A simulation study and application of our method to a real data analysis are illustrated in Section 4. Section 5 includes concluding remarks and discussions. All technical details are included in the Appendix.

2. METHOD

Consider the following linear model: $Y = \beta_0 + \mathbf{X}^\mathsf{T}\boldsymbol{\beta} + e$ where Y is the response variable, $\mathbf{X} = (X_1, \ldots, X_p)^\mathsf{T}$ is a $p \times 1$ random vector of covariates with a compact support $\Omega_\mathbf{X} \subset \mathcal{R}^p$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\mathsf{T}$ is a $p \times 1$ vector of regression coefficients, and e is a random variable with mean 0 which is uncorrelated with $\mathbf{X}$. Let $\mathrm{E}_{Y|\mathbf{X}}$ (or $\mathrm{Var}_{Y|\mathbf{X}}$) denote expectations (or variances) with respect to (w.r.t.) the conditional distribution of Y given $\mathbf{X}$, let $\mathrm{E}_\mathbf{X}$ denote expectations w.r.t. the distribution of $\mathbf{X}$, and let $\mathrm{E}_{(Y,\mathbf{X})}$ (or $\mathrm{Var}_{(Y,\mathbf{X})}$) denote expectations (or variances) w.r.t. the joint distribution of $(Y, \mathbf{X})$.

In this section, we propose a two-step procedure: in the first step, we test the null hypothesis of homogeneity of error variances, namely

$$H_0 : \mathrm{Var}(e \mid \mathbf{X} = \mathbf{x}) = \mathrm{Var}(Y \mid \mathbf{X} = \mathbf{x}) \text{ is constant for any } \mathbf{x} \in \Omega_\mathbf{X}. \tag{1}$$

If the null hypothesis H_0 is rejected, in the second step, we detect which covariate $\mathrm{Var}(e \mid \mathbf{X} = \mathbf{x})$ might depend on. The test statistics are constructed under the homoscedasticity assumption based on comparisons between sandwich and model-based estimators for the variances of individual regression coefficient estimators. First, we present these two different types of variance estimator.

2.1. Sandwich and Model-Based Variance Estimators Under Homoscedasticity

Denote a data set by $\mathfrak{D} = \{(Y_i, \mathbf{X}_i), i = 1, \ldots, n\}$ consisting of n independent replicates of $(Y, \mathbf{X})$. To simplify the formulae, we consider the situation where the covariates have expectation 0 and there is no intercept in the model, so that the LM becomes $Y_i = \mathbf{X}_i^\mathsf{T}\boldsymbol{\beta} + e_i$. In implementing the procedures, we centre the responses and covariates at their sample means, so $\sum_{i=1}^n Y_i = 0$ and $\sum_{i=1}^n X_{ij} = 0$ for $j = 1, \ldots, p$. To estimate $\boldsymbol{\beta}$, we solve the following OLS estimating equation:

$$\sum_{i=1}^n \boldsymbol{\Psi}_i(\boldsymbol{\beta}) = \sum_{i=1}^n \frac{\mathbf{X}_i(Y_i - \mathbf{X}_i^\mathsf{T}\boldsymbol{\beta})}{\sigma^2} = \mathbf{0}, \tag{2}$$

where $\sigma^2 > 0$ is an arbitrary constant. This estimating equation, Equation (2), essentially involves only the conditional mean and variance of Y_i given $\mathbf{X}_i$, $\mathrm{E}_{Y|\mathbf{X}}(Y_i) = \mathbf{X}_i^\mathsf{T}\boldsymbol{\beta}$ and $\mathrm{Var}_{Y|\mathbf{X}}(Y_i) = \sigma^2$, instead of any fully specified distributions of Y_i conditional on $\mathbf{X}_i$. Let $\widehat{\boldsymbol{\beta}}$ be the OLS estimator of $\boldsymbol{\beta}$, which is the solution to the Equation (2). We assume (A1) $\widehat{\boldsymbol{\beta}}$ is consistent in the sense that $\widehat{\boldsymbol{\beta}}$ converges in probability to a vector $\boldsymbol{\beta}^*$ which is the solution to $\mathrm{E}_{Y|\mathbf{X}}\{\boldsymbol{\Psi}_i(\boldsymbol{\beta})\} = \mathbf{0}$. If conditional mean structure $\mathrm{E}_{Y|\mathbf{X}}(Y_i) = \mathbf{X}_i^\mathsf{T}\boldsymbol{\beta}$ is correctly specified, $\boldsymbol{\beta}^*$ is the true value of $\boldsymbol{\beta}$.

Now we present the sandwich and model-based estimators of the asymptotic variance of $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$. Let $\mathcal{X}$ be an $n \times p$ matrix with $\mathbf{X}_i^T$ as its ith row. The estimator $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ converges in distribution to a normal random variable with mean 0, and its asymptotic covariance matrix can be estimated by the so-called "heteroscedasticity-consistent" covariance matrix estimator proposed by White (1980),

$$\mathfrak{S}_{\widehat{\boldsymbol{\beta}}} = \mathbf{S}_n^{-1}\mathbf{V}_n\mathbf{S}_n^{-1},\ \mathbf{S}_n = n^{-1}\mathcal{X}^\mathsf{T}\mathcal{X},\ \mathbf{V}_n = n^{-1}\mathcal{X}^\mathsf{T}\mathcal{R}\mathcal{X}, \tag{3}$$

where $\mathcal{R} = \mathrm{diag}\{r_1^2, \ldots, r_n^2\}$ with residuals $r_i = Y_i - \mathbf{X}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}$. This is also called "sandwich" covariance matrix estimator. On the other hand, under the null hypothesis of homoscedasticity, the asymptotic variance of $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ is estimated by $\mathfrak{M}_{\widehat{\boldsymbol{\beta}}} = \sigma^2\{n^{-1}\mathcal{X}^\mathsf{T}\mathcal{X}\}^{-1}$, the so-called "model-based" estimator. In practice, the unknown parameter σ^2 may be replaced with a consistent estimator, for example, $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n r_i^2$. Thus, the model-based covariance matrix estimator

with estimated σ^2 is given by

$$\mathfrak{M}_{\widehat{\beta}} = \widehat{\sigma}^2 \{n^{-1}\mathcal{X}^\intercal \mathcal{X}\}^{-1}. \tag{4}$$

The diagonal elements of $\mathfrak{S}_{\widehat{\beta}}$ and $\mathfrak{M}_{\widehat{\beta}}$, denoted by $\mathcal{V}_j^s$ and $\mathcal{V}_j^m$, are the sandwich and model-based estimators for the variance of individual regression coefficient estimator $\sqrt{n}(\widehat{\beta}_j - \beta_j^*)$, respectively, $j = 1, \ldots, p$. If the homoscedasticity assumption is true, the model-based estimator is asymptotically equivalent to the sandwich estimator; otherwise, there exist discrepancies between these two types of variance estimators. In the next subsection, we propose covariate-specific IR statistics by comparing $\mathcal{V}_j^s$ with $\mathcal{V}_j^m$, $j = 1, \ldots, p$.

In addition to the assumption (A1), we also need to make the following assumptions: (A2) $X_{i1}, \ldots, X_{ip}$ are uncorrelated, (A3) $\mathrm{E}_{\mathbf{X}}(X_{ij}^2 \mid X_{ik}, k \neq j) = E(X_{ij}^2)$ for all $j = 1, \ldots, p$, that is, the variance of each covariate does not depend on other covariates, (A4) $\mathrm{E}_{Y|\mathbf{X}}(Y_i^4)$ is bounded.

2.2. Covariate-Specific IR Statistics

Define X_j-specific IR statistic as $\mathrm{IR}_j = \mathcal{V}_j^s/\mathcal{V}_j^m$ with $\mathcal{V}_j^m \neq 0$. In Lemma 1, we show that this IR statistic can be expressed as a weighted quadratic form in residuals r_i. Let $\mathcal{H} = \mathcal{X}\left(\mathcal{X}^\intercal\mathcal{X}\right)^{-1}\mathcal{X}^\intercal$, and $\mathcal{H}_{(-j)} = \mathcal{X}_{(-j)}\left(\mathcal{X}_{(-j)}^\intercal\mathcal{X}_{(-j)}\right)^{-1}\mathcal{X}_{(-j)}^\intercal$, where $\mathcal{X}_{(-j)}$ is the matrix resulting from deleting the jth column of the matrix $\mathcal{X}$.

Lemma 1. *For $j = 1, \ldots, p$, the X_j-specific IR statistic can be expressed as follows:*

$$\mathrm{IR}_j = \sum_{i=1}^{n} \frac{w_i^{(j)} r_i^2}{\widehat{\sigma}^2},$$

where $w_i^{(j)} = h_{ii} - h_{ii}^{(-j)}$, h_{ii} is the ith diagonal element of $\mathcal{H}$, and $h_{ii}^{(-j)}$ is the ith diagonal element of $\mathcal{H}_{(-j)}$.

Note that $\sum_{i=1}^n w_i^{(j)} = 1$. The proof of Lemma 1 is given in the Appendix. In the following lemma, we establish that the weights are stochastically equivalent to a quantity which depends only on X_{ij}.

Lemma 2. *Let $A_j = \mathrm{E}_{\mathbf{X}}(X_{ij}^2) > 0$ and $\bar{w}_i^{(j)} = n^{-1}X_{ij}^2/A_j$. Then for each j, $\sup_i |w_i^{(j)} - \bar{w}_i^{(j)}| = o_p(n^{-1})$, where $w_i^{(j)}$ is given in Lemma 1.*

The proof of Lemma 2 is given in the Appendix. In the following Theorem 1, we establish the asymptotic distribution of the covariate-specific IR statistics under the null hypothesis, Equation (1). Let us introduce some notation.

Let $e_i = Y_i - \mathbf{X}_i^\intercal\boldsymbol{\beta}^*$. Under the assumption (A1), $\mathrm{E}_{Y|\mathbf{X}}(e_i) = 0$. Let $\nu^2 = \mathrm{E}_{(Y,\mathbf{X})}(e_i^2)$ and $\nu_j^2 = \mathrm{E}_{(Y,\mathbf{X})}(e_i^2X_{ij}^2)/A_j$. Note that neither ν^2 nor ν_j^2 depends on i through the operation of expectation w.r.t. the joint distribution of $(Y_i, \mathbf{X}_i)$. In the Appendix, we also show that $\mathcal{W}_j = \sqrt{n}(\mathrm{IR}_j - 1) = \sqrt{n}(\nu_j/\nu - 1) + n^{-1/2}\sum_{i=1}^n \mathcal{R}_{ij} + o_p(1)$ where $\mathcal{R}_{ij}$ are i.i.d. random variables with mean 0. Suppose that a general form of the conditional variance of Y_i given $\mathbf{X}_i$ is given by

$$\mathrm{Var}_{Y|\mathbf{X}}(Y_i) = \mathrm{E}_{Y|\mathbf{X}}(e_i^2) = \kappa^2 V(\mathbf{X}_i), \tag{5}$$

where κ^2 is an arbitrary constant and $V(\cdot)$ is a certain smooth function. Under the null hypothesis, Equation (1), $V(\mathbf{X}_i) \equiv 1$ and $\nu_j^2 = \nu^2 = \kappa^2$, implying $\mathrm{E}(\mathcal{W}_j) = o(1)$. The asymptotic distribution of $\mathcal{W}_j$ under the null hypothesis is given in the following.

Theorem 1. *Suppose that Assumptions (A1)–(A4) are satisfied. Under the null hypothesis, Equation (1), for all $j = 1, \ldots, p$, the transformed X_j-specific IR test statistic $\mathcal{W}_j = \sqrt{n}(\mathrm{IR}_j - 1)$ converges in distribution to a Gaussian random variable with mean 0 and variance*

$$\Sigma_j = E_{(Y,\mathbf{X})}\left[\left\{\left(\frac{X_{ij}^2}{A_j} - \frac{\nu_j^2}{\nu^2}\right)\left(\frac{e_i^2}{\nu^2} - 1\right) - \left(\frac{\nu_j^2}{\nu^2} - 1\right)\right\}^2\right]. \tag{6}$$

The proof of Theorem 1 is given in the Appendix. It is worth pointing out that the asymptotic variance, Equation (6), is obtained under a general variance structure given in Equation (5). However, its analytic closed form is not available because of the involvement of the fourth moment $\mathrm{E}_{Y|\mathbf{X}}(Y_i^4)$. The explicit form of this fourth moment, in general, is unknown. Following Zhou, Song, & Thompson (2012), we suggested a perturbation resampling method to obtain Σ_j numerically. Such a perturbation method, similar to the so-called wild bootstrap (Wu, 1986; Mammen, 1992), has been widely used in survival analysis (see, e.g., Park & Wei, 2003; Cai, Tian, & Wei, 2005) in a situation where the asymptotic variances are difficult to calculate. Specifically, in the implementation, we define the perturbed counterpart of the test statistic $\mathcal{W}_j$ as

$$\mathcal{W}_j^* = \sqrt{n}\sum_{i=1}^{n}\left\{\left(w_i^{(j)} - \frac{1}{n}\widetilde{\mathrm{IR}}_j\right)\left(\frac{r_i^2}{\widehat{\sigma}^2} - 1\right) - \frac{1}{n}\left(\widetilde{\mathrm{IR}}_j - 1\right)\right\}\xi_i, \tag{7}$$

where $\widetilde{\mathrm{IR}}_j$ is the observed value of the statistic IR_j given the data set $\mathfrak{D} = \{(Y_i, \mathbf{X}_i), i = 1, \ldots, n\}$, and $\Xi = \{\xi_1, \ldots, \xi_n\}$ is a set of i.i.d. random variables with mean 0 and variance 1, for example, random variables from a standard normal distribution. In addition. Ξ is independent of $\mathfrak{D}$. Conditional on the data $\mathfrak{D}$, every term in Equation (7) is fixed except that Ξ are random variables, and consequently, $\mathrm{E}(\mathcal{W}_j^* \mid \mathfrak{D}) = 0$ and

$$\mathrm{Var}(\mathcal{W}_j^* \mid \mathfrak{D}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(nw_i^{(j)} - \widetilde{\mathrm{IR}}_j\right)\left(\frac{r_i^2}{\widehat{\sigma}^2} - 1\right) - \left(\widetilde{\mathrm{IR}}_j - 1\right)\right\}^2.$$

According to Lemma 2, $\sup_i \left|w_i^{(j)} - n^{-1}X_{ij}^2/A_j\right| = o_p(n^{-1})$, and executing similar arguments in the proof of Theorem 1, we can show that $\mathrm{Var}(\mathcal{W}_j^* \mid \mathfrak{D})$ converges in probability to Σ_j as $n \to \infty$. With B perturbed samples of $\mathcal{W}_j$, denoted by $\{\mathcal{W}_j^{*(b)}, b = 1, \ldots, B\}$, the empirical X_j-specific P-value, denoted by $\mathcal{P}_{X_j}$, is $B^{-1}\sum_{b=1}^{B} I\{\mathcal{W}_j^{*(b)} \geq \widetilde{W}_j\}$, where $\widetilde{W}_j$ is the observed statistic $\mathcal{W}_j$ based on the actual data.

When using these covariate-specific IR test statistics together to test the null hypothesis H_0, Equation (1), to control overall type I error, a Bonferroni correction may be adopted to adjust for multiplicity of the testing. Specifically, given a significance level α, the rejection rule is

$$\mathfrak{C}_{X_j}: \text{ the null hypothesis is rejected if } \mathcal{P}_{\min,\mathrm{X}} = \min\{\mathcal{P}_{X_1}, \ldots, \mathcal{P}_{X_p}\} \leq \alpha/p. \tag{8}$$

2.3. Pooled IR Statistics

The asymptotic equivalence between the sandwich and model-based variance estimators can be explained by the information unbiasedness in the theory of estimating functions. Given the estimating function $\Psi_i(\boldsymbol{\beta})$ in Equation (2), $\lim_{n\to\infty}\mathbf{S}_n = \mathrm{E}_{\boldsymbol{\theta}^*}\{-\partial\boldsymbol{\Psi}_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\}$ and $\lim_{n\to\infty}\mathbf{V}_n = \mathrm{Cov}_{\boldsymbol{\theta}^*}\{\boldsymbol{\Psi}_i(\boldsymbol{\beta})\}$ are called the *sensitivity* and *variability* information matrices associated with the parameter $\boldsymbol{\beta}$, respectively. When the conditional mean and variance of the response variable in

the estimating function are correctly specified, the sensitivity matrix is equal to the variability matrix. This equality is referred to as information equivalence (White, 1982) or information unbiasedness (Lindsay, 1982). Zhou, Song, & Thompson (2012) proposed an IR statistic based on a multiplicative comparison between the sensitivity matrix and variability matrix, given as

$$\mathrm{IR}_{\mathrm{pool}} = \frac{\mathrm{tr}\{\mathbf{S}_n^{-1}\mathbf{V}_n\}}{p} = \frac{\mathrm{tr}\{\mathfrak{M}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathfrak{S}_{\widehat{\boldsymbol{\beta}}}\}}{p}, \tag{9}$$

where $\mathrm{tr}(\cdot)$ is the trace of a matrix. Equation (9) is in fact equivalent to a comparison between the sandwich and model-based covariance matrix estimators. This IR statistic can also be written as a weighted quadratic form in residuals r_i, namely $\mathrm{IR}_{\mathrm{pool}} = \sum_{i=1}^{n} r_i^2 w_i^{\mathrm{pool}}/\widehat{\sigma}^2$, where $w_i^{\mathrm{pool}} = h_{ii}/p$ contains the information pooled from all the covariates, and so it is called the pooled IR statistic. Under homogeneity of error variances, Zhou, Song, & Thompson (2012) showed that the IR statistic $\sqrt{n}(\mathrm{IR}_{\mathrm{pool}} - 1)$ is asymptotically distributed as a normal random variable with mean 0, and its asymptotic variance can be numerically approximated using perturbation resampling. Let $\mathcal{P}_{\mathrm{pool}}$ be the P-value of the pooled IR test, and the rejection rule for testing the null hypothesis H_0, Equation (1), using the pooled IR test alone is

$$\mathfrak{C}_{\mathrm{pool}}: \text{ the null hypothesis is rejected if } \mathcal{P}_{\mathrm{pool}} \leq \alpha. \tag{10}$$

This method has been shown to have greater test power than both the information matrix test of White (1980) and the "in-and-out-of-sample" likelihood ratio test of Presnell & Boos (2004), in addition to a satisfactory type-I error control.

2.4. Two-Step Procedure of Profiling Heteroscedasticity

The purpose of the first step is to detect any form of heteroscedasticity, and either covariate-specific IR tests with Bonferroni adjustment, Equation (8), or the pooled IR test, Equation (10), can serve for this purpose. Some numeric evidence in the simulation study in Section 4.1 suggested that if the error variances depend on only one covariate, the covariate-specific IR tests appeared more sensitive to detect heteroscedasticity than the pooled IR test; on the other hand, if the error variances depend on more than one covariate, the pooled IR test appeared to perform better. In practice, given the fact that the form of the error variance is never known, it seems natural to combine these two tests to reach a compromise. This combination results in a "hybrid" test, denoted by $\mathfrak{C}_{\mathrm{hybrid}}$. Specifically, define a P-value $\mathcal{P}_{\mathrm{hybrid}} = 2\min\{\mathcal{P}_{\mathrm{pool}}, p\mathcal{P}_{\mathrm{min,X}}\}$, and the rejection rule is

$$\mathfrak{C}_{\mathrm{hybrid}}: \text{ the null hypothesis is rejected if } \mathcal{P}_{\mathrm{hybrid}} \leq \alpha. \tag{11}$$

This hybrid strategy helps to control overall type I error. This is because under the null hypothesis, $Pr(\mathcal{P}_{\mathrm{hybrid}} \leq \alpha) \leq Pr(\mathcal{P}_{\mathrm{pool}} \leq \alpha/2) + Pr\left\{\mathcal{P}_{\mathrm{min,X}} \leq (\alpha/2)/p\right\} \leq \alpha/2 + \alpha/2 = \alpha$, meaning that the probability of type I error is bounded by α.

If the null hypothesis is rejected, the second step needs to be carried out to identify covariates associated with the heteroscedasticity. Adjusting for multiple testing, the covariates with corresponding covariate-specific IR P-values less than α/p can be regarded as the covariates associated with error variances. This can be justified by the asymptotic distribution of the covariate-specific IR statistics under heteroscedastic variance structures. Consider the variance structure Equation (5) depending on $\mathbf{Z}_i$, a sub-vector of $\mathbf{X}_i$, say $\mathbf{Z} = (X_{i,j_1}, \ldots, X_{i,j_q})$, where $\mathcal{J} = \{j_1, \ldots, j_q\} \subset \{1, \ldots, p\}$. This leads to $v^2 = \kappa^2 \mathrm{E}_{\mathbf{X}}\{V(\mathbf{Z}_i)\}$ and $v_j^2 = \kappa^2 \mathrm{E}_{\mathbf{X}}\left\{V(\mathbf{Z}_i)X_{ij}^2\right\}/A_j$. If the covariate X_{ij} is not included in $\mathbf{Z}_i$, that is, $j \notin \mathcal{J}$, due to Assumption (A3) $\mathrm{E}_{\mathbf{X}}(X_{ij}^2 \mid X_{ik}, k \neq j) = \mathrm{E}_{\mathbf{X}}(X_{ij}^2)$,

$\mathrm{E}_{\mathbf{X}}\left\{V(\mathbf{Z}_i)X_{ij}^2\right\}/A_j = \mathrm{E}_{\mathbf{X}}\{V(\mathbf{Z}_i)\}\,\mathrm{E}_{\mathbf{X}}(X_{ij}^2)/A_j = \mathrm{E}_{\mathbf{X}}\{V(\mathbf{Z}_i)\}$, and consequently, $\nu_j^2 = \nu^2$. The X_j-specific IR test statistic $\mathcal{W}_j$ converges in distribution to a Gaussian random variable with mean 0. Given a significance level α, the false positive rate for individual covariate, that is, the probability of wrongly identifying X_{ij} as the covariate associated with error variances is at most α/p. If the covariate X_{ij} is included in $\mathbf{Z}_i$, that is, $j \in \mathcal{J}$, the statistic $\mathrm{IR}_j - 1$ converges in distribution to a random variable with mean $\nu_j^2/\nu^2 - 1$. This leads to a high probability that X_{ij} is chosen. Larger sample sizes and/or larger distance of ν_j^2/ν^2 from 1 leads to higher sensitivity. Note that when the covariates are approximately normally distributed, Assumption (A2) would imply Assumption (A3). However, for other situations, violation of Assumption (A3) might lead to a false positive identification. For example, if the variance of X_{ij} outside $\mathbf{Z}_i$ depends on some component in $\mathbf{Z}_i$, the X_j-specific IR test statistic might converge in distribution to a Gaussian random variable with a nonzero mean, which might lead to a small P-value, even though X_{ij} is not associated with the variance structure. Thus we recommend to check Assumption (A3) only on the covariates which are identified as associated with the variance structure.

3. GENERALIZATIONS OF COVARIATE-SPECIFIC IR STATISTICS

In some applications, some covariates may explain the differences in the variance of the response variable across subjects rather than the differences in the mean of the response variable (Zalesny & Farace , 1986). Our method may be extended to a more general situation where a set of m explanatory variables $\mathbf{C}_i = (C_{i1}, \ldots, C_{im})^{\mathsf{T}}$ are considered as candidate covariates for the error variances, and variables in $\mathbf{C}_i$ are possibly not in $\mathbf{X}_i$. Let $\mathbb{C}$ be an $n \times m$ matrix with $\mathbf{C}_i^{\mathsf{T}}$ as the ith row, and let r_i be the OLS residuals from regressing Y_i on $\mathbf{X}_i$. Based on this candidate set, the pooled IR statistic and covariate-specific IR statistics are given as

$$\mathrm{IR}_{\mathrm{pool}} = \sum_{i=1}^{n} r_i^2 w_i^{\mathrm{pool}}/\widehat{\sigma}^2 \quad \text{and} \quad \mathrm{IR}_j = \sum_{i=1}^{n} r_i^2 w_i^{(j)}/\widehat{\sigma}^2, \tag{12}$$

where $w_i^{\mathrm{pool}} = c_{ii}/m$ with c_{ii} being the diagonal elements of $\mathbb{C}\left(\mathbb{C}^{\mathsf{T}}\mathbb{C}\right)^{-1}\mathbb{C}$, and $w_i^{(j)} = c_{ii} - c_{ii}^{(-j)}$ with $c_{ii}^{(-j)}$ being the diagonal element of $\mathbb{C}_{(-j)}\left(\mathbb{C}_{(-j)}^{\mathsf{T}}\mathbb{C}_{(-j)}\right)^{-1}\mathbb{C}_{(-j)}$ and $\mathbb{C}_{(-j)}$ being the matrix resulting from deleting the jth column from the matrix $\mathbb{C}$. Note that if $\mathbf{C}_i \equiv \mathbf{X}_i$, these IR statistics in Equation (12) are the same as those in Lemma 1 and Equation (9). Details on the derivation of Equation (12) are provided in the Appendix.

The covariate-specific IR statistics may be extended to covariate-set-specific statistics. Specifically, consider an index set $\mathcal{J}_k \subset \{1, \ldots, p\}$ with dimension l_k. Let $\mathbf{C}_{i,\mathcal{J}_k} = \{C_{ij} : j \in \mathcal{J}_k\}$ denote an l_k-dimensional sub-vector of $\mathbf{C}_i$ and let $\mathbf{C}_{i,(-\mathcal{J}_k)} = \{C_{ij} : j \notin \mathcal{J}_k\}$. The covariate set $\mathbf{C}_{i,\mathcal{J}_k}$-specific IR statistic is given as $\mathrm{IR}_{\mathcal{J}_k} = \sum_{i=1}^{n} r_i^2 w_i^{(\mathcal{J}_k)}/\widehat{\sigma}^2$, $w_i^{(\mathcal{J}_k)} = c_{ii} - c_{ii}^{(-\mathcal{J}_k)}$ where $c_{ii}^{(-\mathcal{J}_k)}$ are the diagonal elements of $\mathbb{C}_{(-\mathcal{J}_k)}\left(\mathbb{C}_{(-\mathcal{J}_k)}^{\mathsf{T}}\mathbb{C}_{(-\mathcal{J}_k)}\right)^{-1}\mathbb{C}_{(-\mathcal{J}_k)}$ with $\mathbb{C}_{(-\mathcal{J}_k)}$ being the resulting matrix with the column set $\mathcal{J}_k$ deleted from $\mathbb{C}$. If any element in $\mathbf{C}_{i,\mathcal{J}_k}$ is uncorrelated with any element in $\mathbf{C}_{i,(-\mathcal{J}_k)}$, the weights $w_i^{(\mathcal{J}_k)}$ are stochastically equivalent to $n^{-1}\mathbf{C}_{i,\mathcal{J}_k}{}^{\mathsf{T}} A_{\mathcal{J}_k}^{-1}\mathbf{C}_{i,\mathcal{J}_k}$ where $A_{\mathcal{J}_k} = \mathrm{E}\left(\mathbf{C}_{i,\mathcal{J}_k}\mathbf{C}_{i,\mathcal{J}_k}{}^{\mathsf{T}}\right)$.

In practice, under the assumption (A2) that $X_{i1}, \ldots, X_{ip}$ are uncorrelated, is a strong assumption. For experimental data, the covariate space could be designed as orthogonal, and the proposed procedure can be applied directly. For observational data, the correlation structure amongst covariates could be learned from the data, such as a sample correlation matrix. The proposed procedure might work adequately for the situations when the sample correlation between covariates is modest. However, in situations where strong correlation is present, the weight $w_i^{(j)}$ of the C_j-specific IR statistic would converge to a quantity which does not depend on C_{ij} alone. In

using the covariate-specific IR statistics, the true dependence of error variances on some covariates might be masked or weakened by other covariates which are correlated with these covariates truly associated with the error variances. We suggest a screening mechanism which successively examines each covariate $\mathbb{C}_j = (C_{1j}, \ldots, C_{nj})^{\mathsf{T}}$, $j = 1, \ldots, p$ as the "target" covariate. Starting with $\mathbb{C}_1$, we transform $\mathbb{C}$ to $\mathbb{C}^{\dagger} = (\mathbb{C}_1, \Delta)$ where Δ is an $n \times (m-1)$ matrix whose column vectors are perpendicular to $\mathbb{C}_1$. This can be done by regressing each of the column vectors of the matrix $\mathbb{C}_{(-1)}$ on $\mathbb{C}_1$, and taking the resulting residuals, say $\{\boldsymbol{\delta}_j, j = 2, \ldots, m\}$, to form the matrix Δ. Based on the new design matrix $\mathbb{C}^{\dagger}$, we construct the C_1-specific IR statistic. Note that $\mathrm{IR}_{\mathrm{pool}}$ based on $\mathbb{C}^{\dagger}$ is same as the one based on $\mathbb{C}$. Similarly, we obtain the P-values for the C_j-specific IR tests for $j = 2, \ldots, m$. This algorithm allows the true relationship of each covariate with the error variances to be singled out.

4. NUMERICAL ILLUSTRATION

4.1. Simulation Study

Consider the model: $Y_i = 1 + X_{i1} + X_{i2} + X_{i3} + e_i$ where the variance σ_i^2 of e_i takes the form of $g(\sigma_i^2) = \phi_0 + \mathbf{W}_i^{\mathsf{T}}\boldsymbol{\phi}$, where $g(\cdot)$ is a known smooth function, $\mathbf{W}_i$ is a $q \times 1$ vector function of a subset of $\{X_{i1}, X_{i2}, X_{i3}\}$, and $\boldsymbol{\phi}$ is a $q \times 1$ vector. Two families of error variance functions considered in Cook & Weisberg (1983) are examples of these forms: $\log(\sigma_i^2) = \sum_j \phi_j X_{ij}$ and $\log(\sigma_i^2) = \sum_j \phi_j \log(X_{ij})$ assuming $X_{ij} > 0$. Both the variability of covariates and the magnitude of $\boldsymbol{\phi}$ affect heterogeneity across σ_i^2, where $\boldsymbol{\phi}$ is called the *effect size*. For simplicity, we generate uncorrelated covariates with variance 1 and vary the values of $\boldsymbol{\phi}$. Specifically, we generate three independent sets $(\widetilde{X}_{1j}, \ldots, \widetilde{X}_{nj})$, $j = 1, 2, 3$, of random variables from a uniform distribution between 0 and 1. Let $\bar{x}_j$ and s_j be the sample mean and sample standard deviation of $\{\widetilde{X}_{1j}, \ldots, \widetilde{X}_{nj}\}$, respectively, and let $X_{ij} = (\widetilde{X}_{ij} - \bar{x}_j)/s_j$. The covariates $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3})$ are used to generate the responses Y_i. In the simulation, we generate the errors e_i from a normal distribution with mean 0 and homogeneous error variance $H_0 : \sigma_i^2 = 1$ as well as the following heteroscedastic variance structures $\sigma_i^2 = V(\mathbf{X}_i)$:

(i) σ_i^2 depends on $\mathbf{X}_{i1}$ only: $1 + X_{i1}^2$ and $1 + 3X_{i1}^2$;
(ii) σ_i^2 depends on both $\mathbf{X}_{i1}$ and $\mathbf{X}_{i2}$: $1 + 3X_{i1}^2 + X_{i2}^2$ and $1 + 3X_{i1}^2 + 3X_{i2}^2$;
(iii) σ_i^2 depends on all the covariates: $1 + 3X_{i1}^2 + 3X_{i2}^2 + 3X_{i3}^2$;
(iv) σ_i^2 depends on covariates in an exponential functional form: $\exp(3X_{i1})$ and $\exp(3X_{i1} + 3X_{i2})$.

The sample size is set to be 200 and 400. All the simulation results are obtained based on 10,000 replications for each variance structure and each sample size; to implement the perturbation resampling method, 1,000 perturbed resamples are generated from each data set and we generate ξ_i from a standard Normal distribution.

We investigate three tests: (i) the pooled IR test alone, which is based on $\mathfrak{C}_{\mathrm{pool}}$ in Equation (10); (ii) covariate-specific IR tests with Bonferroni adjustment, which is based on $\mathfrak{C}_{X_j}$ in Equation (8); and (iii) the hybrid test, which is based on $\mathfrak{C}_{\mathrm{hybrid}}$ in Equation (11). The empirical frequencies of rejecting the null hypothesis using these three tests at the significance level $\alpha = 0.05$ under each of the variance structures are reported in Tables 1 and 2. First, under the null hypothesis, by McNemar's test, we confirm that the empirical type I errors of the hybrid test are significantly different from those of the other two tests since the P-values of the McNemar's test are all less than 0.001. The empirical type I errors of $\mathfrak{C}_{\mathrm{pool}}$ are generally inflated whilest those of $\mathfrak{C}_{X_j}$ are conservative with type I error rate less than α. The empirical type I errors of $\mathfrak{C}_{\mathrm{hybrid}}$ are closer to the nominal level compared with $\mathfrak{C}_{\mathrm{pool}}$ and $\mathfrak{C}_{X_j}$. Secondly, when the error variances depend on only one covariate, for example $\sigma_i^2 = 1 + X_{i1}^2$, the rejection criterion $\mathfrak{C}_{X_j}$ is more powerful than

TABLE 1: Empirical relative frequencies (multiplied by 100) of rejecting the null hypothesis amongst 10,000 replicates under various variance structures when the sample size is 200.

	$\mathfrak{C}_{\text{pool}}$	$\mathfrak{C}_{X_j}$	$\mathfrak{C}_{\text{hybrid}}$	White	GQ	BPG
H_0	6.3	3.9	5.0	3.4	14.0	4.7
$1 + X_{i1}^2$	60.6	95.2	91.9	76.8	19.6	18.1
$1 + 3X_{i1}^2$	89.4	100.0	99.9	98.7	25.2	30.0
$\exp(3X_{i1})$	89.5	100.0	100.0	100.0	46.4	100.0
$1 + 3X_{i1}^2 + X_{i2}^2$	95.3	99.3	99.2	94.2	22.3	26.0
$1 + 3X_{i1}^2 + 3X_{i2}^2$	97.9	97.7	98.3	92.8	21.5	26.0
$\exp(3X_{i1} + 3X_{i2})$	98.8	98.3	99.3	100.0	38.6	100.0
$1 + 3X_{i1}^2 + 3X_{i2}^2 + 3X_{i3}^2$	99.2	89.2	98.5	84.8	19.7	23.5

$\mathfrak{C}_{\text{pool}}$ represents the pooled IR test alone; $\mathfrak{C}_{X_j}$ represents the covariate-specific IR tests with the Bonferroni correction; $\mathfrak{C}_{\text{hybrid}}$ represents the hybrid test. "White" represents the White's test; "GQ" represents the Goldfeld–Quandt test; "BPG" represents the Breusch–Pagan–Godfrey test.

the criterion $\mathfrak{C}_{\text{pool}}$ despite of being conservative. When the error variances depend on more than one covariate, the rejection criterion $\mathfrak{C}_{\text{pool}}$ appears more powerful but may be due to its inflated type I errors. In both of these two scenarios, our proposed hybrid test $\mathfrak{C}_{\text{hybrid}}$ in most of the cases outperforms the other two criteria $\mathfrak{C}_{\text{pool}}$ and $\mathfrak{C}_{X_j}$. Thirdly, the test power based on all the three criteria increases for the larger sample size.

To compare our method with some of the existing methods, we modify the White's test, Goldfeld–Quand (GQ) test, and Breusch–Pagan–Godfrey (BPG) test suitable to test for homoscedasticity and to identify the covariates associated with the error variances. The details of these three tests can be found in the Appendix, and their empirical frequencies of rejecting the null hypothesis are also listed in Tables 1 and 2. The results show that amongst these three tests, the empirical type I errors of the BPG test are closest to the nominal level. Compared with the

TABLE 2: Empirical relative frequencies (multiplied by 100) of rejecting the null hypothesis amongst 10,000 replicates under various variance structures when the sample size is 400.

	$\mathfrak{C}_{\text{pool}}$	$\mathfrak{C}_{X_j}$	$\mathfrak{C}_{\text{hybrid}}$	White	GQ	BPG
H_0	5.7	3.8	4.5	3.2	14.7	5.0
$1 + X_{i1}^2$	90.5	100.0	99.9	99.0	20.5	18.9
$1 + 3X_{i1}^2$	99.6	100.0	100.0	100.0	25.9	30.8
$\exp(3X_{i1})$	99.3	100.0	100.0	100.0	46.4	100.0
$1 + 3X_{i1}^2 + X_{i2}^2$	100.0	100.0	100.0	100.0	23.2	27.0
$1 + 3X_{i1}^2 + 3X_{i2}^2$	100.0	100.0	100.0	100.0	22.0	26.8
$\exp(3X_{i1} + 3X_{i2})$	100.0	100.0	100.0	100.0	38.5	100.0
$1 + 3X_{i1}^2 + 3X_{i2}^2 + 3X_{i3}^2$	100.0	99.8	100.0	99.6	20.3	24.0

$\mathfrak{C}_{\text{pool}}$ represents the pooled IR test alone; $\mathfrak{C}_{X_j}$ represents the covariate-specific IR tests with the Bonferroni correction; $\mathfrak{C}_{\text{hybrid}}$ represents the hybrid test. "White" represents the White's test; "GQ" represents the Goldfeld–Quandt test; "BPG" represents the Breusch–Pagan–Godfrey test.

TABLE 3: Empirical relative frequencies (multiplied by 100) of identifying each of the covariates as a covariate associated with the error variances when the sample size is 200.

	X_1			X_2			X_3		
	IR	White	BPG	IR	White	BPG	IR	White	BPG
$1 + X_{i1}^2$	91.5	6.2	19.0	1.4	1.4	8.4	1.3	1.5	8.4
$1 + 3X_{i1}^2$	99.9	8.3	28.0	1.4	1.4	12.8	1.3	1.4	12.3
$\exp(3X_{i1})$	99.9	100.0	100.0	2.3	1.1	44.3	2.1	1.0	44.5
$1 + 3X_{i1}^2 + X_{i2}^2$	98.6	7.2	23.2	21.6	2.7	14.2	1.4	1.4	10.2
$1 + 3X_{i1}^2 + 3X_{i2}^2$	79.2	4.9	18.7	80.1	5.0	18.8	1.4	1.4	9.7
$\exp(3X_{i1} + 3X_{i2})$	87.6	86.6	100.0	88.7	86.8	100.0	6.9	0.4	66.5
$1 + 3X_{i1}^2 + 3X_{i2}^2 + 3X_{i3}^2$	48.9	3.4	14.3	50.2	3.9	14.8	49.5	4.0	15.1

"White" represents the White's test; "BPG" represents the Breusch–Pagan–Godfrey test.

BPG test, the empirical test size of the White's test is generally smaller than the nominal level, but that of the GQ test is seriously inflated.

Tables 3 and 4 summarize the empirical frequencies of identifying each of the covariates $\{X_{i1}, X_{i2}, X_{i3}\}$ as one associated with the error variances under each of the heteroscedastic variance structures. Firstly, our method has high sensitivity, which increases as the sample size increases. For example, under the error variance $\sigma_i^2 = 1 + X_{i1}^2$, for sample size 200, with 91.5% success rate, X_{i1} is identified as the covariate associated with the variance amongst the 10,000 replicates; for the sample size 400, the sensitivity increases to 99.9%. Secondly, the results show high specificity. For example, under the error variance $\sigma_i^2 = 1 + X_{i1}^2$, for sample size 200, with 98.6% success rate, X_2 is correctly identified as covariate not associated with the variance. However, we observe that the specificity for X_{i3} under variance structures of exponential form are relatively lower, that is, the false positive rates appear inflated. Further investigation is required.

TABLE 4: Empirical relative frequencies (multiplied by 100) of identifying each of the covariates as a covariate associated with the error variances when the sample size is 400.

	X_1			X_2			X_3		
	IR	White	BPG	IR	White	BPG	IR	White	BPG
$1 + X_{i1}^2$	99.9	6.6	19.5	1.4	1.4	8.8	1.1	1.4	8.3
$1 + 3X_{i1}^2$	100.0	8.6	28.5	1.4	1.3	13.2	1.2	1.3	12.7
$\exp(3X_{i1})$	100.0	100.0	100.0	1.5	1.3	45.7	1.8	1.2	44.8
$1 + 3X_{i1}^2 + X_{i2}^2$	100.0	7.2	23.8	43.2	2.8	14.9	1.1	1.4	10.6
$1 + 3X_{i1}^2 + 3X_{i2}^2$	98.7	5.1	19.8	98.7	5.3	20.1	1.2	1.4	10.1
$\exp(3X_{i1} + 3X_{i2})$	98.4	98.0	100.0	98.5	98.1	100.0	4.7	0.8	70.3
$1 + 3X_{i1}^2 + 3X_{i2}^2 + 3X_{i3}^2$	83.0	3.8	15.1	83.4	4.0	15.7	82.7	4.2	15.5

"White" represents the White's test; "BPG" represents the Breusch–Pagan–Godfrey test.

Thirdly, we observe that the sensitivity depends on the effect size of the covariate on the error variances. We focus on the scenarios with the sample size 200. The variance functions $\sigma_i^2 = 1 + X_{i1}^2$ and $\sigma_i^2 = 1 + 3X_{i2}^2$ are in the form of $\sigma_i^2 = 1 + \phi X_{i1}^2$, and the parameter ϕ characterizes the effect size. The results in Table 3 show that the true positive rate of X_{i1} is 91.5% with $\phi = 1$, and it increases to 99.9% with $\phi = 3$. Comparing the three variance structures with the same effect sizes of X_{i1}: $\sigma_i^2 = 1 + 3X_{i1}^2$, $\sigma_i^2 = 1 + 3X_{i1}^2 + 3X_{i2}^2$ and $\sigma_i^2 = 1 + 3X_{i1}^2 + 3X_{i2}^2 + 3X_{i3}^2$, their empirical true positive rates of X_{i1} are 99.9%, 79.2%, and 48.9%, respectively. Even though the effect sizes of X_{i1} are the same across these three variance functions, their true positive rates differ. This is because the effect size does not fully characterize the marginal contribution of the corresponding covariate to the error variances in the presence of other covariates. For a better measure, we define covariate-specific standardized variance functions. Take X_{i1} as an example. Let $\widetilde{V}(X_{i1}) = \mathrm{E}_{\mathbf{X}_{(-1)}}\{V(\mathbf{X}_i)|X_{i1}\}/\mathrm{E}_{\mathbf{X}}\{V(\mathbf{X}_i)\}$ where $\mathrm{E}_{\mathbf{X}_{(-1)}}$ denotes the expectation w.r.t. the distribution of $\mathbf{X}_{(-1)}$, a sub-vector of $\mathbf{X}$ deleting its first element; call $\widetilde{V}(X_{i1})$ the X_1-specific standardized variance function. The X_1-specific standardized functions of $1 + 3X_{i1}^2$, $1 + 3X_{i1}^2 + 3X_{i2}^2$, and $1 + 3X_{i1}^2 + 3X_{i2}^2 + 3X_{i3}^2$ are $(1/4) + (3/4)X_{i1}^2$, and $(1/7) + (3/7)X_{i1}^2$, and $(1/10) + (3/10)X_{i1}^2$, where the marginal effect sizes of X_{i1} are 3/4, 3/7 and 3/10, respectively. Corresponding to these decreasing marginal effect sizes, the empirical true positive rates decrease, 99.9%, 79.2% and 48.9%. Thus, the sensitivity increases with increasing marginal effect sizes in the covariate-specific standardized variance functions. These marginal effect sizes are able to better capture the marginal contributions of the corresponding covariates, compared with the effect sizes in the original variance functions.

Fourthly, it appears that the covariate-specific IR statistics are powerful to detect covariates associated with heteroscedastic variance structures in a variety of forms. This is because that our method does not require any explicit form of the error variance functions.

Finally, the empirical identification frequencies of X_{i1}, X_{i2} and X_{i3} via the White's test and the BPG test are also reported in Tables 3 and 4. Compared with these two tests, our method performs better in general. Specifically, both White's test and the BPG test have higher sensitivity in identifying the true dependant covariates for variance functions in an exponential form. However, neither of these two tests performs well for the variance functions in quadratic forms.

4.2. Data Analysis

To illustrate the effectiveness of our method, we analyze a data set collected from a retrospective study of inpatients who underwent initiation of renal replacement therapy for a diagnosis of acute kidney injury (AKI) during the period of November, 2007 to October, 2008 (Heung et al. , 2012). This was the first study ever that collected biomarker data from patients with severe AKI who were treated by dialysis. This data consists of 170 patients with several baseline characteristics including age, gender (96 males and 74 females), baseline weight, and three important clinical measures to reflect profiles of renal function: baseline serum creatinine (mg/dL), albumin at admission (g/dL), and haemoglobin at admission (mg/dL). Note that the baseline weight used in the analysis is defined as the average outpatient weight recorded within 3 months preceding hospitalization.

One aim of this analysis was to characterize one baseline renal function outcome of hospitalized AKI patients with the other two lab measurements related to renal functions, adjusting for confounding covariates (age, gender and baseline weight). Such relationship has not been well studied in the literature for severe inpatient AKI patients. Amongst these associations, it is of great interest to investigate the relationship of creatinine with albumin and haemoglobin, adjusting for age, gender and baseline weight. This is because creatinine is a primary clinical variable used in the estimation of glomerular filtration rate, the most widely used outcome for renal function.

Prior to the analysis, a log transformation was applied to the baseline weight and creatinine. Figure 1 shows the histograms and pairwise scatter plots as well as pairwise correlations of the

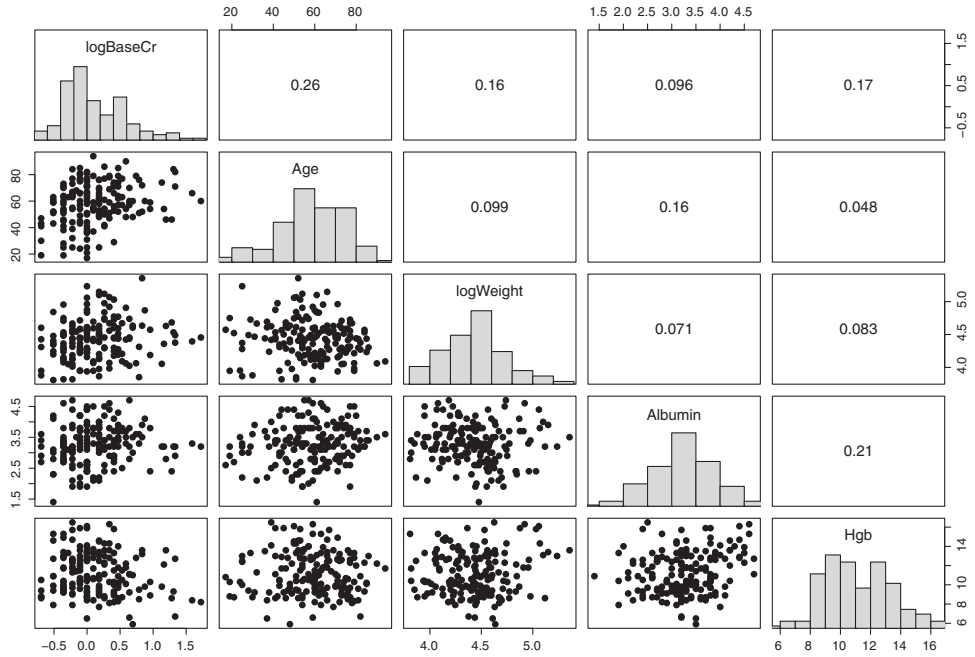

FIGURE 1: Acute kidney injury data: Histograms of log baseline serum creatinine (logBaseCr), age (Age), log baseline weight (logWeight), albumin at admission (Albumin), and haemoglobin at admission (Hgb) as well as their pairwise scatter plots and pairwise correlations. [Color figure can be seen in the online version of this article, available at http://wileyonlinelibrary.com/journal/cjs]

continuous variables used in the analysis. We regress the log baseline serum creatinine on age, gender, log baseline weight, albumin and haemoglobin. The OLS estimates of the regression coefficients as well as the model-based and sandwich estimates of the standard errors (SE) are reported in Table 5. The results show that age, log baseline weight and haemoglobin are significant factors. The model-based and sandwich SE estimates appear different on some variables such as albumin, which might be an indirect indication of heteroscedasticity.

TABLE 5: The OLS and WLS estimates of the regression coefficients and their associated standard errors.

	OLS			WLS		
	Est	$SE^m(SE^s)$	*P*-value	Est	SE	*P*-value
Age	0.0070	0.0022 (0.0019)	<0.01	0.0054	0.0020	<0.01
Gender	0.1130	0.0731 (0.0752)	0.133	0.1398	0.0659	0.036
logWeight	0.2578	0.1207 (0.1099)	0.019	0.2790	0.1072	0.010
Albumin	0.0755	0.0539 (0.0429)	0.079	0.0889	0.0432	0.041
Hgb	−0.0421	0.0155 (0.0152)	<0.01	−0.0466	0.0139	<0.01
AIC		207.1			185.2	
BIC		229.0			207.1	

For the OLS estimation, the model-based estimates SE^m and sandwich estimates SE^s of the standard errors are reported. The *P*-values are obtained using the sandwich standard error estimtes.

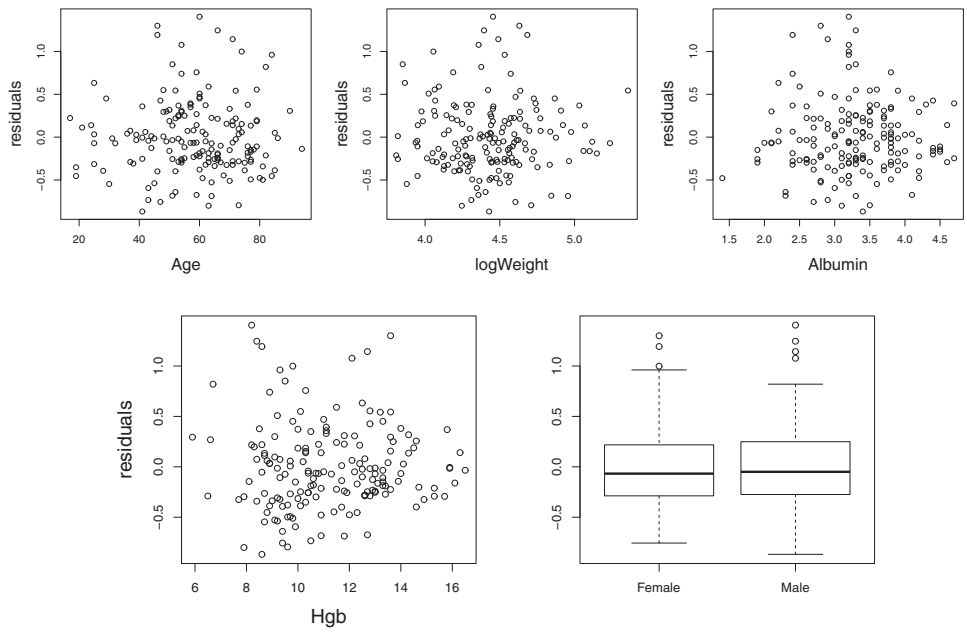

FIGURE 2: Acute kidney injury data: Residual plots versus age, log baseline weight (logWeight), albumin, haemoglobin (Hgb) and gender.

Figure 2 comprises the residual plots against the five covariates. Some patterns of association could be observed between the error variances and age, log baseline weight and albumin, but the patterns do not appear very strong. We applied the proposed two-step procedure to examine these patterns rigorously. The P-value for the pooled IR test was 0.0018. Figure 1 shows weak correlations amongst the covariates, and thus it seems reasonable to use the covariate-specific IR tests without running the covariate-wise screening algorithm described in Section 3. The covariate-specific P-values were found to be 0.086, 0.25, 0.25, 0.0004 and 0.93 for age, gender, log baseline weight, albumin and haemoglobin, respectively. The hybrid P-value was 0.0036. Given the significance level 0.05, the null hypothesis of homoscedasticity was rejected. The variable albumin has been identified as a covariate associated with the error variances, which agreed with the result in Table 5 that the discrepancy between the two SE estimates for the variable albumin was the largest amongst the five covariates.

Next we investigated the influence of incorporating heteroscedasticity in the estimation of the regression coefficients. Since the relationship of the error variances with albumin appeared nonlinear in the residual plot, we obtained the nonparametric estimates of the error variances via kernel smoothing (Wand & Jones , 1995). The plot of the estimated error variances as a function of albumin is shown in Figure 3, and this plot also confirms the nonlinear relationship of the error variances with albumin. Note that the kernel estimates of the error variances was obtained using the *ksmooth* function in R, with the bandwith chosen by using the *np.gcv* funtion in R. In addition, to investigate the influence of the endpoints on the estimated variance, we fit the nonparametric models using the observations of albumin with different proportions of endpoints deleted. The resulting estimated variances appear similar. Thus, removing various proportions of endpoints has little influence on the estimated variance structure.

The estimated error variances $\widehat{\sigma}_i^2$ were used as the weights in the WLS estimation, and the resulting estimates with estimated standard errors (e.s.e.) were reported in Table 5. In contrast to the homoscedastic model, the magnitudes of the coefficients estimates were slightly different

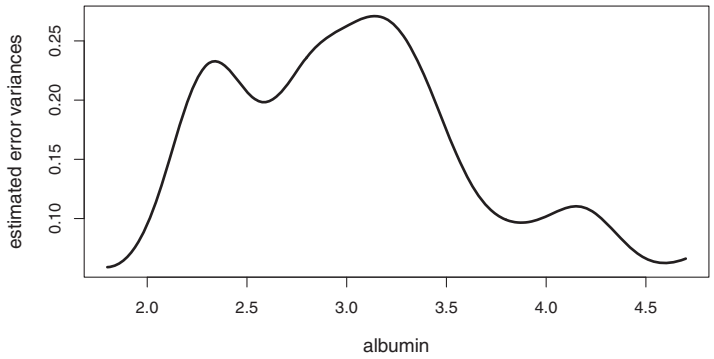

FIGURE 3: Acute kidney injury data: Plot of the nonparametric estimates of the error variances via the kernel smoothing versus albumin.

in the heteroscedastic model, and the e.s.e.'s tended to be smaller. Two variables, gender and albumin, were significant when accounting for the heteroscedasticity in the LM, but they were insignificant in the homoscedastic model. In addition, the AIC and BIC of the heteroscedastic model were smaller than those given by the homoscedastic model. These results all indicated that correctly identifying the dependence of error variances on covariates is critically important to yield appropriate estimation and inference in the regression analysis.

5. CONCLUDING REMARKS

We have proposed a powerful screening procedure to detect the presence and sources of variance heterogeneity in LMs. The proposed procedure is constructed based on two types of IR statistics, both relying on comparisons between sandwich and model-based covariance matrices of regression coefficients estimators. Compared with existing tests for heteroscedasticity, the novelty of our methods lies in the fact that they do not need to assume any specific models for the error variances. Thus, our method is flexible to deal with general heteroscedastic variance structures. This is illustrated through the simulation studies, in which we also show that the proposed testing procedure can control the type I error and produce high sensitivity.

We note that the finite sample performance of our proposed method might not be stable since the asymptotic distribution might not be a good approximation to the finite sample distribution of the covariate-specific IR statistics. Further investigation is desirable. In addition, the covariate-specific IR statistics are not invariant to different linear transformations on the design matrix. Wrongly chosen transformations can give misleading results, but, on the other hand, the method permits a transformation that is closely pertinent to a relevant space for better understanding of heteroscedasticity profiles. The covariate-wise screening algorithm proposed in Section 3 provides one approach to basing the profiling on the covariate space. In another context, one may perform a singular value decomposition on the design matrix to extract principal components that align with differences in the variance. The problem of optimal rotation on covariates deserves future research.

Other future work includes extending the proposed method to profiling heterogeneity in variances across different subjects in longitudinal data using a linear mixed-effects model (LMM) $\mathbf{Y}_i = \beta_0 + \mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\mathbf{b}_i + \mathbf{e}_i$ where $\mathbf{b}_i$'s are *random effects*, typically assumed to be random with zero mean and covariance matrix Σ_b. This LMM leads to the conditional variance of $\mathbf{Y}_i$: $\mathrm{Cov}(\mathbf{Y}_i \mid \mathbb{Z}_i) = \sigma^2 + \mathbb{Z}_i\Sigma_b\mathbb{Z}_i^{\mathsf{T}}$, which depends on $\mathbb{Z}_i$, the covariates associated with random effects. The model could be generalized to the form $\mathbf{Y}_i = \beta_0 + \mathbb{X}_i\boldsymbol{\beta} + h(\mathbb{Z}_i; \mathbf{b}_i) + \mathbf{e}_i$, where $h(\cdot)$ is an arbitrary smooth function. Misspecification of the random effects structures may affect the

estimation efficiency of the fixed effects. A key component of correct specification of the random effects structure is correct identification of elements in the covariates which are associated with the random effects. The proposed procedure can be extended to detect covariates associated with the random effects in mixed effects models.

APPENDIX

Proof of Lemma 1. By the QR decomposition (Golub & Van Loan, 1996), the matrix $\mathcal{X}$ can be decomposed into a product $\mathcal{X} = \mathbf{QR}$ of an orthogonal $n \times p$ matrix $\mathbf{Q}$ and a $p \times p$ upper triangular matrix $\mathbf{R}$. Let $\mathbf{U} = \mathbf{R}^{-1}$ which is also an upper triangular matrix. The sandwich and model-based covariance matrix estimators of $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$, Equations (3) and (4), can be expressed as $\mathfrak{S}_{\widehat{\boldsymbol{\beta}}} = n\mathbf{U}\left(\mathbf{Q}^\mathsf{T}\mathcal{R}\mathbf{Q}\right)\mathbf{U}^\mathsf{T}$ and $\mathfrak{M}_{\widehat{\boldsymbol{\beta}}} = n\widehat{\sigma}^2\mathbf{U}\mathbf{U}^\mathsf{T}$.

Let us start with the sandwich and model-based estimators of the variance of $\sqrt{n}(\widehat{\beta}_p - \beta_p^*)$. Re-express the matrix $\mathbf{Q} = \left(\mathbf{Q}_{(-p)}, \mathbf{Q}_p\right)$, where $\mathbf{Q}_{(-p)}$ is an $n \times (p-1)$ matrix consisting of the first $p-1$ columns of $\mathbf{Q}$ (equivalently deleting the pth column), and $\mathbf{Q}_p = (Q_{1,p}, \ldots, Q_{n,p})^\mathsf{T}$ is the last column. Note that $\mathbf{Q}_{(-p)}$ is also an orthogonal matrix. Then, we re-partition the matrix $\mathbf{U} = \begin{pmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ \mathbf{0}_{p-1}^\mathsf{T} & u_{22} \end{pmatrix}$, where $\mathbf{U}_{11}$ is a $(p-1) \times (p-1)$ upper triangular matrix, $\mathbf{U}_{12}$ is a $(p-1) \times 1$ vector, and u_{22} is a scalar. Consequently, the sandwich and model-based variance estimators of $\sqrt{n}(\widehat{\beta}_p - \beta_p^*)$ can be written as $\mathcal{V}_p^s = nu_{22}^2 \sum_{i=1}^n \mathbf{Q}_p^T\mathcal{R}\mathbf{Q}_p$ and $\mathcal{V}_p^m = n\widehat{\sigma}^2 u_{22}^2$. The X_p-specific IR statistic is given by

$$\mathrm{IR}_p = \mathcal{V}_p^s/\mathcal{V}_p^m = \sum_{i=1}^n Q_{i,p}^2 r_i^2/\widehat{\sigma}^2.$$

Next, we prove that $Q_{i,p}^2 = h_{ii} - h_{ii}^{(-p)}$, where h_{ii} and $h_{ii}^{(-p)}$ are the diagonal elements of the hat matrices $\mathcal{H}$ and $\mathcal{H}_{(-p)}$. To this end, we note that $\mathcal{H} = \mathbf{QQ}^\mathsf{T} = \mathbf{Q}_{(-p)}\mathbf{Q}_{(-p)}^\mathsf{T} + \mathbf{Q}_p\mathbf{Q}_p^\mathsf{T}$, where the ith diagonal element of $\mathbf{Q}_p\mathbf{Q}_p^\mathsf{T}$ is $Q_{i,p}^2$. In addition, the QR decomposition of $\mathcal{X}$ leads to $\mathcal{X}_{(-p)} = \mathbf{Q}_{(-p)}\mathbf{U}_{11}^{-1}$, and consequently $\mathcal{H}_{(-p)} = \mathbf{Q}_{(-p)}\mathbf{Q}_{(-p)}^\mathsf{T}$. Thus, $h_{ii} - h_{ii}^{(-p)} = Q_{ip}^2$.

For $j = 1, \ldots, p-1$, switching β_j with β_p leads to the switch between the jth and pth column in $\mathcal{X}$ as well as the switch between the jth and pth diagonal elements of $\mathfrak{S}_{\widehat{\boldsymbol{\beta}}}$ and $\mathfrak{M}_{\widehat{\boldsymbol{\beta}}}$. In addition, the hat matrix $\mathcal{H}$ and the residuals r_i's remain the same during the permutation. Thus, the results for IR_p can be generalized to IR_j, for $j = 1, \ldots, p-1$. ■

Proof of Lemma 2. Let $\mathbf{X}_{i,(-j)}$ denote the subvector of $\mathbf{X}_i$ with the jth element deleted. Let $A_j = \mathrm{E}_\mathbf{X}(X_{ij}^2)$, for $j = 1, \ldots, p$. Due to $\mathrm{E}_\mathbf{X}(X_{ij}) = 0$ and Assumption (A2) $X_{i1}, \ldots, X_{ip}$ are uncorrelated, we can show that $\mathrm{E}_\mathbf{X}(\mathbf{X}_i\mathbf{X}_i^\mathsf{T}) = \mathrm{diag}\{A_j, j = 1, \ldots, p\}$ and $\mathrm{E}_\mathbf{X}(\mathbf{X}_{i,(-j)}\mathbf{X}_{i,(-j)}^\mathsf{T}) = \mathrm{diag}\{A'_j, j' = 1, \ldots, p, j' \neq j\}$. Let $\mathbf{A} = \mathrm{E}_\mathbf{X}(\mathbf{X}_i\mathbf{X}_i^\mathsf{T})$ and $\mathbf{A}_{(-j)} = \mathrm{E}_\mathbf{X}(\mathbf{X}_{i,(-j)}\mathbf{X}_{i,(-j)}^\mathsf{T})$.

Let $\mathbf{A}_n = n^{-1}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\mathsf{T}$ and $\mathbf{A}_{(-j),n} = n^{-1}\sum_{i=1}^n \mathbf{X}_{i,(-j)}\mathbf{X}_{i,(-j)}^\mathsf{T}$. The quantities h_{ii} and $h_{ii}^{(-j)}$ can be written as, respectively, $h_{ii} = n^{-1}\mathbf{X}_i^\mathsf{T}\mathbf{A}_n^{-1}\mathbf{X}_i$ and $h_{ii}^{(-j)} = n^{-1}\mathbf{X}_{i,(-j)}^\mathsf{T}\mathbf{A}_{(-j),n}^{-1}\mathbf{X}_{i,(-j)}$. Let $\bar{h}_{ii} = n^{-1}\mathbf{X}_i^\mathsf{T}\mathbf{A}^{-1}\mathbf{X}_i$ and $\bar{h}_{ii}^{(-j)} = n^{-1}\mathbf{X}_{i,(-j)}^\mathsf{T}\mathbf{A}_{(-j)}^{-1}\mathbf{X}_{i,(-j)}$. By the Central Limit Theorem (CLT), we have $|n^{-1}\sum_{i=1}^n X_{ij}X_{ik} - \mathrm{E}_\mathbf{X}(X_{ij}X_{ik})| = O_p(n^{-1/2})$ for $j, k = 1, \ldots, p$. Consequently, we can show that $\sup_i |h_{ii} - \bar{h}_{ii}| = O_p(n^{-3/2}) = o_p(n^{-1})$ and $\sup_i |h_{ii}^{(-j)} - \bar{h}_{ii}^{(-j)}| = O_p(n^{-3/2}) = o_p(n^{-1})$.

Note that $\bar{h}_{ii} = n^{-1}\sum_{j=1}^p X_{ij}^2/A_j$ and $\bar{h}_{ii}^{(-j)} = n^{-1}\sum_{j'=1,\ldots,p,j'\neq j} X_{ij'}^2/A_{j'}$. Thus, $\bar{w}_i^{(j)} = \bar{h}_{ii} - \bar{h}_{ii}^{(-j)} = n^{-1}X_{ij}^2/A_j$. We can show that $\sup_i |w_i^{(j)} - \bar{w}_i^{(j)}| \leq \sup_i |h_{ii} - \bar{h}_{ii}| + \sup_i |h_{ii}^{(-j)} - \bar{h}_{ii}^{(-j)}| = O_p(n^{-3/2}) = o_p(n^{-1})$. ■

Proof of Theorem 1. First, we show the following results: (**R1**) $\sup_i |w_i^{(j)}| = O_p(n^{-1})$ and $\sup_i |\bar{w}_i^{(j)}| = O_p(n^{-1})$; (**R2**) $r_i^2 = e_i^2 + \mathcal{C}_i e_i + O_p(n^{-1})$, where $\mathcal{C}_i$ is $O_p(n^{-1/2})$. To prove the result (**R1**), we first can show $\sup_i |\bar{w}_i^{(j)}| = O_p(n^{-1})$ using the result $\bar{w}_i^{(j)} = n^{-1} X_{ij}^2 / A_j$ in the proof of Lemma 2 and the assumption that the support $\Omega_{\mathbf{X}}$ of $\mathbf{X}_i$ is compact. Using the result of Lemma 2, $\sup_i |w_i^{(j)} - \bar{w}_i^{(j)}| = o_p(n^{-1})$, we can prove $\sup_i |w_i^{(j)}| \le \sup_i |w_i^{(j)} - \bar{w}_i^{(j)}| + \sup_i |\bar{w}_i^{(j)}| = O_p(n^{-1})$. To prove the result (**R2**), since $r_i = e_i - \mathbf{X}_i^{\mathsf{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ and $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = O_p(n^{-1/2})$, we can write $r_i = e_i + \mathcal{O}_i$, where $\mathcal{O}_i = O_p(n^{-1/2})$. Thus, $r_i^2 = e_i^2 + \mathcal{C}_i e_i + O_p(n^{-1})$, where $\mathcal{C}_i = 2\mathcal{O}_i = O_p(n^{-1/2})$.

Let $Q_1 = \sum_{i=1}^n w_i^{(j)} r_i^2$ and $Q_2 = n^{-1} \sum_{i=1}^n r_i^2$ denote the numerator and denominator of the X_j-specific IR statistic. Let $\bar{Q}_1 = \sum_{i=1}^n \bar{w}_i^{(j)} e_i^2 = n^{-1} \sum_{i=1}^n X_{ij}^2 e_i^2 / A_j$ and $\bar{Q}_2 = n^{-1} \sum_{i=1}^n e_i^2$. Next we show the result (**R3**) $|\sqrt{n}(Q_1 - \bar{Q}_1)| = O_p(1)$ and $|\sqrt{n}(Q_2 - \bar{Q}_2)| = O_p(1)$. To this end, first of all, using the result (**R2**) above, we can show that $\sqrt{n}(Q_2 - \bar{Q}_2) = n^{-1/2} \sum_{i=1}^n \mathcal{C}_i e_i + O_p(n^{-1/2})$. Note that $n^{-1/2} \sum_{i=1}^n \mathcal{C}_i e_i$ is bounded by a constant multiplied by $n^{-1} \sum_{i=1}^n e_i = O_p(n^{-1/2})$. Thus, $|\sqrt{n}(Q_2 - \bar{Q}_2)| = O_p(1)$. Using the result (**R1**) and similar arguments, we can show that $|\sqrt{n}(Q_1 - \bar{Q}_1)| = O_p(1)$.

By the Taylor expansion of $\mathrm{IR}_j = Q_1/Q_2$, we have $\sqrt{n}(\mathrm{IR}_j - 1) = \sqrt{n}(\nu_j^2/\nu^2 - 1) + \mathbb{T}_1 + \mathbb{T}_2$ where

$$\mathbb{T}_1 = \sqrt{n} \left\{ \frac{1}{\nu^2}(Q_1 - \nu_j^2) - \frac{\nu_j^2}{(\nu^2)^2}(Q_2 - \nu^2) \right\}, \tag{A.1}$$

$$\mathbb{T}_2 = o_p(\sqrt{n}|Q_1 - \nu_j^2| + \sqrt{n}|Q_2 - \nu^2|). \tag{A.2}$$

We can show that the term in Equation (A.2) is $\mathbb{T}_2 = o_p(\sqrt{n}|Q_1 - \bar{Q}_1| + \sqrt{n}|Q_2 - \bar{Q}_2| + \sqrt{n}|\bar{Q}_1 - \nu_j^2| + \sqrt{n}|\bar{Q}_2 - \nu^2|) = o_p(1)$ by the result (**R3**) and the facts that $\sqrt{n}(\bar{Q}_1 - \nu_j^2) = O_p(1)$ and $\sqrt{n}(\bar{Q}_2 - \nu^2) = O_p(1)$ by CLT. Using the similar arguments in proving the result (**R3**) and $\sum_{i=1}^n w_i^{(j)} = 1$, we can show that the term in Equation (A.1) is

$$\begin{aligned}
\mathbb{T}_1 &= \sqrt{n} \left(\sum_{i=1}^n w_i^{(j)} \frac{e_i^2}{\nu^2} - \frac{\nu_j^2}{\nu^2} \frac{1}{n} \sum_{i=1}^n \frac{e_i^2}{\nu^2} \right) + o_p(1) \\
&= \sqrt{n} \sum_{i=1}^n \left\{ \left(w_i^{(j)} - \frac{\nu_j^2}{\nu^2} \frac{1}{n} \right) \left(\frac{e_i^2}{\nu^2} - 1 \right) - \frac{1}{n} \left(\frac{\nu_j^2}{\nu^2} - 1 \right) \right\} + o_p(1).
\end{aligned}$$

Further using the result (**R1**), we can show

$$\begin{aligned}
\mathbb{T}_1 &= \sqrt{n} \sum_{i=1}^n \left\{ \left(\bar{w}_i^{(j)} - \frac{\nu_j^2}{\nu^2} \frac{1}{n} \right) \left(\frac{e_i^2}{\nu^2} - 1 \right) - \frac{1}{n} \left(\frac{\nu_j^2}{\nu^2} - 1 \right) \right\} + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \left(\frac{X_{ij}^2}{A_j} - \frac{\nu_j^2}{\nu^2} \right) \left(\frac{e_i^2}{\nu^2} - 1 \right) - \left(\frac{\nu_j^2}{\nu^2} - 1 \right) \right\} + o_p(1).
\end{aligned}$$

Let $\mathcal{R}_{ij} = ((X_{ij}^2)/(A_j) - (\nu_j^2)/(\nu^2))((e_i^2)/(\nu^2) - 1) - ((\nu_j^2)/(\nu^2) - 1)$. Note that $\{\mathcal{R}_{ij}, i = 1, \ldots, n\}$ are independent random variables with $\mathrm{E}_{(Y,\mathbf{X})}(\mathcal{R}_{ij}) = 0$ and $\Sigma_j = \mathrm{Var}_{(Y,\mathbf{X})}(\mathcal{R}_{ij}) = \mathrm{E}_{(Y,\mathbf{X})}(\mathcal{R}_{ij}^2)$. Under the null hypothesis, corresponding to $V(\mathbf{X}_i) \equiv 1$ in the general variance structure Equation (5), $\mathrm{E}_{Y|\mathbf{X}}(e_i^2) = \kappa^2$ and $\mathrm{E}_{Y|\mathbf{X}}(e_i^2 X_{ij}^2)/A_j = \kappa^2 X_{ij}^2 / A_j$. Thus, we have $\nu^2 = \nu_j^2 = \kappa^2$, and the test statistic $\sqrt{n}(\mathrm{IR}_j - 1) = n^{-1/2} \sum_{i=1}^n \mathcal{R}_{ij} + o_p(1)$. By the central limit theorem,

$\sqrt{n}(\mathrm{IR}_j - 1)$ converges in distribution to a Gaussian random variable with mean 0 and variance Σ_j. ■

Derivation of Equation (12). The statistics in Equation (12) are essentially derived from the following procedure. First we consider an estimating function of the form

$$\sum_{i=1}^{n} \frac{1}{\sigma^2} \mathbf{C}_i (e_i - \mathbf{C}_i^{\mathsf{T}} \boldsymbol{\gamma}) = 0, \tag{A.3}$$

where $e_i = Y_i - \mathbf{X}_i^{\mathsf{T}} \boldsymbol{\beta}^*$ and $\boldsymbol{\gamma}$ is an m-dimensional vector of coefficients. Let $\widehat{\boldsymbol{\gamma}}$ be the solution to the estimating equation, Equation (A.3), and the resulting residuals are $\widetilde{r}_i = e_i - \mathbf{C}_i^{\mathsf{T}} \widehat{\boldsymbol{\gamma}}$, and the variance parameter estimate is $\widetilde{\widehat{\sigma}}^2 = n^{-1} \sum_{i=1}^{n} \widetilde{r}_i^2$. Applying our method given in Section 2, based on Equation (A.3), we can have

$$\mathrm{IR}_{\mathrm{pool}} = \frac{\sum_{i=1}^{n} \widetilde{r}_i^2 w_i^{\mathrm{pool}}}{\widetilde{\widehat{\sigma}}^2} \quad \text{and} \quad \mathrm{IR}_j = \frac{\sum_{i=1}^{n} \widetilde{r}_i^2 w_i^{(j)}}{\widetilde{\widehat{\sigma}}^2}. \tag{A.4}$$

With the mean structure $\mathrm{E}_{Y|\mathbf{X}}(Y_i) = \mathbf{X}_i^{\mathsf{T}} \boldsymbol{\beta}$ correctly specified, we expect that $\widehat{\boldsymbol{\gamma}}$ converges in probability to 0. This implies that the residuals $\widetilde{r}_i$ and $\widetilde{\widehat{\sigma}}^2$ are stochastically equivalent to the residuals r_i and $\widehat{\sigma}^2$, respectively. Moreover, the IR statistics in Equation (A.4) are stochastically equivalent to the statistics in Equation (12). In the literature, several methods have been suggested to test for the adequacy of the mean structure with no need of correctly specifying the variance structure (Hansen, 1982). ■

White's test, Goldfeld–Quandt test and Breusch–Pagan–Godfrey test. Consider model: $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i$, $i = 1, \ldots, n$. According to the description in Gujarati & Porter (2009), we present the details of the White's general heteroscedasticity test, Goldfeld–Quandt test and Breusch–Pagan–Godfrey test.

White's general heteroscedasticity test.

Step 1. Given the data, we obtain the OLS estimates of $\beta_0, \beta_1, \ldots, \beta_p$ and the residuals $r_1, \ldots, r_n$.
Step 2. We then run the following (auxiliary) regression:

$$r_i^2 = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_{ij} + \sum_{j=1}^{p} \alpha_{(j,j)} X_{ij}^2 + \sum_{j_1 \neq j_2} \alpha_{(j_1,j_2)} X_{i,j_1} X_{i,j_2} + e_i.$$

Step 3. Define the test statistic $\mathcal{T} = n \times R^2$ where R^2 is the coefficient of determination for the auxiliary regression model. Under the null hypothesis of homoscedasticity, $\mathcal{T}$ converges in distribution to a chi-square distributed random variable with $2p + p(p-1)/2$ degrees of freedom (d.f.).
Step 4. If the calculated value of the test statistic exceeds the critical chi-square value at the chosen level of significance, the null hypothesis is rejected.

To detect the covariates associated with the error variances, we modify the White's test described above. For each $j = 1, \ldots, p$, we run the auxiliary regression:

$$r_i^2 = \alpha_0 + \alpha_j X_{ij} + \alpha_{(j,j)} X_{ij}^2 + e_i,$$

and calculate the value of the test statistic $\mathcal{T}$. If this value exceeds the critical value of a chi-square random variable with 2 d.f., the conclusion is that the heteroscedastic error variances depend on X_{ij}.

Goldfeld–Quandt test. This test is applicable if one assumes that the heteroscedastic variance σ_i^2 is positively related to one of the explanatory variables in the regression models. For each covariate X_{ij}, $j = 1, \ldots, p$:

Step 1. order or rank the observations according to the values of X_{ij}, beginning with the lowest X_{ij} value.

Step 2. Omit c middle observations, where c is specified a priori, and divide the $(n - c)$ observations into two groups, each of $(n - c)/2$ observations. In the simulation study of this paper, we omit 20% of all the observations.

Step 3. Fit separate OLS regressions with an intercept and all the covariates $X_{i1}, \ldots, X_{ip}$ to the first $(n - c)/2$ observations and the last $(n - c)/2$ observations, and obtain the respective residual sums of squares RSS_1 and RSS_2. These two RSS's each have $\nu = (n - c)/2 - (p + 1)$ d.f.

Step 4. Compute the ratio $\lambda = (RSS_1/\nu)/(RSS_2/\nu)$. If we assume e_i are normally distributed, and if the assumption of homoscedasticity is valid, then λ follows the F distribution with numerator and denominator d.f. each of $(n - c)/2 - (p + 1)$. If the computed λ is greater than the critical value of F distribution at the level of significance, we can conclude that there is evidence that the error variances depend on X_{ij}.

This test is conducted for each covariate. If the test results show that the error variances depend on one or multiple covariates, we reject the hypothesis of homoscedasticity.

Breusch–Pagan–Godfrey test.

Step 1. Estimate $\beta_0, \beta_1, \ldots, \beta_p$ by OLS and obtain the residuals $r_1, \ldots, r_n$.

Step 2. Obtain $\widehat{\sigma}^2 = \sum_{i=1}^n r_i^2/n$.

Step 3. Construct p_i defined as $p_i = r_i^2/\widehat{\sigma}^2$.

Step 4. Regress p_i on $X_{i1}, \ldots, X_{ip}$ as

$$p_i = \alpha_0 + \alpha_1 X_{i1} + \ldots + \alpha_p X_{ip} + e_i.$$

Step 5. Obtain the explained sum of squares from the regression in Step 4 and define $\Theta = ESS/2$. Assuming the e_i are normally distributed, one can show that if homoscedasticity is true, the statistic Θ converges in distribution to a chi-square random variable with p d.f. If the computed value of Θ is greater than the critical value of a χ_p^2 random variable at the level of significance, we reject the null hypothesis of homoscedasticity.

To detect the covariates the error variances might depend on, we modify the Breusch–Pagan–Godfrey test described above. For each covariate X_{ij}, $j = 1, \ldots, p$, we run the regression of p_i:

$$p_i = \alpha_0 + \alpha_1 X_{ij} + e_i,$$

and compute the value of the test statistic Θ. If this value exceeds the critical value of a χ_1^2, the conclusion is that the heteroscedastic error variances depend on X_{ij}. ■

ACKNOWLEDGEMENTS

The authors are grateful to Dr. Richard J. Cook and Dr. Grace Y. Yi from the University of Waterloo for valuable comments on the thesis on which this paper is based. The authors appreciate the constructive suggestions from Dr. Charmaine Dean from Western University of Ontario and Dr. Tianxi Cai of the Harvard School of Public Health. The authors also thank the two anonymous referees, the AE and Editor for their valuable comments that have led to an improvement in the theory and application presented in the paper. This work is supported by grants from the Natural Science and Engineering Research Council of Canada to the first author and the third author, and the US National Science Foundation to the second author.

BIBLIOGRAPHY

Anscombe, F. (1961). Examination of residuals. *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, 1, 1–3.

Bickel, P. (1978). Using residuals robustly I: Tests for heteroscedasticity, nonlinearity. *The Annals of Statistics*, 6, 266–291.

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

Cai, T., Tian, L., & Wei, L. (2005). Semiparametric Box–Cox power transformation models for censored survival observations. *Biometrika*, 92, 619–632.

Carroll, R. & Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *The Annals of Statistics*, 10, 429–441.

Carroll, R. & Ruppert, D. (1988). *Transformation and Weighting in Regression*, Chapman & Hall, New York.

Cook, R. & Weisberg, S. (1982), *Residuals and Influence in Regression*, Vol. 5, Chapman & Hall, New York.

Cook, R. & Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika*, 70, 1–10.

Diblasi, A. & Bowman, A. (1997). Testing for constant variance in a linear model. *Statistics and Probability Letters*, 33, 95–103.

Glejser, H. (1969). A new test for heteroskedasticity. *Journal of the American Statistical Association*, 64, 316–323.

Goldfeld, S. & Quandt, R. (1965). Some tests for homoscedasticity. *Journal of the American Statistical Association*, 60, 539–547.

Golub, G. H. & Van Loan, C. F. (1996). *Matrix Computations*, 3rd ed., Johns Hopkins University Press, USA.

Gujarati, D. & Porter, D. (2009). *Basic Econometrics*, 5th ed., McGraw-Hill Irwin, Boston.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 50, 1029–1054.

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

Heung, M., Wolfgram, D., Kommareddi, M., Hu, Y., Song, & Ojo, A. (2012). Fluid overload at initiation of renal replacement therapy is associated with lack of renal recovery in patients with acute kidney injury. *Nephrology Dialysis Transplantation*, 27, 956–961.

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

Kuk, A. (1999). Nonparametrically weighted least squares estimation in heteroscedastic linear regression. *Biometrical Journal*, 41, 401–410.

Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, 84, 309–326.

Lindsay, B. G. (1982). Conditional score functions: Some optimality results. *Biometrika*, 69, 503–512.

Mak, T. (1992). Estimation of parameters in heteroscedastic linear models. *Journal of the Royal Statistical Society, Series B*, 54, 649–655.

Mammen, E. (1992). Bootstrap, wild bootstrap, and asymptotic normality. *Probability Theory and Related Fields*, 93, 439–455.

Müller, H. & Stadtmuller, U. (1987). Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics*, 15, 610–625.

Müller, H.-G. & Zhao, P.-L. (1995). On a semiparametric variance function model and a test for heteroscedasticity. *The Annals of Statistics*, 23, 946–967.

Park, R. (1996). Estimation with heteroscedastic error terms. *Econometrica*, 34, 888–888.

Park, Y. & Wei, L. (2003). Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika*, 9, 717–723.

Presnell, B. & Boos, D. D. (2004). The IOS test for model misspecification. *Journal of the American Statistical Association*, 99, 216–227.

Ruppert, D., Wand, M., Holst, U., & Hössjer, O. (1997). Local polynomial variance-function estimation. *Technometrics*, 39, 262–273.

Verbyla, A. (1993). Modelling variance heterogeneity: residual maximum likelihood and diagnostics. *Journal of the Royal Statistical Society, Series B*, 55, 493–508.

Wand, M. P. & Jones, M. C. (1995). *Kernel Smoothing*, Chapman and Hall, London.

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

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.

Wu, C. (1986). Jackknife, bootstrap and other resampling methods in statistics. *The Annals of Statistics*, 14, 1261–1350.

Zalesny, M. D. & Farace, R. V. (1986). A field study of social information processing. *Human Communication Research*, 13, 268–290.

Zhou, Q., Song, P. X. K., & Thompson, M. E. (2012). Information ratio test for model misspecification in quasi-likelihood inference. *Journal of the American Statistical Association*, 107, 205–213.

Received 3 April 2013
Accepted 30 January 2015