Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation

ANDREW F. HAYES
Ohio State University, Columbus, Ohio

AND

LI CAI
University of North Carolina, Chapel Hill, North Carolina

Homoskedasticity is an important assumption in ordinary least squares (OLS) regression. Although the estimator of the regression parameters in OLS regression is unbiased when the homoskedasticity assumption is violated, the estimator of the covariance matrix of the parameter estimates can be biased and inconsistent under heteroskedasticity, which can produce significance tests and confidence intervals that can be liberal or conservative. After a brief description of heteroskedasticity and its effects on inference in OLS regression, we discuss a family of heteroskedasticity-consistent standard error estimators for OLS regression and argue investigators should routinely use one of these estimators when conducting hypothesis tests using OLS regression. To facilitate the adoption of this recommendation, we provide easy-to-use SPSS and SAS macros to implement the procedures discussed here.

Ordinary least squares (OLS) regression is arguably the most widely used method for fitting linear statistical models. An OLS regression model takes the familiar form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ip} + \varepsilon_i, \quad (1)$$

where Y_i is case i's value on the outcome variable, β_0 is the regression constant, X_{ij} is case i's score on the jth of p predictor variables in the model, β_j is predictor j's partial regression weight, and ε_i is the error for case i. Using matrix notation, Equation 1 can be represented as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (2)$$

where $\mathbf{y}$ is an $n \times 1$ vector of outcome observations, $\mathbf{X}$ is an $n \times (p + 1)$ matrix of predictor variable values (including a column of ones for the regression constant), and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors, where n is the sample size and p is the number of predictor variables. The p partial regression coefficients in $\boldsymbol{\beta}$ provide information about each predictor variable's unique or partial relationship with the outcome variable. Researchers are often interested in testing the null hypothesis that a specific element in $\boldsymbol{\beta}$ is zero or constructing a confidence interval for that element using a sample-derived estimate combined with an estimate of the sampling variance of the estimate.

The validity of the hypothesis tests and confidence intervals as implemented in most statistical computing packages depends on the extent to which the model's assumptions are met. The assumptions of the OLS regression model include that (1) the Y_is are generated according to the model specified in Equation 1, (2) the X values are fixed (rather than random), (3) the errors are uncorrelated random variables with (4) zero means, and (5) constant variance, the latter assumption known as *homoskedasticity*. Assumptions (1) and (4) require book-length treatment but in effect state simply that the model is correctly specified (see, e.g., Berry, 1993; White, 1980). When the predictors are random variables rather than fixed, assumption (2) can be relaxed without causing major problems by interpreting the partial regression coefficients as "conditional" on values of X, provided that the conditional expectation of ε given the values of X is zero. Unfortunately, this conditional expectation may well be nonzero, such as when predictors contain measurement error. Measurement error in predictors can bias OLS regression estimates, a topic that is beyond the scope of the present article. A nice introduction can be found in Berry (1993).

The focus of our attention in this article is on the homoskedasticity assumption, an assumption that cannot be treated lightly. We briefly overview the effects of heteroskedasticity on hypothesis tests and confidence intervals based on OLS regression coefficient estimates, some of the methods for detecting violations of this assumption, and suggested remedies. We then describe a family of heteroskedasticity-consistent standard error estimators

A. F. Hayes, hayes.338@osu.edu

that, if used, could alleviate the anxiety an investigator might feel about the validity and power of hypothesis tests based on OLS regression coefficients in the presence of heteroskedasticity. We end with a description and example use of some SPSS and SAS macros we developed that allow investigators using these popular programs to employ a heteroskedasticity-consistent estimator of the regression coefficient standard errors in their regression analyses.

What Is Heteroskedasticity and What Are Its Effects on Inference?

The assumption of homoskedasticity states that the variance of the regression errors is constant. Imagine having available the population regression model linking outcome variable Y to a set of p predictor variables (X_j, $j = 1$ to p) using the population regression coefficients. Now, imagine obtaining a subset of all members of the population with the same values on the predictor variables and applying the population regression model to those members of the population. This would allow you to generate a set of regression errors conditioned on those values of X. Define the variance of the errors as σ^2. The homoskedasticity assumption states that the variance of the regression errors is σ^2 regardless of which set of values of the p predictor variables is used to generate those errors. When this assumption is violated, we say that the errors are *heteroskedastic*, a condition known as *heteroskedasticity*. The homoskedasticity assumption implies that the variance of the errors is unrelated to any predictor or any linear combination of the predictor variables. However, homoskedasticity could be violated, even if the magnitude of the error variances is not a function of the predictors in the regression model, a situation referred to as heteroskedasticity of *unknown form* (White, 1980).

Heteroskedasticity can take a number of different forms and result from a variety of different processes. For example, the variability in the errors may increase or decrease linearly as a function of one or more of the predictors, or variability might be larger for moderate values of one or more of the predictors. Such patterns of errors can occur if a model is misspecified as a result of, for example, a failure to include necessary interactions between predictors. Heteroskedasticity can also be the result of modeling bounded or limited outcome variables using OLS regression, or if the reliability of measurement of the outcome is related in some fashion to one or more of the predictor variables (Berry, 1993; Downs & Rocke, 1979; Perry, 1986).

When the homoskedasticity assumption is violated, the usual OLS regression estimator of the partial regression coefficients is unbiased and strongly consistent under heteroskedasticity (White, 1980). However, it is less efficient (i.e., it has a larger sampling variance) than some alternative estimators. Nevertheless, violations of this assumption can invalidate statistical inferences. Such inferences are usually made using a hypothesis test, deriving the p value for the ratio of the sample estimate of the regression coefficient of interest to an estimate of the standard error of the estimate, or by using the estimated standard error to

derive a confidence interval. The standard error estimator for regression coefficient j implemented in most statistical packages—the estimator that assumes homoskedasticity (OLSE)—is

$$\text{OLSE}_j = \sqrt{\frac{MS_{\text{residual}}}{n\left(s_j^2\right)\left(1 - R_j^2\right)}}, \qquad (3)$$

where s_j is the sample standard deviation of X_j, R_j is the multiple correlation estimating X_j from the other ($p - 1$) X variables in the model, and MS_{residual} is the mean squared residual. The extent of the problem produced by heteroskedasticity depends on both the form and the severity of heteroskedasticity. When the errors are heteroskedastic, OLSE is both *biased* and *inconsistent*. The net result is Type I error inflation or reduced statistical power for tests of hypotheses involving the regression coefficients, and inaccuracy in the estimation of the upper and lower bounds on confidence intervals. Increasing the sample size does not eliminate the bias; indeed, it can actually exacerbate the problems produced by heteroskedasticity when using OLSE (Hayes, 1996; Long & Ervin, 2000).

Although it is largely impossible to construct generalizable rules about the extent to which inferences from OLS regression are going to be affected by heteroskedasticity, the existing literature provides some guidance (e.g., Duncan & Layard, 1973; Edgell & Noon, 1984; Hayes, 1996; Kowalski, 1973; Long & Ervin, 2000; Rasmussen, 1989). First, relatively mild heteroskedasticity is not going to produce profound problems and is unlikely to swing the outcome of an analysis drastically one way or the other. Of course, there is no way of defining "mild" without relying on rules of thumb of uncertain value. Furthermore, even mild heteroskedasticity can have notable effects on inference in large samples. Second, if the errors are less variable around the extremes of one of the predictors in the model, OLSE tends to overestimate the standard error and therefore its use produces confidence intervals that are too wide, p values that are too large, and hypothesis tests that are lower in power than they are when the homoskedasticity assumption is met. Third, if the errors are more variable at either relatively high or relatively low values of a predictor variable (or both), OLSE tends to underestimate the standard error, resulting in confidence intervals that are too narrow, p values that are too small, and invalid hypothesis tests.

Heteroskedasticity can also affect tests of significance on the multiple correlation coefficient, R. The null hypothesis that population R equals zero is usually tested by converting R to an F statistic

$$F = \frac{(n - p - 1)R^2}{(p)\left(1 - R^2\right)}, \qquad (4)$$

which is then evaluated in reference to the $F(p, n - p - 1)$ distribution. But use of F as the null reference distribution for generating the p value assumes homoskedasticity of the errors. Violations of this assumption can produce a hypothesis test that is either liberal or conservative, depending on the form and severity of the heteroskedasticity.

Given that heteroskedasticity can affect the validity or power of statistical tests when using OLS regression, it behooves researchers to test the tenability of this assumption. Although the assumption of homoskedasticity applies to the unknowable errors, the assumption is often tested by reliance on the sample residuals, e, which are the observed discrepancies between Y and $\hat{Y}$ using sample-derived estimates of the vector of regression parameters ($\hat{\beta}$) and applying Equation 5 to a sample from the population of interest. That is, $e_i = Y_i - \hat{Y}_i$, where

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_p X_{pi}. \qquad (5)$$

In matrix form, $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$, and thus $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\beta}$, where $\mathbf{e}$ is a vector of sample residuals.

The simplest test of the homoskedasticity assumption is the "eyeball" test, in which the residuals from the regression model are plotted against $\hat{Y}$ (or, alternatively, against one or more of the predictor variables X) in a scatterplot. The homoskedasticity assumption is deemed to be satisfied if the dispersion of the residuals appears to be the same across all values of $\hat{Y}$ or X. A discernible pattern in the variability of the residuals suggests a violation of the homoskedasticity assumption. But as Darlington (1990) and Weisberg (1980) discuss, a visual examination of residuals can be a useful means of detecting only blatant heteroskedasticity.

A more rigorous method is to use an empirical test of the homoskedasticity assumption. One approach is to use either the absolute values of the residuals (raw, standardized, or Studentized) or their squares as the outcome variable in a new regression estimating their size from $\hat{Y}$ or from the original predictors (see, e.g., Darlington, 1990, pp. 360–362). The homoskedasticity assumption is deemed satisfied if no systematic relationships can be found. Alternatively, one of the existing formal tests of homoskedasticity, such as those introduced by Breusch and Pagan (1979), Cook and Weisberg (1983), Goldfeld and Quandt (1965), or White (1980), could be used; but, unfortunately, these tests are designed to detect only certain forms of heteroskedasticity, such as when the variance increases as a function of one or more of the predictors. In the case of White's test, rejection of the null hypothesis is ambiguous because this implies either heteroskedasticity, correlation between the residuals and one or more of the predictors, or both.

In the event that heteroskedasticity exists, various statistical alternatives have been recommended for making sound inferences in its presence. Perhaps the easiest and most widely recommended method is the use of some kind of variance stabilizing transformation of Y. Commonly recommended transformations include $Y^{1/2}$, $\log(Y)$, or Y^{-1}, or the Box–Cox transformation (Box & Cox, 1964), but many others are possible (see, e.g., Carroll & Ruppert, 1988; Cook & Weisberg, 1999). But, typically, it is difficult to know which transformation is optimal without some kind of a priori belief about the functional form of the heteroskedasticity, which may not be well approximated by any transformation.

An alternative approach is weighted least squares (WLS), discussed in a variety of regression books (e.g., Cook & Weisberg, 1983; Draper & Smith, 1981). WLS regression allows the investigator to weight each case differently in the derivation of the sum of squared residuals. When the variance of errors is related to one or more of the predictors by a constant multiplier, weights can be chosen that produce parameter estimates that are more efficient if the researcher has an accurate model of the process generating heteroskedasticity. The gain in efficiency in the estimation of standard errors through WLS regression is a function of how well the weights match the actual (as opposed to the assumed) form of heteroskedasticity. If the weights do not correspond closely to the form of heteroskedasticity, the WLS standard error estimates may still be inaccurate as a result of heteroskedasticity that is not accounted for by the weights chosen.

WLS is a special case of generalized least squares (GLS) (see, e.g., Mardia, Kent, & Bibby, 1979). GLS estimation is widely used in fields such as econometrics, where strong theories usually exist about the functional form of heteroskedasticity. If the investigator has knowledge about the underlying mechanism that generates heteroskedasticity, GLS estimators can be applied and consistency of the standard error estimator regained; but GLS, like WLS, assumes some knowledge of the functional form of the heteroskedasticity. When the form of heteroskedasticity is not known, estimated generalized least squares (EGLS, also called *feasible generalized least squares*) can be used. EGLS refers to a two-stage process where the structure of heteroskedasticity is formally modeled in the first step to generate the weights to be used in a subsequent WLS stage, but this method has the same shortcoming as WLS, in that the model for heteroskedasticity may be incorrect. And even after using WLS, heteroskedasticity may remain, leaving the investigator back on square one, with concerns about the effects of heteroskedasticity on statistical inference.

Heteroskedasticity-Robust Inference Using a Heteroskedasticity-Consistent Standard Error Estimator

An alternative and highly appealing method of reducing the effects of heteroskedasticity on inference is to employ a heteroskedasticity-consistent standard error (HCSE) estimator of OLS parameter estimates (Hinkley, 1977; Long & Ervin, 2000; MacKinnon & White, 1985; White, 1980). With this approach, the regression model is estimated using OLSs, but an alternative method of estimating the standard errors is employed that does not assume homoskedasticity. The appeal of this method lies in the fact that, unlike such methods as WLS, it requires neither knowledge about nor a model of the functional form of the heteroskedasticity. It does not require the use of an arbitrary transformation of Y, and no intensive computer simulation is necessary (such as is required when bootstrapping, an approach others have recommended; e.g., Cribari-Neto, 2004; Godfrey, 2006; Godfrey & Orne, 2004; Wilcox, 2005).

From what we can gather, HCSE estimators are largely unknown outside of statistics and econometrics. Few of the major statistical-method books used by scientists

outside of these fields discuss them, and a citation analysis of some of the seminal works in HCSE estimators reveals that HCSE estimators are rarely cited outside of statistics and economics. So it is safe to assume that most researchers, even those who follow the methodology literature closely, probably know little about HCSE estimators. We believe this is unfortunate, because the literature suggests that they can be quite useful in alleviating the worries that a researcher might have about the effects of heteroskedasticity on inferential tests in OLS regression. Indeed, some econometricians have even gone so far as to claim that with the advent of HCSE estimators, researchers can now "report new statistics that work, regardless of the kind of heteroskedasticity present in the population" (Wooldridge, 2000, p. 249). They are, therefore, worth knowing about and, as we argue below, worth being used routinely, at least as a means of checking whether statistical inferences from a regression analysis are being affected by heteroskedasticity.

Before progressing, we should acknowledge the position that heteroskedasticity can be thought of as a sign that a model is misspecified. However, as MacCallum (2003) points out, all models are wrong to some extent. About the best that we will ever be able to do with a mathematical model of an outcome variable is to approximate the process generating our observations. If we insist that models be correctly specified before we attempt to draw inferences from them, we may as well stop testing hypotheses using models, because we will never achieve perfect specification. When inferential tests using OLS regression assuming homoskedasticity are performed, the assumption is that the errors follow a particular model (i.e., that the elements in $\boldsymbol{\varepsilon}$ have constant variance); or when WLS, GLS, or EGLS is used, the model of heteroskedasticity is assumed to be correct. By advocating the use of an HCSE estimator in OLS regression, we are not arguing that other methods are not useful; they are, to be sure. We are simply acknowledging an alternative way of approaching the problem of inference in OLS regression, one that does not impose a constraint (either assumed or modeled) on the structure of the errors. That is, we ask the reader to consider why one should make an assumption about the structure of the errors when methods exist that allow for valid inference with fewer (and arguably more realistic) assumptions.

We now introduce the details of several of the HCSE estimators that have been proposed. (This section of the paper is necessarily technical and readers not interested in the fine details can skip to the next section.) HCSE estimators are derived from an estimate of the variance–covariance matrix of the regression coefficients, $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$, that does not assume homoskedasticity of the errors. Going back to Equation 2, the error vector $\boldsymbol{\varepsilon}$ is the starting point of HCSE estimators. The assumption of homoskedasticity combined with the assumption of uncorrelated errors is a formal statement that the variance–covariance matrix of the errors, $\boldsymbol{\Phi}$, is equal to $\sigma^2\,\mathbf{I}$. By simple covariance algebra, $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Phi}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$, but when $\boldsymbol{\Phi} = \sigma^2\,\mathbf{I}$, this expression reduces to $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, estimated in a sample as $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = MS_{\text{residual}}\,(\mathbf{X}'\mathbf{X})^{-1}$. This is the usual variance–

covariance matrix estimator that assumes homoskedasticity. The squares of the estimated standard errors for the regression coefficients reside in the main diagonal of $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$, and thus the estimated standard errors of the coefficients are simply the square root of the diagonal elements and the same as the values yielded by Equation 3. But when the errors are heteroskedastic, $\boldsymbol{\Phi} \neq \sigma^2\,\mathbf{I}$, and the OLS estimator of the variance–covariance matrix of the regression coefficients becomes biased, resulting in hypothesis tests that are either liberal or conservative, depending on the form of heteroskedasticity.

Over the last 25 years or so, several heteroskedasticity-consistent estimators of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$ have been proposed. Building upon the earlier work of Eicker (1963, 1967) and Huber (1967), White (1980) argued to place the ith squared error into the ith row of the diagonal of the $\boldsymbol{\Phi}$ matrix, using the OLS residuals as estimators of the errors. So $\hat{\boldsymbol{\Phi}} = \text{diag}[e_i^2]$, which is a diagonal matrix with the squared OLS residuals on the main diagonal and all off diagonal entries zero. The HC0 estimator of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$ is defined as

$$\text{HC0} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{diag}[e_i^2]\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \qquad (6)$$

and the entries on the main diagonal of HC0 are the estimated squared standard errors of the regression coefficients. Dividing the regression coefficients by these standard errors produces a ratio used to derive p values for hypothesis testing.[1] Equation 6 shows how estimators in this family have come to be known as *sandwich estimators*: $\mathbf{X}'\,\text{diag}[e_i^2]\,\mathbf{X}$ is like meat between two slices of bread, i.e., $(\mathbf{X}'\mathbf{X})^{-1}$.

For small sample sizes, the standard errors from HC0 are quite biased, usually downward, and this results in overly liberal inferences in regression models (see, e.g., Bera, Suprayitno, & Premaratne, 2002; Chesher & Jewitt, 1987; Cribari-Neto, Ferrari, & Cordeiro, 2000; Cribari-Neto & Zarkos, 2001; Furno, 1996). But HC0 is a consistent estimator when the errors are heteroskedastic; that is, the bias shrinks with increasing sample size. Three alternative estimators, HC1, HC2, and HC3, are all asymptotically equivalent to HC0 but have far superior small sample properties relative to HC0 (Long & Ervin, 2000; MacKinnon & White, 1985). A newer estimator, HC4, is preferred when there are cases with high leverage.

These alternative HC estimators seek some form of adjustment to the squared OLS residuals before constructing the $\hat{\boldsymbol{\Phi}}$ matrix: HC1, an estimator derived by Hinkley (1977), is a simple degree-of-freedom adjustment to HC0. For HC1, every squared OLS residual is multiplied by a factor of $n/(n - p - 1)$. Therefore, the HC1 estimator of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$ is

$$\text{HC1} = \frac{n}{n - p - 1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{diag}\left[e_i^2\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \quad (7)$$

HC2 is similar to HC1 in logic, but instead of a degree-of-freedom correction, the ith squared OLS residual is weighted by the reciprocal of $(1 - h_{ii})$, where $h_{ii} = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{x}_i'$. The h_{ii}s are also known as *leverage* values and are the diagonal elements in the "hat" matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'$. The use of leverage adjusted residuals is based on an extensive literature on the finite-sample bias

of HC0 (e.g., Chesher & Jewitt, 1987; Cribari-Neto, 2004; Kauermann & Carroll, 2001). The bias is primarily due to the existence of points of high leverage in the **X** matrix. So the HC2 estimator of $\Sigma_{\hat{\beta}}$ is defined as

$$\text{HC2} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{diag}\left[\frac{e_i^2}{1 - h_{ii}}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \qquad (8)$$

Notice that HC2 is equal to HC0 with e_i^2 weighted by $1/(1 - h_{ii})$, and HC2 and HC1 are equivalent when the **X** matrix reflects a balanced experimental design.

As a historical note, the variants of the HC methods are intimately related to a technique known as *the jackknife*. As MacKinnon and White (1985) noted, Hinkley (1977) derived HC1 using a weighted jackknife argument and they used a similar apparatus to obtain what they called HC3. However, the estimator that has come to be widely known as HC3 is actually an approximation proposed by Davidson and MacKinnon (1993) but simpler to compute than the earlier one presented in MacKinnon and White (1985). The HC3 estimator of $\Sigma_{\hat{\beta}}$ is defined as

$$\text{HC3} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{diag}\left[\frac{e_i^2}{\left(1 - h_{ii}\right)^2}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \qquad (9)$$

Notice that HC3 weights each squared OLS residual by a factor of $1/(1 - h_{ii})^2$ rather than $1/(1 - h_{ii})$. Using simulations, Long and Ervin (2000) evaluated the empirical power functions of the t tests of the regression coefficients, using both the ordinary OLS estimator and the four HC methods discussed thus far; they recommended that HC3 always be used because it can keep the test size at the nominal level regardless of the presence or absence of heteroskedasticity (and there is only a slight loss of power associated with HC3 when the errors are indeed homoskedastic). Cribari-Neto, Ferrari, and Oliveira's (2005) simulation results also suggest the superiority of HC3 over its predecessors.

The performance of HC3 does depend to some extent on the presence or absence of points of high leverage in **X** (e.g., Chesher & Jewitt, 1987; Kauermann & Carroll, 2001; Wilcox, 2001). Furthermore, research has shown (e.g., Long & Ervin, 2000; Sudmant & Kennedy, 1990) that HC3 can have a liberal bias in very small samples. In addition, Wilcox (2001) pointed out that when the predictor variables are from heavy-tailed distributions and the errors are from light-tailed distributions, HC3 may fail for certain forms of heteroskedasticity.

The most recent proposal, HC4, was derived by Cribari-Neto (2004) with the explicit aim of taking large leverage values into consideration before constructing $\hat{\boldsymbol{\Phi}}$. He proposed to let

$$\text{HC4} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{diag}\left[\frac{e_i^2}{\left(1 - h_{ii}\right)^{\delta_i}}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \qquad (10)$$

where

$$\delta_i = \min\left\{4, \frac{nh_{ii}}{p + 1}\right\}.$$

The exponent δ_i controls the level of "discounting" for the ith observation, and with the truncation point at 4, the maximum amount of weighting is twice as much as HC3. Cribari-Neto's (2004) simulations show that HC4 can outperform HC3 in terms of test size control when there are high leverage points and nonnormal errors.

It is important to acknowledge that the HCSE estimators described above are not necessarily a panacea for the inferential problems that heteroskedasticity can produce. These estimators have their own limitations (see, e.g., Kauermann & Carroll, 2001; Wilcox, 2001). In particular, the sampling variability of the HCSE estimators described below is often larger than model-based estimators such as those arising from the use of WLS, GLS, or EGLS, if the functional form of heteroskedasticity is correctly specified. Indeed, the use of an HCSE estimator, together with OLS parameter estimates, can be viewed as a special case of the method of generalized estimating equations (GEE, see, e.g., Kauermann & Carroll, 2001), where the working model for the error covariance structure is the same as what is assumed in OLS, but GEE also allows one to specify more elaborate models for the error covariance structure. The advantage of the GEE approach over the HCSE approach is that when the working model of heteroskedasticity is more accurate than the simple error covariance structure model assumed by OLS, one can obtain consistent standard errors but also regain some efficiency in the parameter estimates. The only constraint in GEE is that the size of the error variance has to be a function of one or more of the X variables in order to set up the working model. However, when White (1980) popularized his HC estimator, what he had in mind was heteroskedasticity of an unknown form, including the case in which the size of the error variance is not a function of the predictors. Furthermore, it is apparent that the advantages of explicitly modeling heteroskedasticity using GEE (or EGLS, or WLS) presume some confidence that the model for heteroskedasticity is correct. In almost all areas of statistical inference, including the theory of linear models, the efficiency of estimators and the validity of one's inferences are dependent in part upon the number of assumptions that one is willing to make. HCSE estimators are derived under a minimal set of assumptions about the errors. By using HCSE estimators, the data analyst is to some extent trading efficiency for consistency. As Kauermann and Carroll (2001) succinctly pointed out, this is a price that one has to pay.

Using an HCSE Estimator: An Example

To illustrate the application of an HC estimator, we analyze a data set using OLS regression, first estimating the standard errors of the regression coefficients using OLSE, and then using HC3 and HC4. The data come from a study relating seventh-grade students' grade point average (GPA) to IQ, gender (1 = female, 2 = male), and self-concept, as measured on the Piers–Harris children's self-concept scale. The data are described in the data appendix of Moore and McCabe's (2003) popular introductory statistics text, and can be downloaded as file

name "concept.dat" from www.stat.purdue.edu/people/mccabe/. We focus here on only three of the subscales of the self-concept scale: behavior, popularity, and anxiety. Table 1 displays the OLS regression estimates as well as the estimated standard errors and p values for each regression coefficient (with the p value derived from the ratio of the regression coefficient to its standard error and using the t distribution with $n - p - 1$ degrees of freedom). As can be seen, GPA is significantly related to both IQ and students' self-concepts, with students' perceptions of their popularity and aspects of their behavior significantly predicting GPA, even after controlling for IQ and gender. But for popularity self-concept, the p value just barely leads to a rejection of the null hypothesis. It is in such a situation that a researcher needs to be especially concerned about the effects of assumption violations, because assumption violations can move standard errors and p values in unpredictable ways.[2]

Importantly, there is compelling evidence of violation of the homoskedasticity assumption. A residual scatterplot reveals less variability in the standardized residuals for higher predicted GPA values. The Breusch–Pagan test of homoskedasticity confirms this eyeball test [$\chi^2(5) = 16.51$, $p < .001$], as does an analysis of the correlation between the absolute value of the standardized residuals and the estimated GPA from the model ($r = -0.39$, $p < .001$). Thus, there is reason to be suspicious of the accuracy of the standard errors and, therefore, of the p values. Square root and natural log transformations of GPA were tried, but the heteroskedasticity remained. Furthermore, a couple of the cases have large leverage values—in excess of 3 times the average—and thus are noteworthy in terms of their potential influence on inference (Furno, 1996).

This inferential ambiguity can be navigated by doing hypothesis tests on the basis of a standard error estimator that does not assume homoskedasticity. Table 1 shows the results using the HC3 and HC4 estimators. Note that the standard errors are quite similar for some predictors but noticeably different for others. We can be pretty confident that there is a relationship between IQ and GPA because the regression estimate is statistically different from zero, regardless of how the standard error is estimated; but conclusions about self-concept differ. When heteroskedastic-

ity is managed using the HC3 estimator, the partial relationship between popularity self-concept and GPA is not statistically significant. HC4 confirms the results from the use of HC3.

This example illustrates that heteroskedasticity can affect hypothesis tests and confidence intervals when the homoskedasticity assumption is violated, a fact well established in the literature. At the same time, it illustrates that heteroskedasticity need not incite panic. Not every test in the analysis reported here was influenced by heteroskedasticity. Of course, it is nearly impossible to know how much of an effect heteroskedasticity will have on each of the inferential tests in a standard OLS regression. We believe that HCSE estimators, HC3 and HC4 in particular, can function as a sedative of sorts for the data analyst anxious about the possible effects of heteroskedasticity on the validity of inferences and power of hypothesis tests in OLS regression. Thus, we echo Long and Ervin's (2000) recommendation that investigators should employ a heteroskedasticity-consistent estimator such as HC3—if not as a matter of routine, at least as a means of double-checking one's inferences to see if they might be influenced by heteroskedasticity. In the event that a few cases exhibit high leverage, HC4 can be used instead.

But is this recommendation realistic, given that most statistical programs used by researchers do not allow the investigator to implement this recommendation? In a survey of statistical packages at the time, Long and Ervin (2000) noted that only STATA and TSP offer the newer HC methods (HC2 and HC3) as an option for estimating standard errors, and that a package exists for the freely available R software that implements all 4 of these HC methods. But the two most popular statistical packages in wide use throughout the sciences, SPSS and SAS, offer virtually no opportunity for the data analyst interested in using an HC estimator. SAS offers only the HC0 estimator, and SPSS currently has no built-in procedures for estimating standard errors using any of the HC methods. To remedy this situation, two macros are provided in the Appendix that bring HCSE estimators to users of SPSS and SAS. An example output from the SPSS macro applied to the self-concept data set using HC3 can be found in Figure 1. As can be seen, the macro produces the OLS regression weights along with the HC standard error estimates

Table 1
OLS Regression Analysis Estimating GPA Using Standard Error Estimates Assuming Homoskedasticity (OLS) and Not Assuming Homoskedasticity (HC3 and HC4)

| | | Standard Error Estimator | | | | | |
| | | OLSE | | HC3 | | HC4 | |
	$\hat{\beta}$	SE	p	SE	p	SE	p
Constant	−4.119	1.528		1.565		1.629	
IQ	0.086	0.014	<.001	0.015	<.001	0.016	<.001
Gender	−0.546	0.393	.169	0.340	.112	0.335	.107
SC_Behavior	0.169	0.068	.015	0.077	.031	0.080	.038
SC_Popularity	0.179	0.088	.046	0.099	.076	0.101	.079
SC_Anxiety	−0.061	0.088	.491	0.088	.494	0.089	.497
$R^2 = 0.550$		$F(5,72) = 17.676$ $p < .001$		$F(5,72) = 14.874$ $p < .001$		$F(5,72) = 14.190$ $p < .001$	

SPSS command:

```
hcreg dv=gpa/iv=iq sex sc_beh sc_pop sc_anx/method=3/covmat=1/test=3.
```

```
Run MATRIX procedure:

HC Method
 3

Criterion Variable
 GPA

Model Fit:
       R-sq              F            df1            df2              p
      .5505       14.8735        5.0000        72.0000          .0000

Heteroskedasticity-Consistent Regression Results
              Coeff        SE(HC)             t         P>|t|
Constant    -4.1185        1.5650       -2.6316         .0104
IQ            .0864         .0149        5.8151         .0000
SEX          -.5460         .3397       -1.6071         .1124
SC_BEH        .1692         .0768        2.2035         .0308
SC_POP        .1790         .0993        1.8037         .0755
SC_ANX       -.0607         .0884        -.6868         .4944

Heteroskedasticity-Consistent Covariance Matrix
            Constant            IQ           SEX        SC_BEH        SC_POP        SC_ANX
Constant      2.4493        -.0150        -.0269        -.0571         .0047         .0047
IQ            -.0150         .0002        -.0017        -.0003        -.0004         .0000
SEX           -.0269        -.0017         .1154         .0059         .0040        -.0064
SC_BEH        -.0571        -.0003         .0059         .0059         .0008        -.0010
SC_POP         .0047        -.0004         .0040         .0008         .0099        -.0063
SC_ANX         .0047         .0000        -.0064        -.0010        -.0063         .0078

Setwise Hypothesis Test
          F           df1           df2              p
     2.8576        3.0000       72.0000          .0429

Variables in Set:
   SC_BEH    SC_POP     SC_ANX

------ END MATRIX -----
```

Figure 1. Output from SPSS macro using the HC3 standard error estimator.

and the corresponding t and p values. The standard errors can be used to generate confidence intervals that can be interpreted in the usual manner. The estimated covariance matrix of the regression coefficients is also provided in the output as well as R^2 and F. Although R^2 is not affected by the use of an HC estimator, the F ratio will be different when using one of the HC estimators, as discussed next.

Testing a Hypothesis Requiring More Than a Single Parameter Estimate

Not all hypotheses of interest to researchers can be tested in reference to a single parameter estimate. Researchers are often interested in testing the null hypothesis that each variable in a set of predictors are all unrelated

to the outcome variable, controlling for one or more additional variables. For example, in a regression model of the form $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + e$, the researcher might want to know whether β_3 and β_4 are both zero, which tests whether X_3 and X_4 improve the fit of the model or significantly reduce the error in estimation. In that case, the standard practice is to estimate two nested regression models, one with (model 2: the full model above) and one without the predictors of interest (model 1: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$), and then test the significance of the increase in R^2 resulting from the addition of the predictors in model 2. This increase is converted to an F ratio and the p value computed using the $F(q, df_{res2})$ distribution, where q is the number of variables added to model 1 to generate model 2, and df_{res2} is the residual degrees of freedom from model 2. For this test, F is defined as

$$F = \frac{df_{res2}\left(R^2_{model\ 2} - R^2_{model\ 1}\right)}{q\left(1 - R^2_{model\ 2}\right)}. \qquad (11)$$

This F statistic can be expressed more generally in matrix form as

$$F = \frac{\left(\mathbf{L'\hat{\beta}}\right)'\left(\mathbf{L'\hat{\Sigma}_{\hat{\beta}}L}\right)^{-1}\left(\mathbf{L'\hat{\beta}}\right)}{q}, \qquad (12)$$

where $\mathbf{L}$ is a $(p + 1) \times q$ nonrandom matrix specifying the constraints on the parameters under the null hypothesis, $\hat{\beta}$ is a $(p + 1) \times 1$ vector of regression coefficient estimates, and $\hat{\Sigma}_{\hat{\beta}}$ is the estimated variance–covariance matrix of the regression coefficients (see, e.g., Mardia et al., 1979). For example, if the consecutive rows in $\hat{\beta}$ correspond to sample estimates of β_0 and β_1 through β_4 respectively, the contrast matrix for the example above would be a 5×2 matrix containing all zeros except for ones in entries (4,1) and (5,2). Equations 11 and 12 are equivalent when $\hat{\Sigma}_{\hat{\beta}}$ is set to $MS_{residual}(\mathbf{X'X})^{-1}$ under homoskedasticity. But using an HC estimator of $\hat{\Sigma}_{\hat{\beta}}$ yields a test that does not assume homoskedasticity. The p value can still be derived using the $F(q, df_{res2})$ distribution. Equation 12 can also be used to derive the F ratio for the test that all p population regression coefficients are zero, which corresponds to the null hypothesis that population R^2 equals zero (Darlington, 1990). In that case, $q = p$ and the $\mathbf{L}$ matrix is defined as a $p \times p$ identity matrix vertically concatenated below a $1 \times p$ vector of zeros. Godfrey (2006) also considered this F statistic but used a double bootstrap to derive its p value.

The macros listed in the Appendix have an optional subcommand that tests sets of predictors. This option allows the researcher to test the null hypothesis that the regression coefficients for the last q variables in the predictor variable list are all 0, using one of the heteroskedasticity-consistent estimators of the variance–covariance matrix of the regression coefficients described earlier. Using the SPSS macro, the null hypothesis that all the regression coefficients for self-concept are equal to zero was tested using the HC3 estimator, and the output is displayed at the bottom of Figure 1. As can be seen, this null hypothesis can be rejected [$F(2,72) = 2.858, p = .043$].

Conclusion

The purpose of this article was to provide an overview of heteroskedasticity-consistent standard error estimators for the least squares linear regression model, and to provide a set of macros to implement them in SPSS and SAS. We argue that the use of an HCSE estimator, rather than assuming homoskedasticity when conducting inferential tests in OLS regression, can give the researcher greater comfort in the validity and power of those tests. For this reason, we argue that an HC estimator, preferably HC3 or HC4, should be routinely used in linear regression models, if it is not used as the default method of standard error estimation, researchers would be well advised to at least double-check the results from the use of the OLSE estimator against the results obtained with an HC estimator, to make sure that conclusions are not compromised by heteroskedasticity. Excitingly, research continues to further improve upon the estimators discussed here (e.g., Cai & Hayes, in press), and other promising methods, such as bootstrapping—sometimes combined with HCSE estimators (Cribari-Neto, 2004; Godfrey, 2006; Godfrey & Orne, 2004; Wilcox, 2005)—have been investigated. The bootstrap provides more refined approximations to the distributions of test statistics than do classical large-sample approximations, but it does so at the cost of a dramatic increase of the amount of computational time required to obtain a p value. Efficient algorithms and software implementation of these newer methods are in order.[3] Perhaps in time, HCSE estimators and future improvements upon them will become the default implemented in popular software, eliminating the need for researchers to worry about heteroskedasticity when using OLS regression.

We conclude with a reminder. In this article, we have treated heteroskedasticity as if it were nothing but an impediment to valid inference. However, we remind the reader that heteroskedasticity is more than a nuisance; it may itself have substantively important interpretations (Berry, 1993; Carroll, 2003; Downs & Rocke, 1979; Olejnik, 1988; Weisberg, 1980). Although we do believe that there is value in the use of HC3 or HC4 when testing hypotheses and forming confidence intervals in linear regression models, we do not mean to suggest that researchers should not bother looking for evidence of heteroskedasticity. There is little point is deciding whether or not to use HC methods based on a test of heteroskedasticity (cf. Long & Ervin, 2000), but such a test is still worth doing, because the existence of heteroskedasticity suggests that the data are not being adequately explained by the statistical model of conditional means being estimated. Unless heteroskedasticity is thought to be the result only of measurement error related to the predictor variable(s), it should be a warning to the researcher about the adequacy of the model estimated. Thus, a test for heteroskedasticity can be used not only as a test of model adequacy, but also as a test of model assumptions. The discovery of heteroskedasticity can also function as an impetus for future research to understand the process producing heteroskedasticity, such as latent interactions (Perry, 1986), or to motivate a quest for an alternative model that better fits the data. When

heteroskedasticity is discovered, we should not simply ask "What can I do to make the problem go away?" without also asking "What does heteroskedasticity tell me about the process I am studying?"

AUTHOR NOTE

Correspondence concerning this article should be addressed to A. F. Hayes, School of Communication, Ohio State University, 154 N. Oval Mall, Columbus, OH 43210 (e-mail: hayes.338@osu.edu).

REFERENCES

BERA, A. K., SUPRAYITNO, T., & PREMARATNE, G. (2002). On some heteroskedasticity-robust estimators of variance-covariance matrix of the least-squares estimators. *Journal of Statistical Planning & Inference*, **108**, 121-136.

BERRY, W. D. (1993). *Understanding regression assumptions*. Newbury Park, CA: Sage.

BOX, G. E. P., & COX, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society B*, **26**, 211-243.

BREUSCH, T. S., & PAGAN, A. R. (1979). A simple test for heteroskedasticity and random coefficient variation. *Econometrica*, **47**, 1287-1294.

CAI, L., & HAYES, A. F. (in press). A new test of linear hypotheses in OLS regression under heteroskedasticity of unknown form. *Journal of Educational & Behavioral Statistics*.

CARROLL, R. J. (2003). Variances are not always nuisance parameters. *Biometrics*, **59**, 211-220.

CARROLL, R. J., & RUPPERT, D. (1988). *Transformation and weighting in regression*. New York: Chapman and Hall.

CHESHER, A., & JEWITT, I. (1987). The bias of a heteroskedasticity consistent covariance matrix estimator. *Econometrica*, **55**, 1217-1222.

COOK, R. D., & WEISBERG, S. (1983). Diagnostics for heteroskedasticity in regression. *Biometrika*, **70**, 1-10.

COOK, R. D., & WEISBERG, S. (1999). *Applied regression including computing and graphics*. New York: Wiley.

CRIBARI-NETO, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics & Data Analysis*, **45**, 215-233.

CRIBARI-NETO, F., FERRARI, S. L. P., & CORDEIRO, G. M. (2000). Improved heteroskedasticity-consistent covariance matrix estimators. *Biometrika*, **87**, 907-918.

CRIBARI-NETO, F., FERRARI, S. L. P., & OLIVEIRA, W. A. S. C. (2005). Numerical evaluation of tests based on different heteroskedasticity-consistent covariance matrix estimators. *Journal of Statistical Computation & Simulation*, **75**, 611-628.

CRIBARI-NETO, F., & ZARKOS, S. G. (2001). Heteroskedasticity-consistent covariance matrix estimation: White's estimator and the bootstrap. *Journal of Statistical Computation & Simulation*, **68**, 391-411.

DARLINGTON, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill.

DAVIDSON, R., & MACKINNON, J. G. (1993). *Estimation and inference in econometrics*. Oxford: Oxford University Press.

DOWNS, G. W., & ROCKE, D. M. (1979). Interpreting heteroskedasticity. *American Journal of Political Science*, **23**, 816-828.

DRAPER, N. R., & SMITH, H. (1981). *Applied regression analysis* (2nd ed.). New York: Wiley.

DUNCAN, G. T., & LAYARD, W. M. J. (1973). A Monte-Carlo study of asymptotically robust tests for correlation coefficients. *Biometrika*, **60**, 551-558.

EDGELL, S. E., & NOON, S. M. (1984). Effect of violation of normality on the *t* test of the correlation coefficient. *Psychological Bulletin*, **95**, 576-583.

EICKER, F. (1963). Asymptotic normality and consistency of the least squares estimator for families of linear regression. *Annals of Mathematical Statistics*, **34**, 447-456.

EICKER, F. (1967). Limit theorems for regression with unequal and dependent errors. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Berkeley, CA: University of California Press.

FURNO, M. (1996). Small sample behavior of a robust heteroskedasticity consistent covariance matrix estimator. *Journal of Statistical Computation & Simulation*, **54**, 115-128.

GODFREY, L. G. (2006). Tests for regression models with heteroskedasticity of unknown form. *Computational Statistics & Data Analysis*, **50**, 2715-2733.

GODFREY, L. G., & ORME, C. D. (2004). Controlling the finite sample significance levels of heteroskedasticity-robust tests of several linear restrictions on regression coefficients. *Economics Letters*, **82**, 281-287.

GOLDFELD, S. M., & QUANDT, R. E. (1965). Some tests for homoskedasticity. *Journal of the American Statistical Association*, **60**, 539-547.

HAYES, A. F. (1996). Permutation test is not distribution-free: Testing H_0: $\rho = 0$. *Psychological Methods*, **1**, 184-198.

HINKLEY, D. V. (1977). Jackknifing in unbalanced situations. *Technometrics*, **19**, 285-292.

HUBER, P. J. (1967). The behavior of maximum likelihood estimation under nonstandard conditions. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Berkeley, CA: University of California Press.

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.

KOWALSKI, C. J. (1973). On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. *Applied Statistics*, **21**, 1-12.

LONG, J. S., & ERVIN, L. H. (2000). Using heteroskedasticity consistent standard errors in the linear regression model. *American Statistician*, **54**, 217-224.

MACCALLUM, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research*, **38**, 113-139.

MACKINNON, J. G., & WHITE, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, **29**, 305-325.

MARDIA, K. V., KENT, J. T., & BIBBY, J. M. (1979). *Multivariate analysis*. New York: Academic Press.

MOORE, D. S., & McCABE, G. P. (2003). *Introduction to the practice of statistics* (4th ed.). New York: Freeman.

OLEJNIK, S. F. (1988). Variance heterogeneity: An outcome to explain or a nuisance factor to control? *Journal of Experimental Education*, **56**, 193-197.

PERRY, D. K. (1986). Looking for heteroskedasticity: A means of searching for neglected conditional relationships. In M. L. McLaughlin (Ed.), *Communication yearbook 9* (pp. 658-670). Beverly Hills, CA: Sage.

RASMUSSEN, J. L. (1989). Computer-intensive correlational analysis: Bootstrap and approximate randomization techniques. *British Journal of Mathematical & Statistical Psychology*, **42**, 103-111.

SUDMANT, W., & KENNEDY, P. (1990). On inference in the presence of heteroskedasticity without replicated observations. *Communication in Statistics–Simulation & Computation*, **19**, 491-504.

WEISBERG, S. (1980). *Applied linear regression*. New York: Wiley.

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

WILCOX, R. R. (2001). Comment. *The American Statistician*, **55**, 374-375.

WILCOX, R. R. (2005). *Introduction to robust estimation and hypothesis testing*. New York: Academic Press.

WOOLDRIDGE, J. M. (2000). *Introductory econometrics: A modern approach*. Cincinnati, OH: South-Western College Publishing.

WU, C. F. J. (1986). Jackknife bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, **14**, 1261-1295.

NOTES

1. There are mixed opinions and inconsistencies in the literature about whether this ratio should be evaluated with reference to the *t* distribution or to the standard normal distribution for hypothesis testing. Whereas MacKinnon and White (1985, p. 314, Table 1 note) appear to use the standard normal distribution in their simulations, both Wilcox (2005, p. 420) and Wu (1986) use a *t* with $(n - p)$ *df* instead. Kauermann and Carroll (2001) reason that in finite samples, the distribution of this ratio is unknown, not always closely approximated by the standard normal,

and often, though not always, better approximated by the *t* distribution. In the macros we describe later, we use the $t (n - p)$ distribution for generating the *p* value for hypothesis tests.

2. It could be argued that GPA is not suitable as an outcome in OLS regression given that it is limited. Of course, limited outcome variables are routinely used as outcomes in OLS regression models, with unknown (and probably small in many circumstances) effect on the veracity of the substantive conclusions that investigators reach about the relationships of interest. We would argue that for such an outcome, the use of an HCSE method is particularly important if OLS regression is used, as the homoskedasticity assumption would be more likely to be violated when there are constraints on the upper and/or lower bounds of the outcome.

3. Wilcox (2005) describes some S-plus and R functions that implement some of the bootstrap methods that have been recently proposed.

APPENDIX
SPSS and SAS Macros for Heteroskedasticity-Robust Regression Using the HC Methods

SPSS and SAS have powerful matrix algebra-based sets of commands similar in their features to such programming languages as GAUSS, R, and MATLAB. These command sets, combined with the ability to generate new command syntax through the macro definition procedures, allow the user to add tremendous functionality to SPSS and SAS. Here we provide a set of macros that implement the HC methods described in this article in SPSS and SAS.

Each time a new SPSS session is started, the macro must first be activated by running the program. Once the program is run, a new command, HCREG, will be added to the SPSS syntax command set. Once the macro is activated, HCREG will remain a part of the syntax command set until the SPSS session is terminated by quitting SPSS. However, the macro must be reactivated each time SPSS is executed. The syntax of the HCREG command is `HCREG dv = yvar/iv = ivlist/const = c/method = m/covmat = cv/test = q`, where *c* is 1 for regression with a constant and 0 for regression without the constant, *m* is an integer between 0 and 5 corresponding to the HC method described in this article or 5 for ordinary standard errors that assume homoskedasticity, cv is 1 to print the variance–covariance matrix of the parameter estimates or 0 to disable its printing, *yvar* is the name of the criterion variable in the active SPSS data file, and *ivlist* is a list of one or more predictor variables in the SPSS file. The parts of the command underlined above are optional. When those parts are omitted, *c* defaults to 1, *m* defaults to 3, and cv defaults to 0. The "/*test* = *q*" subcommand is also optional and can be omitted. If it is included, the output will include a test that the regression coefficients for the last *q* variables in *ivlist* are all equal to zero. The value of *q* must be between 1 and one less than the number of variables in *ivlist*. There is no limit to the number of predictor variables that can be listed in *ivlist*. For example, the command set `HCREG dv = depress iv = gender age anxiety ptsd/test = 2` produces an OLS regression estimating depress from a constant, gender, age, anxiety, and ptsd and prints standard error estimates and significance tests using the HC3 estimator. It also produces a test of the null hypothesis that the regression coefficients for both anxiety and ptsd are equal to zero.

The macro has no error-handling procedures, so it is important that the data first be screened for such problems as singularities in the predictors and missing data. The presence of a singularity can be easily identified by seeing whether the SPSS OLS regression procedure rejects one or more of the predictors. The macro can handle system missing data represented with the period character ("."") through listwise deletion, but it does not recognize user missing values, so user-defined missing data should be deleted manually prior to the execution of the macro command. Problems during computation will show up with a string of largely unintelligible errors. If any errors appear, do not interpret the output. With large sample sizes, SPSS may produce an "out of memory" error. If this occurs, try increasing the workspace or memory allocated to SPSS (see the SPSS command reference documentation for details on the SET command to accomplish this).

The SAS version of the macro functions identically, although the syntax is slightly different. After the macro is activated, the new command is `%HCREG(data = filename, dv = yvar, iv = xvarlist, const = c, method = m, covmat = cv, test = q)`, where *filename* is the name of an SAS data file and the other parameters are defined as described above. The test parameter is optional, and c, m, and cv default to 1, 3, and 0, respectively. The SAS version of the macro also has no error handling and assumes that missing data are entered as "."

Potential users can request an electronic copy of the macros by sending an e-mail to hayes.338@osu.edu. Alternatively, the macros can be downloaded from www.comm.ohio-state.edu/ahayes/.

SPSS Version
```
DEFINE hcreg (dv =!charend ('/')/iv =!charend ('/')
            /test = !charend('/') !default (0)
            /const = !charend('/') !default(1)
            /method = !charend ('/') !default (3)
            /covmat = !charend('/') !default(0)).
PRESERVE.
set length = none.
SET MXLOOP = 100000000.
MATRIX.
GET x/file = */variables = !dv !iv/names = dv/missing = omit.
compute y=x(:,1).
compute x=x(:,2:ncol(x)).
```

APPENDIX (Continued)

```
compute iv5 = x.
compute pr = ncol(x).
compute n = nrow(x).
compute L = ident(pr).
compute tss=cssq(y)-(((csum(y)&**2)/n)*(!const <> 0)).
do if (!const = 0).
  compute iv = t(dv(1,2:ncol(dv))).
  compute df2 = n-pr.
else.
  compute iv = t({"Constant", dv(1,2:ncol(dv))}).
  compute con = make(n,1,1).
  compute x={con,x}.
  compute df2 = n-pr-1.
  compute L1 = make(1,pr,0).
  compute L = {L1;L}.
end if.
compute dv=dv(1,1).
compute b = inv(t(x)*x)*t(x)*y).
compute k = nrow(b).
compute invXtX = inv(t(x)*x).
compute h = x(:,1).
loop i=1 to n.
  compute h(i,1)= x(i,:)*invXtX*t(x(i,:)).
end loop.
compute resid = (y-(x*b)).
compute mse = csum(resid&**2)/(n-ncol(x)).
compute pred = x*b.
compute ess= cssq(resid).
 do if (!method = 2 or !method = 3).
  loop i=1 to k.
    compute x(:,i) = (resid&/(1-h)&**(1/(4-!method)))&*x(:,i).
  end loop.
 end if.
 do if (!method = 0 or !method = 1).
  loop i=1 to k.
    compute x(:,i) = resid&*x(:,i).
  end loop.
 end if.
 do if (!method = 5).
   loop i=1 to k.
    compute x(:,i) = sqrt(mse)&*x(:,i).
  end loop.
 end if.
do if (!method = 4).
 compute mn = make(n,2,4).
 compute pr3 = n-df2.
 compute mn(:,2) = (n*h)/pr3.
 compute ex=rmin(mn).
  loop i=1 to k.
    compute x(:,i) = (resid&/(1-h)&**(ex/2))&*x(:,i).
  end loop.
 end if.
compute hc = invXtX*t(x)*x*invXtX.
do if (!method = 1).
  compute hc = (n/(n-k))&*hc.
end if.
compute F = (t(t(L)*b)*inv(t(L)*hc*L)*((t(L)*b)))/pr).
compute pf = 1-fcdf(f,pr,df2).
compute r2 = (tss-ess)/tss.
compute pf = {r2,f,pr,df2,pf}.
do if (!method <> 5).
  print !method/title = "HC Method"/format F1.0.
end if.
print dv/title = "Criterion Variable"/format A8.
```

APPENDIX (Continued)

```
print pf/title = "Model Fit:"/clabels = "R-sq" "F" "df1" "df2" "p"/format F10.4.
compute sebhc = sqrt(diag(hc)).
compute te = b&/sebhc.
compute p = 2*(1-tcdf(abs(te), n-nrow(b))).
compute oput = {b,sebhc, te, p}.
do if (!method <> 5).
 print oput/title = 'Heteroskedasticity-Consistent Regression Results'/clabels
       = "Coeff" "SE(HC)" "t" "P>|t|"/rnames = iv/format f10.4.
 else if (!method = 5).
 print oput/title = 'OLS Regression Results Assuming Homoskedasticity'/clabels
       = "Coeff" "SE" "t" "P>|t|"/rnames = iv/format f10.4.
end if.
compute iv2 = t(iv).
do if (!covmat = 1).
 print hc/title = 'Covariance Matrix of Parameter Estimates'/cnames =
      iv/rnames = iv2/format f10.4.
end if.
do if (!test > 0 and !test < pr).
 compute L2 = make(pr-!test+!const,!test,0).
 compute L = {L2;L((pr+1-!test+!const):(pr+!const),(pr-!test+1):(pr))}.
 compute F = (t(t(L)*b)*inv(t(L)*hc*L)*((t(L)*b)))/!test).
 compute pf = 1-fcdf(f,!test,df2).
 compute pf = {f,!test,df2,pf}.
 print pf/title = "Setwise Hypothesis Test"
    /clabels = "F" "df1" "df2" "p"/format F10.4.
 compute iv = t(iv((pr+1-!test+!const):(pr+!const),1)).
 print iv/title = "Variables in Set:"/format A8.
end if.
END MATRIX.
RESTORE.
!END DEFINE.
```

SAS Version

```
%macro hcreg (data=,dv=,iv=,const=1,method=3,covmat=0,test=0);
proc iml;
use &data;
read all var{&dv &iv} into x;
nms={&iv};
xx=(x = .);xx=xx[,+];
j=1;do i=1 to nrow(x);if xx[i,1]=0 then;do;x[j,]=x[i,];j=j+1;end;end;
x=x[1:j-1,];
y=x[,1];
x=x[,2:ncol(x)];
pr=ncol(x);
n=nrow(x);
L=I(pr);
tss=ssq(y)-(((y[+,]*y[+,])/n)*(&const ^= 0));
if (&const = 0) then;
  do;
  df2 = n-pr;
  end;
if (&const = 1) then;
  do;
  nms={constant}||{&iv};
  con=j(n,1,1);
  x=con||x;
  df2 = n-pr-1;
  L1=j(1,pr,0);
  L=L1//L;
  end;
b=inv(x`*x)*x`*y;
k=nrow(b);
invXtx=inv(x`*x);
h=x[,1];
```

APPENDIX (Continued)

```
do i=1 to n;
  h[i,1]=x[i,]*invXtX*(x[i,]`);
end;
resid=(y-(x*b));
pred = x*b;
resid2 = resid##2;
mse = resid2[+,]/(n-ncol(x));
ess=ssq(resid);
if (&method > 1) then;
  do;
  if (&method < 4) then;
    do i = 1 to k;
    x[,i]=(resid/(1-h)##(1/(4-&method)))#x[,i];
    end;
  end;
if (&method < 2) then;
  do i = 1 to k;
  x[,i]=resid#x[,i];
  end;
if (&method = 5) then;
  do i = 1 to k;
  x[,i]=sqrt(mse)#x[,i];
  end;
mn1=j(n,1,4);
if (&method = 4) then;
  pr3=n-df2;
  mn2=(n*h)/pr3;
  ex = mn1 >< mn2;
  do i = 1 to k;
  x[,i]=(resid/(1-h)##(ex/2))#x[,i];
  end;
hc = invXtX*x`*x*invXtX;
if (&method = 1) then;
  do;
  hc=(n/(n-k))*hc;
  end;
F=((L`*b)`*inv(L`*hc*L)*(L`*b))/pr;
pf=1-probf(f,pr,df2);
r2=(tss-ess)/tss;
pf=r2||f||pr||df2||pf;
cn={"Coeff" "SE" "t" "P>|t|"};
if (&method < 5) then;
 do;print "HC Method" &method;
 cn={"Coeff" "SE(HC)" "t" "P>|t|"};end;

print "Criterion Variable" {&dv};
print "Model Fit";
cnms = {"R-sq" "F" "df1" "df2" "p"};
print pf [colname = cnms format = 10.4];
nmst=nms`;
sebhc=sqrt(vecdiag(hc));
te=b/sebhc;
p=2*(1-probt(abs(te),n-nrow(b)));
oput=b||sebhc||te||p;
if (&method < 5) then;do;
print "Heteroskedasticity-Consistent Regression Results";end;
if (&method = 5) then;do;
print "OLS Regression Results Assuming Homoskedasticity";end;
print oput [colname = cn rowname = nmst format=10.4];
if (&covmat = 1) then;
 do;print "Heteroskedasticity-Consistent Covariance Matrix";
 print hc [colname = nms rowname = nmst format=10.4];
 end;
```

APPENDIX (Continued)

```
if (&test > 0) then;
   do;if (&test < pr) then;do;
   L2=j(pr-&test+&const,&test,0);
   L=L2//L[(pr+1-&test+&const):(pr+&const),(pr-&test+1):(pr)];
   F=((L`*b)`*inv(L`*hc*L)*(L`*b))/&test;
   pf=1-probf(f,&test,df2);
   pf=f||&test||df2||pf;
   cnms = {"F" "df1" "df2" "p"};
   print "Setwise Hypothesis Test";
   print pf [colname = cnms format = 10.4];
   nmst = nmst[(pr+1-&test+&const):(pr+&const),1]`;
   print "Variables in Set: " nmst;
   end;
   end;
quit;
%mend hcreg;
```