Heteroskedasticity-robust tests of linear regression coefficients: A review and evaluation of small-sample corrections

James E. Pustejovsky and Gleb Furman
University of Texas at Austin
Educational Psychology Department

May 16, 2016

Abstract

The text of your abstract. 200 or fewer words.

Keywords: heteroskedasticity; sandwich estimator; robust covariance estimator; linear regression; Satterthwaite approximation; saddlepoint approximation; Edgeworth approximation

1 Introduction

Linear regression models, estimated by ordinary least squares (OLS), are one of the most important and ubiquitous tools in applied statistical work. Classically, hypothesis tests and confidence intervals for linear regression coefficients rely on the assumption that the model errors are homoskedastic, or have constant variance for all values of the covariates. However, it can be difficult to diagnose violations of this assumption—particularly in small samples—and so it is often desirable to use methods that do not rely on it. One common solution is to use a heteroskedasticity-consistent covariance matrix estimator (HCCME), which provides an asymptotically consistent estimate of the sampling variance of OLS coefficient estimates for models with heteroskedasticity of an unknown form. Generalizations of HCCME are also available that provide asymptotically consistent variance estimates for auto-correlated errors (Newey & West 1987, 1994) or errors that have an unknown dependence structure within clusters of observations (Liang & Zeger 1986).

HCCMEs were introduced by Huber (1967), Eicker (1967), and White (1980). MacKinnon & White (1985) proposed several variations to improve the finite-sample properties of the HCCMEs. In a large simulation study, Long & Ervin (2000) demonstrated that one of these variations, HC3, performs substantially better than the others. HC3 is the default in software such as the R package `sandwich` (Zeileis 2004), although the original HCCME (commonly called HC0) remains the default in SAS `proc reg` and Stata's `regress` command with `vce(robust)`. More recently, several further variations on the HCCMEs have been proposed (Cribari-Neto 2004, Cribari-Neto et al. 2007, Cribari-Neto & da Silva 2011). For hypothesis testing, HCCMEs are typically used to calculate t-statistics, which are compared to standard normal or $t(n-p)$ reference distributions, where n is the sample size and p is the dimension of the coefficient vector.

Another approach to improving the small-sample properties of hypothesis tests based on HCCMEs is to find a better approximation to the null sampling distribution of the test statistic. Several such approximations have been proposed, including Satterthwaite approximations (Lipsitz et al. 1999), Edgeworth approximations (Rothenberg 1988, Kauermann & Carroll 2001), and saddlepoint approximations (McCaffrey & Bell 2006). Although there is evidence that each of these approximations improves upon the standard, large-sample tests,

their performance has been examined only under a limited range of conditions. Moreover, it appears that these approximations have been developed independently and without reference to previous work, and their performance has never been compared under a common set of conditions. In contrast to the various HC corrections, to our knowledge, none of the distributional approximations are implemented in standard software packages for data analysis.

In this paper, we review the various small-sample approximations for hypothesis tests based on HCCMEs, using a common notation in order to facilitate comparisons among them. In so doing, we identify several further variations on the approximations that have not previously been considered. We then evaluate the performance of these approximations, along with the standard methods, in a large simulation study. The design of the simulation study is modeled on the earlier study of Long & Ervin (2000).

2 Methods

2.1 Model

We will consider the regression model

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i, \tag{1}$$

for $i = 1, ..., n$, where y_i is the outcome, $\mathbf{x}_i$ is a $1 \times p$ row-vector of covariates (including an intercept) for observation i, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ is a $p \times 1$ vector of regression coefficients, and ϵ_i is a mean-zero error term with variance σ_i^2. We shall assume that the errors are mutually independent. For ease of notation, let $\mathbf{y} = (y_1, ..., y_n)'$ denote the $n \times 1$ vector of outcomes, $\mathbf{X} = (\mathbf{x}_1', ..., \mathbf{x}_n')'$ be the $n \times p$ design matrix, and $\boldsymbol{\epsilon}$ be the $n \times 1$ vector of errors with $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, ..., \sigma_n^2)$. Let $\mathbf{M} = (\mathbf{X}'\mathbf{X}/n)^{-1}$. Finally, let $\hat{\boldsymbol{\beta}} = (\mathbf{M})\,\mathbf{X}'\mathbf{y}/n$ denote the vector of OLS estimates and $e_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$ denote the residual for unit i.

The goal is to test a hypothesis regarding a linear combination of the regression coefficients $\mathbf{c}'\boldsymbol{\beta}$, i.e., $H_0 : \mathbf{c}'\boldsymbol{\beta} = k$, with Type-I error rate α. All tests under consideration are

based on the Wald statistic

$$T(\mathbf{V}) = \frac{\mathbf{c}'\hat{\boldsymbol{\beta}} - k}{\sqrt{\mathbf{c}'\mathbf{V}\mathbf{c}}}, \tag{2}$$

where $\mathbf{V}$ is some estimator for Var $\left(\hat{\boldsymbol{\beta}}\right)$.

If the errors are homoskedastic, so that $\sigma_i^2 = \sigma^2$ for $i = 1, ..., n$, then the hypothesis can be tested using a standard t-test. The variance of $\boldsymbol{\beta}$ is then estimated by $\mathbf{V}^{hom} = \hat{\sigma}^2\mathbf{M}$, where $\hat{\sigma}^2 = \left(\sum_{i=1}^n e_i^2\right)/(n-p)$. Under H_0 and assuming that the errors are normally distributed and homoskedastic, the t-statistic follows a t distribution with $n-p$ degrees of freedom. Thus, H_0 is rejected if $\left|T\left(\mathbf{V}^{hom}\right)\right| > F_t^{-1}\left(1 - \frac{\alpha}{2}; n-p\right)$, where $F_t^{-1}(x;\nu)$ is the quantile function for a t distribution with ν degrees of freedom. However, if the errors are instead heteroskedastic, the variance estimator $\mathbf{V}^{hom}$ will be inconsistent and this t-test will generally have incorrect size.

2.2 HCCMEs

Under the general model that allows for heteroskedasticity, the true variance of the OLS estimator is

$$\text{Var}\left(\hat{\boldsymbol{\beta}}\right) = \frac{1}{n}\mathbf{M}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i^2\mathbf{x}_i\mathbf{x}'\right)\mathbf{M} \tag{3}$$

The HCCMEs all involve estimating Var $\left(\hat{\boldsymbol{\beta}}\right)$ by replacing the σ_i^2 with crude estimates involving the squared residuals. Although taken singly, the squared residual e_i^2 is a poor estimate of σ_i^2, together the residuals provide an adequate means of estimating the middle term of Equation (3). The HCCCMEs all have the general form

$$\mathbf{V}^{HC} = \frac{1}{n}\mathbf{M}\left(\frac{1}{n}\sum_{i=1}^n \omega_i e_i^2\mathbf{x}_i\mathbf{x}'\right)\mathbf{M} \tag{4}$$

where $\omega_1, ..., \omega_n$ are correction terms that differ for the various HC estimators. Under weak assumptions, the weak law of large numbers ensures that the middle term in Equation (4) converges to the corresponding term in (3) as the sample size increases. Furthermore, the robust Wald statistic $T\left(\mathbf{V}^{HC}\right)$ converges in distribution to $N(0,1)$ as n increases to infinity. Thus, any asymptotically correct test can be constructed by rejecting H_0 when $\left|T\left(\mathbf{V}^{HC}\right)\right|$ is greater than the $1-\alpha/2$ critical value from a standard normal distribution.

Because this test often has inflated size in small samples, it is common to instead use the critical value from a t distribution with $n - p$ degrees of freedom.

2.3 Rothenberg's Edgeworth approximation

Rothenberg (1988) developed an Edgeworth approximation for the distribution of Wald-type t-statistics based on the HC0 variance estimator. It is straight-forward to generalize the approach to any of the HC estimators. Let

$$
\begin{aligned}
g_i &= \mathbf{x}_i \mathbf{M} \mathbf{c} \\
f_i &= n \mathbf{x}_i \mathbf{M} \mathbf{X}' \mathbf{\Sigma} \mathbf{X} \mathbf{M} \mathbf{c} \\
\mathbf{Q} &= (\mathbf{I} - \mathbf{H}) \, \mathbf{\Sigma} \, (\mathbf{I} - \mathbf{H}) - \mathbf{\Sigma} \\
a &= \frac{\sum_{i=1}^n \omega_i g_i^2 z_i^2}{\left(\sum_{i=1}^n g_i^2 \sigma_i^2\right)^2} \\
b &= \frac{\sum_{i=1}^n \omega_i g_i^2 q_{ii}}{\sum_{i=1}^n g_i^2 \sigma_i^2} \\
\nu &= \frac{2 \left(\sum_{i=1}^n g_i^2 \sigma_i^2\right)^2}{\sum_{i=1}^n \omega_i^2 g_i^4 \sigma_i^4}
\end{aligned}
$$

For an observed value of the test statistic t_{HC}, the corresponding p-value is calculated as

$$
p = 2 \left[1 - \Phi \left[\frac{|t_{HC}|}{2} \left(2 - \frac{1 + t_{HC}^2}{\nu} + a \left(t_{HC}^2 - 1 \right) + b \right) \right] \right],
$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. A further approximation provides a means for calculating a critical value for a specified α-level. Let z_α denote the $1 - \alpha/2$ quantile from a standard normal distribution. Here, the hypothesis test is rejected if t_{HC} is greater than the critical value t_{crit} defined by

$$
t_{crit} = \frac{z_\alpha}{2} \left[2 + \frac{z_\alpha^2 + 1}{\nu} - a \left(z_\alpha^2 - 1 \right) - b \right].
$$

In practice, these testing procedures will need to be based on estimates of the quantities involved. Rothenberg proposed a simple estimate of the degrees of freedom:

$$
\nu_q = \frac{6 \left(\sum_{i=1}^n \omega_i g_i^2 e_i^2\right)^2}{\sum_{i=1}^n \omega_i^2 g_i^4 e_i^4}.
$$

Rothenberg also proposed to calculate a, b, $\mathbf{z}_q$, and $\mathbf{Q}$ by simply replacing the values of σ_i^2 with $\omega_i e_i^2$. Alternately, one could assume that $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, in which case $\mathbf{z} = \mathbf{0}$, $a = 0$,

$$b = -\frac{\sum_{i=1}^n h_i\omega_i g_i^2}{\sum_{i=1}^n g_i^2}, \qquad \text{and} \qquad \nu = \frac{2\left(\sum_{i=1}^n g_i^2\right)^2}{\sum_{i=1}^n \omega_i^2 g_i^4}.$$

2.4 Kauermann and Carroll's Edgeworth approximation

Kauermann & Carroll (2001)

2.5 Satterthwaite approximation

Lipsitz et al. (1999)

2.6 Saddlepoint approximation

McCaffrey & Bell (2006)

3 Simulation study

4 Conclusion

SUPPLEMENTARY MATERIAL

Title: Brief description. (file type)

R-package for MYNEW routine: R-package ?MYNEW? containing code to perform the diagnostic methods described in the article. The package also contains all datasets used as examples in the article. (GNU zipped tar file)

HIV data set: Data set used in the illustration of MYNEW method in Section 3.2. (.txt file)

References

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

Cribari-Neto, F. & da Silva, W. B. (2011), 'A new heteroskedasticity-consistent covariance matrix estimator for the linear regression model', *Advances in Statistical Analysis* **95**(2), 129–146.

Cribari-Neto, F., Souza, T. C. & Vasconcellos, K. L. P. (2007), 'Inference under heteroskedasticity and leveraged data', *Communications in Statistics - Theory and Methods* **36**(10), 1877–1888.

Eicker, F. (1967), Limit theorems for regressions with unequal and dependent errors, *in* 'Proceedings of the Fifth Berkeley symposium on Mathematical Statistics and Probability', University of California Press, Berkeley, CA, pp. 59–82.

Huber, P. J. (1967), The behavior of maximum likelihood estimates under nonstandard conditions, *in* 'Proceedings of the Fifth Berkeley symposium on Mathematical Statistics and Probability', University of California Press, Berkeley, CA, pp. 221–233.

Kauermann, G. & Carroll, R. J. (2001), 'A note on the efficiency of sandwich covariance matrix estimation', *Journal of the American Statistical Association* **96**(456), 1387–1396.

Liang, K.-Y. & Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**(1), 13–22.

Lipsitz, S. R., Ibrahim, J. G. & Parzen, M. (1999), 'A degrees-of-freedom approximation for a t-statistic with heterogeneous variance', *Journal of the Royal Statistical Society: Series D (The Statistician)* **48**(4), 495–506.

Long, J. S. & Ervin, L. H. (2000), 'Using heteroscedasticity consistent standard errors in the linear regression model', *The American Statistician* **54**(3), 217–224.

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

McCaffrey, D. F. & Bell, R. M. (2006), 'Improved hypothesis testing for coefficients in generalized estimating equations with small samples of clusters.', *Statistics in medicine* **25**(23), 4081–98.

Newey, W. K. & West, K. (1987), 'A simple, positive definite, heteroskedasticity and autocorrelation consistent covariance matrix', *Econometrica* **55**(3), 703–708.

Newey, W. K. & West, K. D. (1994), 'Automatic lag selection in covariance matrix estimation', *The Review of Economic Studies* **61**(4), 631–653.

Rothenberg, T. (1988), 'Approximate power functions for some robust tests of regression coefficients', *Econometrica* **56**(5), 997–1019.

White, H. (1980), 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity', *Econometrica* **48**(4), 817–838.

Zeileis, A. (2004), 'Econometric computing with HC and HAC covariance matrix estimators', *Journal of Statistical Software* **11**(10), 1–17.