From Replications to Revelations: Heteroskedasticity-Robust Inference (EXCERPT)

Sebastian Kranz, Ulm University*

December 2024
(First version November 2024)

Abstract

Analysing the Stata regression commands from 4420 reproduction packages of leading economic journals, we find that, among the 40,571 regressions specifying heteroskedasticity-robust standard errors, 98.1% adhere to Stata's default HC1 specification. We then compare several heteroskedasticity-robust inference methods with a large-scale Monte Carlo study based on regressions from 155 reproduction packages. Our results show that t-tests based on HC1 or HC2 with default degrees of freedom exhibit substantial over-rejection. Inference methods with customized degrees of freedom, as proposed by Bell and McCaffrey (2002), Hansen (2024), and a novel approach based on partial leverages, perform best. Additionally, we provide deeper insights into the role of leverages and partial leverages across different inference methods.

1 Motivation and Basic Insights

A considerable body of literature has proposed and recommended different specifications for heteroskedasticity-robust inference. For instance, based on Monte Carlo evidence, Long and Ervin (2000) strongly recommend for sample sizes below 250 observations HC3 standard errors rather than Stata's robust default: HC1.[1]

Is this recommendation widely followed in empirical practice? No, the opposite holds true. We analyse the code of regression commands found in the Stata scripts of 4420 reproduction packages from leading economic journals. 40,571 regression commands specify heteroskedasticity-robust standard errors and 98.1% stick with HC1 standard errors. Appendix A provides more details.

Would inference results substantially differ if the recommendation were more widely followed? We proceed with a subset of 608 regressions from 155 reproduction packages that have fewer than 1000 observations and can be reproduced using the current version of the toolbox *repbox*.[2] Appendix B provides details on the sample selection.

*Special thanks to Lars Vilhuber, Ben Greiner and all other data editors: without your awesome work, studies like this would not be possible. Also many thanks to Michal Kolesár, James MacKinnon, Enrique Pinzone and Michael Vogt for great discussions.

[1] Similar recommendations to adopt more robust inference methods for smaller samples have been made e.g. by MacKinnon and White (1985), who introduced HC1, HC2 and HC3, Chesher and Jewitt (1987), Chesher and Austin (1991), and Cattaneo, Jansson, and Newey (2018).

[2] The repbox tool chain consists of a series of R packages that I am developing to facilitate and semi-automate methodological meta-studies. This paper constitutes my first application of that toolchain. The toolchain automatically reproduces Stata code found in reproduction packages, performing essential tasks such as automatic file path correction. Additionally, it systematically stores information from regression commands and the underlying data sets, enabling the replication and modification of these regressions in both Stata and R. The general endeavor is complex and remains a work in progress. Currently, repbox does not yet robustly work for all reproduction packages that are theoretically reproducible. Thorough development and documentation will take a lot more time. Another part of repbox and area of

Table 1: Significance at 5% level for tests based on HC1 and HC3 robust standard errors

Interval of	$n \leq 250$		$250 < n \leq 1000$	
HC1 p-value	No. tests	Share HC3 $p \leq 0.05$	No. tests	Share HC3 $p \leq 0.05$
(0,0.01]	398	96.5%	253	99.2%
(0.01,0.02]	73	82.2%	30	96.7%
(0.02,0.03]	56	58.9%	34	70.6%
(0.03,0.04]	54	29.6%	17	58.8%
(0.04,0.05]	37	8.1%	22	9.1%

Note: The table includes only those significance tests from the 608 original regressions from 155 reproduction packages whose p-values under HC1 standard errors were below 5%. For each p-value interval, it shows the number of these tests and the fraction that remain statistically significant at the 5% level when HC3 standard errors are used instead. The first two columns show results for regression with sample size $n \leq 250$ and the remaining two columns for regressions with $250 < n \leq 1000$.

For Table 1 we explore significance tests for the null hypothesis that a true regression coefficient β_k is zero. We run each significance test twice: once with HC1 standard errors and once with HC3 standard errors. HC3 standard errors and p-values are always larger than their HC1 counterparts. Table 1 only considers the subsample of the tests for which the HC1 p-value is below 5% and shows for different intervals of that p-value the fraction of tests that are no longer significant at a 5% level if one uses HC3 standard errors instead.

The results show that the switch from HC1 to HC3 has substantial effects on statistical significance. For instance, fewer than 10% of HC1-based tests with p-values between 4% and 5% remain significant when using HC3. This result holds even for regressions with sample sizes ranging from 250 to 1,000 observations.

While these results are insightful, several important questions remain. For instance, to what extent are HC1-based p-values excessively low, and to what extent are HC3-based p-values overly conservative? Beyond HC3 standard errors, alternative methods for improving robust inference have been proposed. These include HC4 standard errors, introduced by Cribari-Neto (2004); HC2 standard errors with alternative calculations of degrees of freedom for the t-test, suggested by Bell and McCaffrey (2002) and Imbens and Kolesár (2016); Hansen's (2024) modified jackknife estimator, which also extend the degrees of freedom adjustments proposed by Bell and McCaffrey (2002); and wild bootstrap inference, as proposed e.g. by Wu (1986) and Roodman et al. (2019).

Do some of these methods perform better than others across typical situations encountered in economic analyses? How heterogeneous is a method's performance across different situations? Do applied researchers encounter scenarios where these established approaches systematically fail, and novel methods offer meaningful improvements?

To address these questions, we employ a large-scale Monte Carlo study. Monte Carlo simulations are a common tool in research on robust inference, but they typically examine only a small set of regression specifications, usually not based on real world data sets.[3] We perform Monte Carlo studies based on a large set of 608 OLS regressions originally conducted in 155 different reproduction packages of published economic articles. Ideally, our approach offers insights that are representative of the situations typically encountered by empirical researchers. Additionally, this broad scope enables us to

ongoing work, which is not yet used in this paper, is the automatic mapping of regressions from reproduction packages to the regression tables displayed in the corresponding articles. Thankfully, the Deutsche Forschungsgemeinschaft (DFG) supports future work on repbox as part of the larger SocEnRep project where it will also benefit from input from colleagues from computer sciences and social sciences.

[3] A notable exception is Young (2022), who conducts a large-scale Monte Carlo study based on instrumental variable regressions extracted from 30 reproduction packages of economic articles. Already, in his previous article, Young (2019), hand-collected reproducible regressions from a large set of 53 reproduction packages, but did not yet base the Monte Carlo simulations on those regressions.

investigate heterogeneity by examining how the performance of robust inference methods varies across the different original regression specifications.

In a nutshell, for each original regression of the form

$$y^o = X\beta^o + \varepsilon^o, \tag{1}$$

we specify a custom data generating process

$$y = X\beta + \varepsilon \tag{2}$$

with the same $n \times K$ matrix of explanatory variables X as in the original regression. The true coefficients, β, are set to zero. In line with the original researchers' assumptions, error terms are heteroskedastic. Concretely, we assume $\varepsilon_i \sim N(0, \sigma_i^2)$ for each observation $i = 1, \ldots, n$. The specification of the standard deviations, σ_i, of the error terms, ε_i, is a detailed process performed separately for each original regression. Initially, multiple candidate FGLS specifications, constructed using random forests, are estimated and calibrated. Subsequently, one candidate is selected by comparing the moments of the original OLS residuals with those of the OLS residuals obtained from Monte Carlo simulations of the various candidates. Details are provided in Appendix C.

For each original regression, we draw $M = 10,000$ Monte Carlo samples and compute the p-values for a t-test of the null hypothesis $\beta_k = 0$ for up to 25 coefficients, β_k, per regression. Coefficients of fixed effects dummies are not tested. Each of the 3280 tested coefficients from the 608 original regressions constitutes a distinct *test situation*, indexed by s.

For each test situation, we compare different test specifications $\tau \in \{\text{IID, HC1, HC2, } \ldots\}$, which vary by the type of standard error and the specification of the degrees of freedom used in the t-distribution. Mathematical background on the different specifications is provided in Section 2.

Let $p_{\tau,s}(m)$ denote the realized p-value for Monte Carlo sample $m = 1, \ldots, M$ in specification τ and test situation s. Our analysis focuses on the 5% significance level. The simulated rejection rate is defined as the proportion of Monte Carlo samples for which the p-value is below 0.05:

$$\pi_{\tau,s}^{0.05} = \frac{1}{M} \sum_{m=1}^{M} I(p_{\tau,s}(m) \leq 0.05) \tag{3}$$

where $I(\cdot)$ is the indicator function. Since the null hypothesis is true in all test situations, p-values should be uniformly distributed under a correctly specified t-test. Consequently, the ideal value of the rejection rate $\pi_{\tau,s}^{0.05}$ is 0.05.

We measure deviations from this ideal value using the excess and lack of the rejection rate, defined as:

$$\text{excess}_{\tau,s} = \max\left(\pi_{\tau,s}^{0.05} - 0.05, 0\right), \tag{4}$$

$$\text{lack}_{\tau,s} = \max\left(0.05 - \pi_{\tau,s}^{0.05}, 0\right). \tag{5}$$

While an excessive rejection rate increases the risk of false discoveries, lack in rejection rates can lead to under-powered significance tests. Excess is generally regarded as more problematic than an equally high lack. However, opinions may differ regarding acceptable levels of excess and the degree of lack one is willing to tolerate for a given reduction in excess.

Figure 1 shows for each specification τ the average excess and lack and their distribution across all 3280 test situations. Consistent with conventional wisdom, average excess decreases when moving in order from HC1, HC2, HC4, to HC3 standard errors, while average lack correspondingly increases.

More surprisingly, in our sample of regressions with no more than 1000 observations, both HC1 and HC2 yield on average more excessive rejection rates than inference based on i.i.d. standard errors, which is consistent only under homoskedasticity.[4] While HC3 is most conservative in the sense of

[4] Thus, adding the *robust* option to a Stata *regress* command, such that HC1 standard errors are used, may, in smaller sample sizes, misleadingly suggest that the resulting standard errors and test results are more conservative than without the *robust* option.

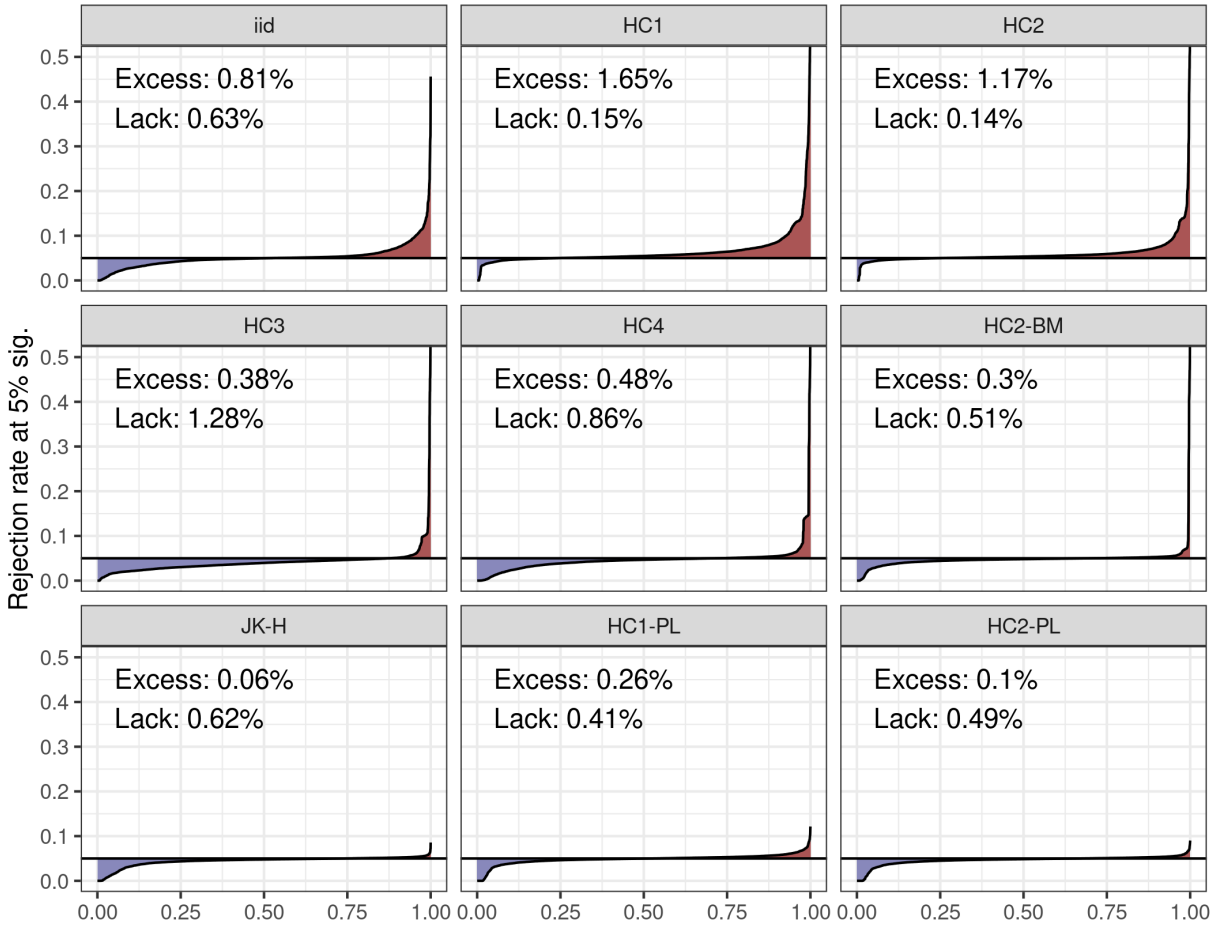

Figure 1: Core Results of Monte Carlo Study

Note: Each pane shows for a different specification of standard errors and degrees of freedom the distribution of rejection rate of t-tests with a 5% significance level across 3280 different regression coefficients from 608 regressions taken from 155 different reproduction packages. Red areas correspond to regression coefficients with excessive rejection rates (above 5%) and blue areas to those with lacking rejection rates (below 5%). For each specification the average excess and lack of the rejection rates across all regression coefficients is reported.

having the largest average lack in rejection rates, it is not the specification with the lowest average excess.

— END OF EXCERPT —