First Draft: 2021 12 17
This draft: 2021 01 09
Current draft: http://urisohn.com/43

JUST RUN ROBUSTER STANDARD ERRORS: A COMMENTARY ON YOUNG (2019)

Uri Simonsohn
ESADE Business School, Barcelona
urisohn@gmail.com

Abstract.
Young (2019) proposes that economists should either run balanced experiments (same number of observations in all conditions) or abandon regression analysis, running randomization tests instead. This recommendation arises from an artifactual result. Young relied on an outdated Stata default for computing robust standard errors. Here I use an alternative (the default in R), known since 1985 to be superior to Stata's default, finding that if anything, randomization tests perform worse than regression with *robuster* standard errors. In addition, Young analyzed 53 published experiments finding that 35% of p<.01 results become p>.01 by eliminating a single observation. This was interpreted as evidence of rampant presence of outliers in economics experiments, but actually, it is approximately what we expect in the *absence* of outliers.

R Code to reproduce all results available from:
https://researchbox.org/545 (use code **MTPQOB**)

I. INTRODUCTION

In a recent and influential article titled "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results", Young (2019) puts forward two results that are striking. First, re-analyzing data from 53 published experimental papers in economics, he finds that the statistical significance of published results is highly unstable; removing a single observation (or single *cluster* of observations when applicable) leads 35% of $p < .01$ results to become $p > .01$. Second, based on simulations, he finds that regression results are often invalid for experiments with uneven number of observations across conditions. Specifically, they exhibit high false-positive rates (above 20% in some cases). Analyzing those same simulated datasets through randomization tests, Young finds a false-positive rate much closer to the nominal 5% (see brief description of randomization tests in footnote 1).[1]

In light of these results, Young calls on economists to either only conduct experiments with a balanced design, where all conditions have the same number of observations, or to analyze unbalanced designs with randomization tests instead of with regressions (see direct quotes in footnote 2).[2] Two years after publication, his article has about 400 google citations, including 6 papers published in the *Quarterly Journal of Economics*.

Here I revisit these results and recommendations. In terms of the impact of removing one observation on a $p < .01$ result, I find that the observed rate at which those results become $p > .01$, is actually about what we should expect in the *absence* of outliers. Moreover, I explain why *p*-hacking is a more likely explanation for instability of statistical significance than is the alluded presence of outliers in the raw data.

[1] Randomization tests consists of simulations were data are resampled in the same way in which random assignment was conducted in an experiment. In simple designs, randomization tests can be achieved by reshuffling the column in the data that contains the condition assignment. After each reshuffle the test-statistic (e.g., t-value for the different of means t-test) is recomputed, and the proportion of shuffled datasets with a test-statistic at least as extreme as that observed in the data is the randomization *p*-value. Randomization tests were first discussed by Fisher (1935) and first developed by Pitman (1937). For a historical overview see David (2008).

[2] Quotes from Young's Conclusions section: "more accurate results might be achieved by breaking the experiment and regression into groups . . . each with a balanced treatment design" (p.596) and "[otherwise] randomization tests . . . provide a means to construct tests with credible finite-sample rejection probabilities" (p.597)

More importantly and consequentially, in terms of the proposed superiority of randomization tests, I show that Young's results arise because he, like most economists, computed robust standard errors relying on an outdated Stata default. When robust standard errors are computed using the default setting in R, a setting that's also available in Stata, a setting shown over 35 years ago to be *robuster* to heteroskedasticity in small samples (MacKinnon and White 1985), the superiority of the randomization tests evaporates. Randomization tests are not superior, Stata's default is inferior.[3]

II. ROBUST AND ROBUSTER STANDARD ERRORS

It has long been recognized that heteroskedasticity biases standard errors in regression models. It typically leads to downward bias: the standard errors are too small, therefore the t-values too big, and the *p*-values too often $p<.05$; an inflated false-positive rate (see brief discussion of when robust SE are smaller vs bigger in footnote 4).[4] A consistent estimator of the standard error in the presence of heteroskedasticity was developed relatively late in the history of statistics (Eicker 1963), and brought to the attention of economists some 17 years later, in the seminal article by White (1980).[5] Within a few years, MacKinnon and White (1985) built on this work, looking for solutions that were not only consistent (a property expected with *infinite* samples), but also unbiased (a property expected with *finite*, possibly small samples). They added three methods for computing standard errors, and referred to the full set of four as HC0, HC1, HC2 and HC3; henceforth, HC*k*s for short.

HC*k*s have in common that they compute standard errors for regression coefficients based on a *weighted* average of the (squared) regression residuals. They differ in what those weights are. For example,

[3] Everything in R can be done in multiple ways; the default I am referring to is the one provided by the popular package 'sandwich' (https://cran.r-project.org/web/packages/sandwich/index.html)

[4] Heteroskedasticity leads to standard errors that are too big, instead of too small, when more extreme values of the predictors lead to less, rather than more, variance in the dependent variable. E.g., if the regression is $y=\mathbf{a}+\mathbf{b}x+e$, and the error term is distributed $e\sim(0, sd = 1/(1+x^2)$, y will have less variance when x is more extreme, robust standard errors will be smaller than uncorrected standard errors. If $\mathbf{b}=0$, the false-positive rate for its estimate, $\hat{b}$, when relying on uncorrected standard errors, is *less* than 5%.

[5] Giving credit for conceptual advancements can be hazardous, so let me quote White (1980) himself, who writes "*results similar to propositions (i) and (ii) of Theorem 1 were stated over a decade ago by Eicker, . . . It is somewhat surprising that these very useful facts have remained unfamiliar to practicing econometricians for so long*" (p.821). MacKinnon (2013) notes that the paper by White (1980) appears to be the most cited article in the economics literature.

HC1 weighs observations by how far they are from the mean, while HC3 weighs by how much leverage (impact on the point estimates) each observation has. Distance from the mean and leverage are related to one another, but are not identical, thus HC1 and HC3 give different results.

Upon reflection, the computations behind HC3, behind weighting residuals by their leverage, make intuitive sense. It is tautologically true that residuals from observations with higher leverage are more consequential for the regression coefficients, and if they impact the coefficient more, then they impact the coefficient's variation more as well (i.e., the standard error). In other words, variation of the dependent variable among observations with high leverage has more impact on the *actual* variability of the regression coefficients than does variation in the dependent variable among observation with low leverage, and HC3 gives those observations with greater leverage, with more impact on true variability, more weight when *estimating* said variability.

MacKinnon and White (1985) relied on simulations to compare the performance of the HC*k*s, concluding rather unambiguously that "HC3 is the . . . estimator of choice" (p.313). Relying on more powerful computers and a more systematic evaluation, Davidson and MacKinnon (1993) also conducted simulations to compare the performance of HC*k*s, and also concluded that HC3 should be used. Relying on even more powerful computers and even more systematic evaluation, Long and Ervin (2000) also conducted simulations to compare the performance of HC*k*s, and also concluded the HC3 should be used, writing that HC3 was "almost always superior to … HC2, HC1 and HC0" (p.222). Given this history of unabated success, it is puzzling and disappointing that Stata, the most popular software for statistical analysis among economists, uses HC1 as a default robust standard errors (e.g., when running '*reg y x, robust*'). It is as if Stata, after being given the option to play a roster of all-pros vs. a roster of Detroit Lions, decided that the latter was the winning arrangement.

It is, moreover, this underdog HC1 procedure, with a 35 year losing streak, that Young (2019) once again observed lose a horse-race inside a computer simulation. Except this time, the long-standing champion, HC3, did not enter the race. In the next section I rerun the simulations performed by Young, now including the procedure statisticians and econometricians have for decades been telling us to use: HC3.

III. ADDING HC3 TO THE SIMULATIONS RAN BY YOUNG (2019)

Young (2019) compares false-positive rates for HC1 vs randomization tests, for a variety of scenarios characterizing economics experiments that differ in sample size, how balanced the observation allocation is, and heterogeneity of effect size (and thus, indirectly, heteroskedasticity).[6] He finds that regression results (with robust standard errors a-la HC1), have false-positive rates that are too high, rejecting the null up to 21% of the time instead of the nominal 5%, while the randomization tests perform much closer to that nominal 5%.

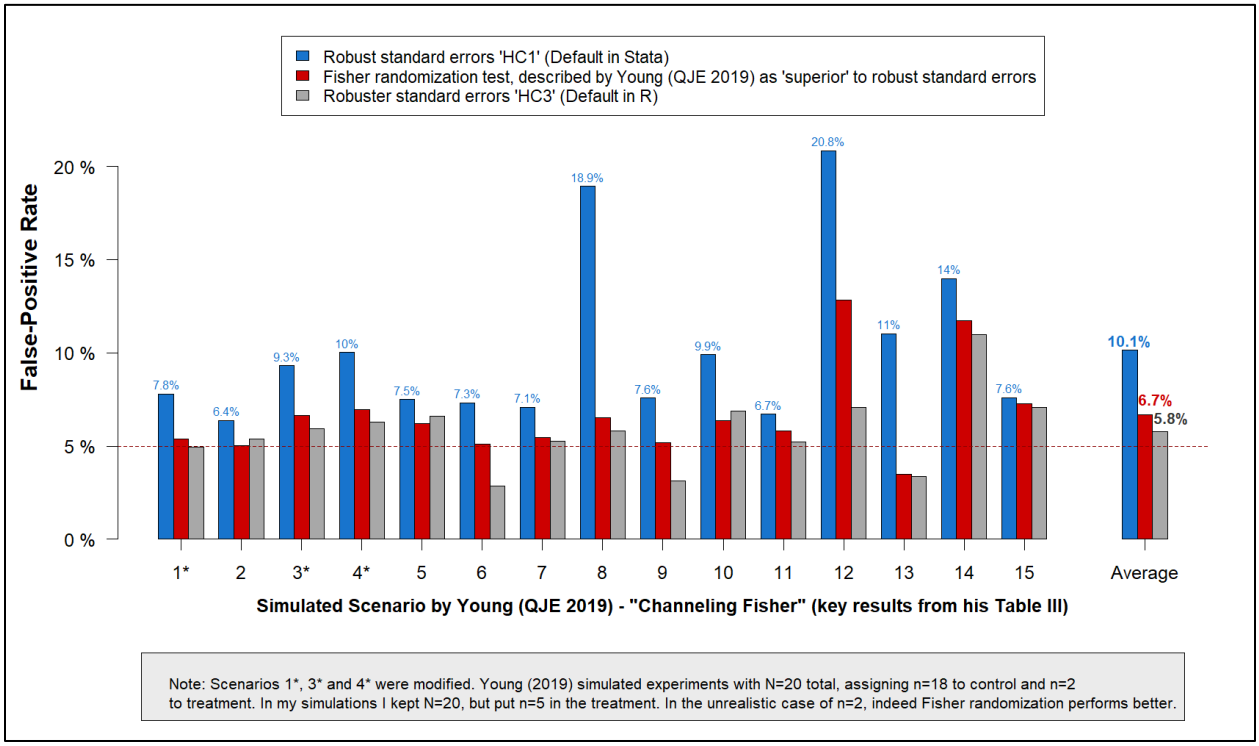
The simulated scenarios are rather extreme (e.g., the most extreme, involve experiments with n=20 participants total, allocating just n=2 to treatment), making Stata's default, HC1, perform worse in this simulated world than it probably would in the real world. While I run very similar simulations to enhance comparability of results, I should note that I do not expect HC1 to perform as poorly as reported here or in Young's paper. Indeed, it is conceivable that the choice of HC1 vs HC3 is not often consequential in the real world. But, to be clear, there is no downside to switching to HC3, and there is potential upside.

In any case, Young's (2019) simulation results are reported in his Table III. In most of the scenarios he considers, robust standard errors (HC1) actually perform well, with about a 5% false-positive rate. Here I focus on the subset of 15 scenarios where they did not perform well, where HC1 has a false-positive rate above 6.5% (see Figure S1 in the supplement for a visual depiction of the subset of analyses in his Table III which I re-run here).

Following the descriptions in his article, I generally reproduced his reported false-positive rates for HC1 and randomization tests (see supplement 2 for a detailed comparison of results). I then also computed results for *robuster* standard errors (HC3). Figure I below shows the results.

[6] Young's simulations include four other approaches for computing *p*-values under heteroskedasticity, but they do not perform best and are not recommended in the article, so I do not discuss them here.

Figure I. False-Positive Rates for 15 Scenarios Simulated by Young (2019)



Notes: The simulations by Young (2019) included calculations based on HC1 robust standard errors and based on randomization tests. I reproduced the posted results and added results based on HC3. The 15 scenarios are a subset of all scenarios included in his Table III, the subset where HC1 obtained a false-positive rate above 6.5%.
R Code to reproduce simulations available from http://researchbox.org/545.13 (use code MTPQOB)

The differences in Figure I between blue and red bars, reproduce Young's results: Stata's default for robust standard errors (HC1) performs worse than do randomization tests in those scenarios. The (mostly) lack of a difference between the red and gray bars show that this is because Stata's default is especially bad, rather than because randomization tests are especially good. If anything, robuster standard errors, HC3, outperform randomization tests in these scenarios (overall false-positive rate of 5.8% < 6.7%)

IV. DO ECONOMICS EXPERIMENTS SUFFER FROM AN OUTLIERS PROBLEM?

Young (2019) re-analyzed data from 53 published economics papers reporting experiments and concluded "one of the central characteristics of my sample [of economics experiments] is its remarkable sensitivity to outliers" (p.566), adding "With the removal of just one observation, 35% of .01-significant reported results in the average paper can be rendered insignificant at that level." (p.567).

Young (2019) does not provide a benchmark to compare that 35% with. Is it actually a surprisingly high number? For example, imagine the datasets of all 53 papers did *not* have "as a central characteristic" the presence of outliers, moreover, imagine they were absolutely *free* of outliers, how often should the post-hoc removal of the single most impactful observation turn a $p < .01$ result into a $p > .01$?

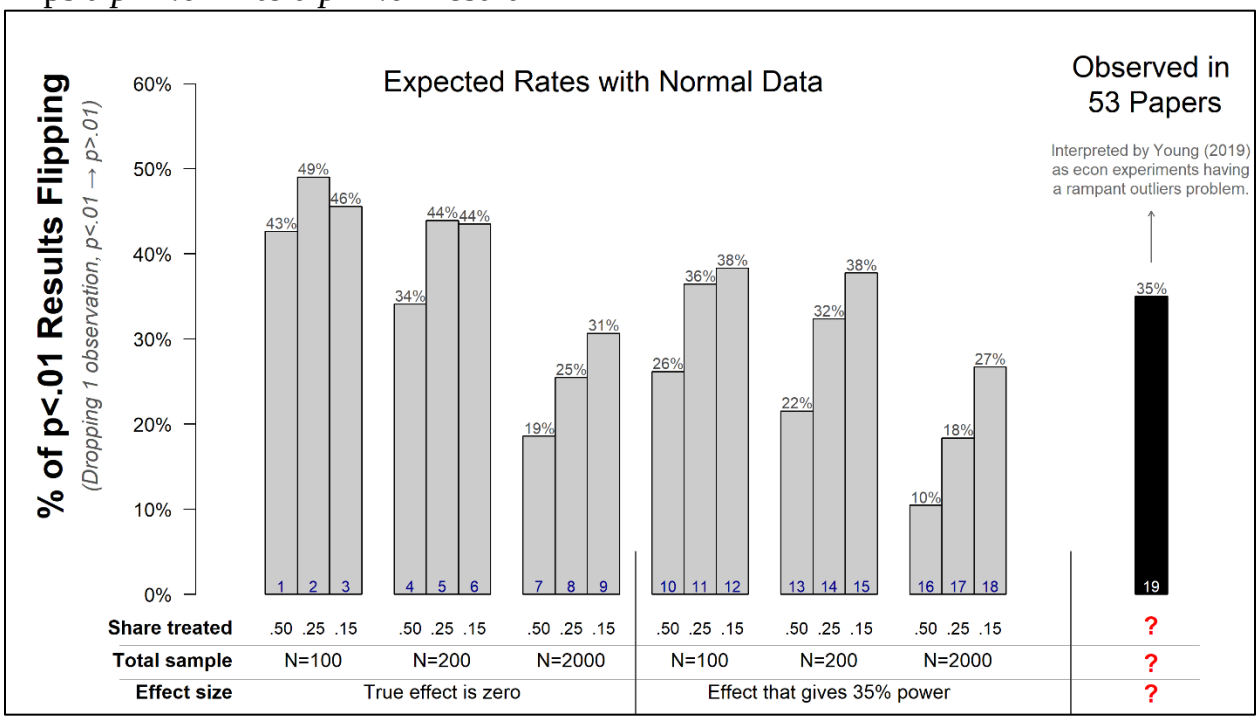
The probability that a p<.01 turns into a p>.01 by removing an observation naturally depends on how many of those p<.01s are close to the threshold, say p=.009 vs far from it, say p=.0000001. The distribution of *p*-values, in turn, depends on several factors, most notably on statistical power (which in turn depends on sample size, effect size, and how balance the design is).

For example, if the true effect is 0, if 'power' is 5%, we expect just as many *p*-values between p=.001 and p=.002 as we do between p=.009 and p=.01 (uniform distribution). In contrast, when power is high, say 80%, we expect many more low rather than high *p*-values. Indeed, we expect about *five* times as many *p*-values between p=.001 and p=.002 as we do between p=.009 and p=.01 (see calculations in Supplement 3). Because we do not know the true effect size, and Young does not report the sample size and allocation to condition of each test in his sample, we do not know underlying power for each test, and thus we cannot conclusively assess whether 35% of results flipping from $p < .01$ to $p > .01$ is too high, too low, or about what we should expected in the absence of outliers. But, we can arrive at an informed calibration, which is what I attempt to do next.

I simulated *normally* distributed data, without adding any extreme values; data free of outliers. Across simulations I vary sample size, allocation to control vs treatment, and true effect size. I consider two effect sizes, zero, and an effect producing 35% power (because, overall, 35% of test results analyzed by Young are p<.05, so roughly speaking, average power in his sample). Before getting to the results, there is an important caveat with relying on average power for the calibration. The relationship between sample size, effect size, and power, is not linear, and therefore, the results in terms of expected flipping from p<.01 to p>.01 are not determined by just knowing average power. It depends on whether *all* tests have 35% power, or some have 90%, some have 5%, and *on average* they have 35%. This is just a calibration exercise.

Figure II below shows the results of this calibration. We see that in general the expected share of p<.01 that become p>.01 by excluding a single observation, in the absolute absence of outliers, is quite high, often higher than that reported by Young to be the case in the experimental economics literature.

Figure II. Expected vs Observed Shares of Tests Where Dropping 1 Observation Flips a $p < .01$ into a $p > .01$ result



Notes: The figure reports simulation results where two conditions, with total sample size N, are drawn from the Normal distribution (no outliers). A t-test is conducted comparing the two means using the full sample and then also excluding the single observations in the simulated study that most impacts the t-test's *p*-value. The bars in the figure depict the share of simulated studies where excluding that single observation eliminated a p<.01 result. The first 9 bars draw data for both conditions from populations with identical means, the next 9 bars from populations with means that differ by an effect big enough to produce 35% power (for obtaining $p<.05$). For example, the 11th bar depicts simulations where the difference in population means gives 35% power to a t-test when one condition has n=25 observations and the other has n=75. The level of power (35%) was set to match the overall share of results analyzed by Young that are p<.05 (his Table V, top-row, second column). The 19th bar depicts the result reported by Young (2019). The "?" below the 19th bar symbolize that we do not have the information necessary to evaluate that 35% result, we don't know sample sizes, shared treated, or true effect size behind that 35% result.
R Code to reproduce figure available from http://researchbox.org/545.1 (use code **MTPQOB**).

In light of results from Figure II, it is unclear what to make of Young's results. It is perhaps the case that 35% is somewhat higher than we would expect for normally distributed data, but it's not obviously the case, nor whether such excess is statistically significant; that is, whether we see a *statistically significant*

higher share of results flipping than we would expect under the null of absolute absence of outliers. But, even if we knew that the 35% number is higher than expected, and knew it was statistically significantly higher than expected, outliers are not the most plausible explanation. P-hacking is.

V. P-HACKING, PUBLICATION BIAS, AND UNSTABLE STATISTICAL SIGNIFICANCE

The calibrations from Figure II were drawn from normally distributed data. This assumption seems stronger than it is, after all, t-tests compare means, not raw data, and thanks to the central limit theorem, even if the underlying dependent variable is not normally distributed, its mean will tend to be, and thus t-test will behave very similarly when applied to data that are normally distributed (e.g., in the simulations) and when applied to data that are not (e.g., to data from 53 economics experiments).

But another assumption in those simulations, a tacit one which probably goes unnoticed by most readers, is much stronger and consequential: the assumption that any given study is analyzed only once by the researcher, and whatever result is obtained in that single analysis is then reported. This is not how data analysis happens in the real world. Instead, we collect data, we analyze it in many ways, and we publish a subset of those analyses. Moreover, we do not report a random subset, we report a cherry-picked subset. We are more likely to report analyses that obtain significance than to report analyses that do not obtain significance. A behavior we have proposed referring to as '*p*-hacking' (Simonsohn, Nelson and Simmons 2014)

Importantly, *p*-hacking can shift distributions of p-values away from what we expect in simulations like those behind Figure II, moving density from lower to higher *p*-values. For example, under the null, when control and treatment have identical means, the distribution of *p*-values is uniform, $p=.0099$ (just below .01) is just as likely as $p=.0001$ (well below .01); but many forms of p-hacking shift that distribution, making it left skewed, such that $p=.0099$ is *more* likely than $p=.0001$ (Simonsohn, Nelson and Simmons 2014). *p*-hacking, then, produces *p*-values that are closer to .01 and thus more likely to flip, after *any* change in analysis, than the simulations assume.

"Economists *p*-hack", as a explanation for a possible excessive lack of robustness of statistically significant results, is an alternative to the explanation by Young which is, paraphrasing, that "economist happen to analyze datasets that have too many outliers". It matters which explanation is right. If we believe that - for some unstated reason- outliers are surprisingly common in economics data, then we need to rely on statistical tools that are robust to outliers. But if outliers are not the problem, if they are just a symptom, of the underlying problem which is p-hacking, then this solution of more robust-to-outliers statistical analysis will not help very much. Researchers can p-hack randomization tests just like they can p-hack regressions results. What we would need is a policy that prevents or at least mitigates p-hacking. That tool consists of (properly implemented) pre-registration.

VI. THE VALUE OF RANDOMIZATION TESTS

The results reported here show that when it comes to comparing means in simple experiments, randomization tests offer no obvious advantages over *robuster* standard errors (HC3). Even tiny (n=20) and unbalanced samples (n=5 vs n=15), with strong heteroskedasticity, are handled appropriately by the latter. Moreover, randomization tests have some unique shortcomings. First, randomization tests typically require that authors implement on their own the coding of the test, instead of relying on an off-the-shelf solution, e.g., simply typing *'reg y x, vce(hc3)'*. If the research design is complex (e.g., nested structure, or stratified samples) or the analysis is complex (e.g., a two-stage estimator) this programing task is far from trivial, increasing the risk of human error. Second, at a more conceptual level, randomization tests provide results *under the null*, and do not provide information regarding precision around an estimate (e.g., standard errors or confidence intervals). This is particularly consequential for nonlinear models (e.g., logit regression). Moreover, as Young (2019) discusses at some length, randomization tests often require imposing a stronger null than the researchers are strictly interested in (imposing a 'sharp' null).

While randomization tests are unnecessary for the majority of cases, they can be enormously valuable for a minority of them. They are useful when no accepted statistical test exists (e.g., to compare maxima instead of means), they are useful to run tests on complex data structures (e.g., to test joint nulls

for non-independent models run on the same dataset (Simonsohn, Simmons and Nelson 2020), or on data with complex network dependencies (Bond, Fariss, Jones, Kramer, Marlow, Settle and Fowler 2012)). Randomization tests are useful for correcting for multiple comparisons with correlated data (Petrondas and Gabriel 1983, Westfall 1993). Randomization tests are useful for testing unusual null hypotheses, such as whether data have been tampered with (Simonsohn 2013). But, for analyzing data from simple experiments with n>5 per cell, randomization tests are at best inconsequential.

VII. CONCLUSIONS

The influential article by Young (2019) called on economists running experiments to either restrict their designs to have the same number of observations across conditions, or to abandon regression analysis. In this article I have argued that that call is unjustified. A much simpler solution is to use robuster standard errors (HC3). Instead of abandoning all the knowledge economists have accumulated about regression analyses, and switch to custom coding randomization tests, economists really just need to type a different 6 letter string at the end of their regression commands (find *',robust'*, replace *',vce(hc3)')*.

Last but not least, Young focused only on experimental data, but heteroskedasticity is hardly a problem that afflicts only experiments. If robust standard errors are insufficiently robust for economics experiments, they are insufficiently robust for economics *non*-experiments. High leverage observations are common, after all, in observational data. Randomization tests are not easy to implement outside experimental settings. The call I make here, to use *robuster* instead of robust standard errors, is equally justified, and equally easy to follow, for experimental and non-experimental data. As MacKinnon and White (1985) put it -during the Reagan administration- "HC3 is the . . . estimator of choice".

References

Bond, Robert M, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler, "A 61-million-person experiment in social influence and political mobilization," *Nature*, 489 (2012), 295-298.

David, Herbert A, "The beginnings of randomization tests," *The American Statistician*, 62 (2008), 70-72.

Davidson, Russell, and James G MacKinnon, *Estimation and inference in econometrics* (New York :: Oxford University Press, 1993).

Eicker, Friedhelm, "Asymptotic normality and consistency of the least squares estimators for families of linear regressions," *The annals of mathematical statistics*, (1963), 447-456.

Fisher, Ronald A, "The Design of Experiments (8th," (Oliver and Boyd, Edinburgh, 1935).

Long, J Scott, and Laurie H Ervin, "Using heteroscedasticity consistent standard errors in the linear regression model," *The American Statistician*, 54 (2000), 217-224.

MacKinnon, James G, "Thirty years of heteroskedasticity-robust inference," in *Recent advances and future directions in causality, prediction, and specification analysis*, (Springer, 2013).

MacKinnon, James G, and Halbert White, "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," *Journal of econometrics*, 29 (1985), 305-325.

Petrondas, Demetrios A, and K Ruben Gabriel, "Multiple comparisons by rerandomization tests," *Journal of the American Statistical Association*, 78 (1983), 949-957.

Pitman, E.J.G, "Significance tests which may be applied to samples from any populations," *Journal of the Royal Statistical Society*, 4 (1937), 119-130.

Simonsohn, Uri, "Just post it: the lesson from two cases of fabricated data detected by statisics alone," *Psychological science*, 24 (2013), 1875-1888.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons, "*p*-curve: A Key to the File Drawer," *Journal of Experimental Psychology: General*, 143 (2014), 534-547.

Simonsohn, Uri, Joseph P Simmons, and Leif D Nelson, "Specification curve analysis," *Nature Human Behaviour*, 4 (2020), 1208-1214.

Westfall, Peter H, *Resampling-based multiple testing: Examples and methods for p-value adjustment* (John Wiley & Sons, 1993).

White, H., "A Heteroskedasticity-Consistent Covariance-Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48 (1980), 817-838.

Young, Alwyn, "Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results," *The Quarterly Journal of Economics*, 134 (2019), 557-598.